Metacognition for a Common Model of Cognition

Jerald D. Kralik[1], Jee Hang Lee[2], Paul S. Rosenbloom[3], Philip C. Jackson, Jr.[4], Susan L. Epstein[5], Oscar J. Romero[6] , Ricardo Sanz[7] , Othalia Larue[8], Hedda R. Schmidtke[9], Sang Wan Lee[1,2], Keith McGreggor[10]

[1]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 34141, South Korea
[2]KI for Health Science and Technology, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 34141, South Korea
[3]Institute for Creative Technologies & Department of Computer Science, University of Southern California, Los Angeles, CA, USA
[4]TalaMind LLC, PMB #363, 55 E. Long Lake Rd., Troy, MI, USA
[5]Department of Computer Science, Hunter College and The Graduate Center of the City University of New York, New York, NY, USA
[6]Department of Machine Learning, Carnegie Mellon University, Pittsburgh, PA, USA
[7]Autonomous Systems Laboratory, Universidad Politécnica de Madrid, Spain
[8]Department of Psychology, Wright State University, Dayton, Ohio, USA
[9]Department of Geography, University of Oregon, Eugene, OR, USA
[10]College of Computing, Georgia Institute of Technology, Atlanta, Georgia, USA

jerald.kralik@kaist.ac.kr, jeehang@kaist.ac.kr, rosenbloom@usc.edu, dr.phil.jackson@talamind.com susan.epstein@hunter.cuny.edu, oscarr@andrew.cmu.edu, ricardo.sanz@upm.es, othalia.larue@wright.edu, hedda.schmidtke@gmail.com, sangwan@kaist.ac.kr, keith.mcgreggor@gatech.edu

Abstract

This paper provides a starting point for the development of metacognition in a common model of cognition. It identifies significant theoretical work on metacognition from multiple disciplines that the authors believe worthy of consideration. After first defining cognition and metacognition, we outline three general categories of metacognition, provide an initial list of its main components, consider the more difficult problem of consciousness, and present examples of prominent artificial systems that have implemented metacognitive components. Finally, we identify pressing design issues for the future.

1. Introduction

The goal of this paper is to begin the development of a consensus model of metacognition that spans all relevant fields, including cognitive science, philosophy, neuroscience, and robotics. In what follows, we first define metacognition, then outline general categories of it, list its major components, discuss its relationship to consciousness, address key design issues, and present case studies with metacognition successfully implemented computationally. Finally, we briefly address the next steps for the project.

2. Metacognition Defined

To define metacognition, we must begin with cognition itself. Cognition is defined differently across fields and contexts. Because an intelligent agent executes a repeating perceive-decide-act cycle, we define *cognition* to capture that cycle, thus incorporating perception and action (Newell, 1990). Here, "perceive" subjects the agent to a continual barrage of signals (e.g., visual, auditory, olfactory) that describe the agent's context. These signals are necessarily a partial description of the environment in which the agent exists. The "decide" portion of the cycle — the focus of this paper — is also treated differently across fields. Here, it is used it broadly, to capture intermediate processes that culminate in a decision about which action(s) to execute. Necessitated by the incomplete, and possibly inconsistent, messages an intelligent agent receives, decision incorporates, but is not restricted to, a wealth of processes. These include, attention, reasoning, learning, planning, imagination, conscious access, and communication and understanding through natural language (Dehaene, 2014). Finally, "act" represents the ultimate outcome of the cognitive cycle, one that typically results in external motor responses (e.g., via muscles, actuators).

Simply put, metacognition is cognition about cognition. Thus it includes, for example, reasoning about reasoning, reasoning about learning, and learning about reasoning (Jackson, 2014; Kralik, 2017; Nelson, 1992). Broadly construed, it is any cognitive process or structure *about* another cognitive process or structure (e.g., data about memory held in memory). Here we focus on cognitive processes applied to cognitive processes (Jackson, 2014), a kind of recursive processing illustrated in Fig. 1. If a particular process is of Type-*X* (where *X* is perception, decision, or action) and receives input from another Type-*X* process, it is considered a metaprocess, and therefore metacognition. In fact, it is considered metacognitive if and only if the process receives input from, sends output to, or both receives from and sends to the same process type.

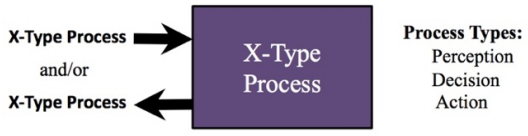

Figure 1. A process is metacognitive if and only if it receives input from, sends output to, or both receives from and sends to the same process type. A process type is perception, decision, or action.

Metacognition addresses what the system knows, the importance of what is known, as well as what has been

remembered and what is worth remembering or forgetting. Its advantages include arbitration underlying competing functions; modulation to help finetune other cognitive processes; safeguards against confusion and errors from lower cognitive processes (especially those designed for efficiency and specialization); and data management to reduce inefficiencies (e.g., removal of obsolete information by forgetting). The next section further clarifies this definition.

3. General Categories of Metacognition

This section delineates general categories of human-level metacognition based on their input and output. It considers these categories in turn and provides prominent examples of each, along with empirical evidence and a description of their key properties. For clarity and brevity, the focal central process is "decide." We align metacognition closely with cognition itself by first identifying cognition as Category 0, the mapping of perception into action (Fig. 2).

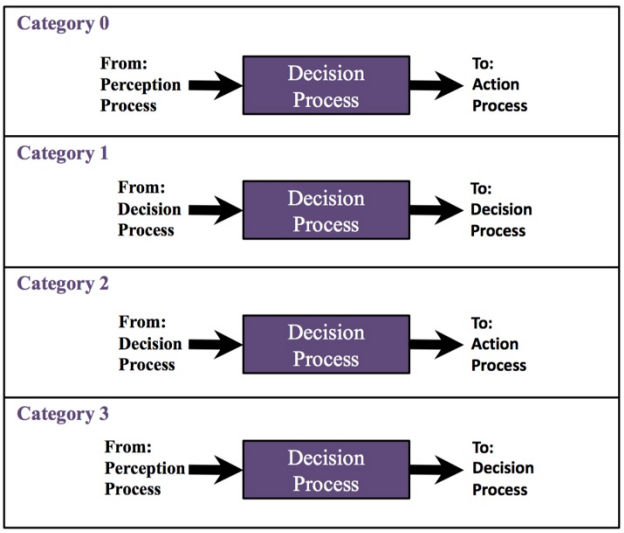

Figure 2. Four categories consistent with Figure 1. Category 0 is cognition itself, with primary input from perception and primary output to action control. Categories 1, 2, and 3 comprise metacognition. Category 1: Primary input and output from and to decision processes. Category 2: Primary input from decision processes; primary output feeds forward to action control. Category 3: Input from perception; primary output to decision processes.

Category 1

For Category 1 metacognition, signals from other decision processes provide the main input; Category 1's output is primarily directed at other decision processes, with the intent to modulate or control them (Fig. 2). Examples from cognitive neuroscience and cognitive science provide a clearer view of Category 1 metacognition.

Arbitration of Model-Free vs Model-Based Reinforcement Learning. Decision neuroscience has proposed two distinct types of reinforcement learning in the human brain: model free (MF) and model based (MB), which account for habitual and goal-directed behavior control, respectively (Doya, 1999; Daw et al., 2005). A "arbitrator" here is a meta-control mechanism between MB and MF systems (Lee et al., 2014). An input to the arbitrator is the estimated reliability of each system, which is computed from the average amount of state prediction errors and reward prediction errors, respectively. The arbitrator then determines the amount of influence MB and MF should have. Neural evidence suggests that ventrolateral prefrontal cortex (vlPFC) computes reliability, which in turn results in the model choice probability (P_{MB}). Given P_{MB}, vlPFC chooses the more reliable of MB or MF to directly control human behavior.

Self-Representation. A self-concept is the ability to situate ourselves in the world and reflect on how we act and feel. It is presumed to arise from the lower processes that define ourselves in the first place, that is, from other processes, including cognitive ones. Cognitive neuroscience research has identified a brain region (ventral medial prefrontal cortex) that mediates this self-concept. It activates when we think about ourselves as opposed to others (Gazzaniga, Ivry, & Mangun, 2013). The way a metacognitive process uses a self-representation determines its category. The next example demonstrates its use in Category 1 metacognition.

Reflection and Self Improvement. Having a self-concept enables multiple important abilities, including that of self-reflection, and thus of self-modification and improvement. Such behavior requires that a system assess the quality of its own performance with respect to some standard or benchmark (including its own past behavior). The FORR cognitive architecture, for example, can manipulate both the value it assigns to its problem-solving mechanisms and the order in which it references them (Epstein & Petrovic, 2011). Moreover, it can develop provisional new heuristics, observe the impact they might have had were they to participate in decisions, and gradually incorporate the most reliable of them into its decision making (Epstein, Freuder, & Wallace, 2005).

Self-control. Category 1 self-control is well illustrated by a high-profile functional imaging study in which the brain activity of dieters was compared to that of others. The study found that dieters' higher-level health or weight-loss self-concept based interest modulated their lower-level valuation process in the striatum. In other words: it actually lowered the value for junk food rather than permit a direct competition between the choice outcomes of two higher-level and lower-level behavioral control systems (Hare, Camerer, & Rangel, 2009).

Artificial Category-1 Metacognition. Further evidence for the use of this category of metacognition in artificial systems include Clarion (Coward & Sun, 2004), MIDCA (Cox et al.,

2016), FORR (Epstein, 1994), Sigma (Rosenbloom, Demski & Ustun, 2016) and Soar (Rosenbloom, Laird & Newell, 1988). The key aspect here is how to develop a representational and processing space that is connected to but separate from the *base space* in which direct interaction with the world occurs. The first two of these architectures provide a separate module for metacognition, while the latter three essentially recur on the base space to do this.

Category 2

For Category 1, input to the metacognitive processes is output from other decision processes (or other signals derived during the decision process), and output is sent to decision processes (rather than, e.g., to actuators). In contrast, Category-2 metacognitive processes do more than merely control/modulate lower-level, behavioral-control systems. Category-2 metacognitive processes are themselves behavioral-control/problem-solving systems; they develop problem representations that lead to decision making and action selection. A key feature of Category-2 metacognition is its use of input from other, lower-level decision processes to inform its own decision making. Social cognition is a particularly salient example of Category 2.

Social cognition. A social setting (e.g., a multiagent system) often requires that an agent have a self-concept, a computational model of human decision making that models the agent within a multiagent environment. Such a self-concept considers possible scenarios (i.e., decision options) with respect to potential social interactions, possibly from a game-theoretic perspective. To the extent that the self-concept models other agents the way it models itself, higher-level social cognition also provides an example of metacognitive elements in higher-level cognition. More specifically, extensive research in social psychology and neuroscience has established that people model each other's beliefs, goals, and intentions, and think about their minds much the way they think about themselves (Gazzaniga et al., 2013). Note, however, that all of social cognition is not necessarily meta-level or even high-level; it too shades from simple (e.g., dominance hierarchies) to complex (e.g., theory of mind).

Social constructs may be based on larger social groups, especially *social rules* (from norms and conventions to laws). Whether such rules are meta-cognitive depends on exactly how they are processed by individuals and/or modeled by artificial systems. Again, this can range from simple (e.g., rules as punishment to avoid) to complex (e.g., moral principles) (Gazzaniga et al., 2013).

Finally, much higher-level human cognition requires sophisticated cognitive machinery to coordinate with other cognitive systems. This is especially clear with social cognition, where almost any problem (e.g., organizing meals, working, raising a family) must pass possible solutions through a social filter. The filter determines whether a possible choice remains viable given the interests and dynamics of others. Such interaction across different content domains (e.g., finding food vs. sociability) ultimately requires sophisticated coordination among a set of cognitive processes, and therefore metacognition.

Artificial Category-2 Metacognition. Evidence for Category-2 metacognition in artificial systems includes developed computational models of social cognition (Alechina et al., 2012; Lee, Kralik, & Jeong, 2018a, 2018b; Lee et al., 2014a, 2014b; Pynadath & Marsella, 2005; Pynadath, Rosenbloom & Marsella, 2014) inspired by socio-cognitive theory on human decision making (Bandura, 2001). For example, regarding higher-level societal understanding, N-2APL (Alechina et al., 2012) and N-Jason (Lee et al., 2014a, 2014b) have metacognitive components that enable decision making with social norms. Their cognitive agents can decide whether to follow their individual goals or deontic goals (related to obligation and permission) triggered by social norms. This allows a cognitive agent to be autonomous over social constructs, that is, it can choose the normative goals or abandon compliance with them.

Category 3

Category-3 metacognition includes processes that receive their input primarily from feedforward representations of environmental stimuli (e.g., from perceptual processes), but primarily project to other cognitive processes (Fig. 2). A prominent example from cognitive neuroscience is context and abstract task relevant information. Evidence from cognitive neuroscience shows that higher-order brain regions (e.g., regions in the prefrontal cortex) provide more sophisticated environmental information to basic decision-making systems (Gazzaniga et al., 2013).

Summary

The categories of metacognition outlined here are meant to clarify the broad possibilities of what constitutes metacognition. In practice, the boundaries between categories themselves can become fuzzy and gradated, especially in sufficiently complex computational systems. Nonetheless, this categorization helps clarify the general characteristics of cognitive and metacognitive processes. The next section provides a list of some of the major components of metacognitive decision processes.

4. Components of Metacognition

This section identifies the components of metacognition for which substantial empirical evidence exists (e.g., Gazzaniga et al., 2013). This list (Fig. 3) currently centers on Category-1 metacognition, with input, central focus, and output all as decision processes (see Fig. 2).

Figure 3. Components of Category 1 Metacognition.

Monitoring. Monitoring occurs when a metacognitive process receives input from the cognitive processes it attempts to influence. In the brain, for example, evidence implicates particular brain regions (e.g., medial prefrontal cortex, and in particular, anterior cingulate) involved in monitoring (Gazzaniga et al., 2013).

Evaluation. Once activity from the monitored cognitive systems is received, the metacognitive system must then evaluate it. A particularly strong example of this is evidence that a region of the prefrontal cortex in the human brain arbitrates among candidate behavioral-control systems via an evaluation process that compares their relative likelihood of success (see Category 1 Section above) (Kowaguchi, Patel, Bunnell, & Kralik, 2016; Lee et al., 2014).

Planning. Because evaluation by higher-level metacognitive control systems is relatively sophisticated, evaluation should include an assessment of future success as well as identification of the best action policies to achieve it. Planning systems can be quite complex. For example, they may have goal hierarchies that require dynamic management to use and update them during task completion (Gazzaniga et al., 2013).

Mental Simulation. Similar to and often in conjunction with planning, mental simulation provides the ability to play out imagined possible scenarios before a given action is chosen. Such simulations require relatively rich *mental models* of the problem environment. Consciousness, described in Section 5, also appears to play an important role in forming mental models; it integrates aspects of the present, the past, and the future as part of a correlated scene.

Control. Category-1 metacognitive processes are dedicated to coordinating (or orchestrating) activities of lower-level behavioral-control systems. They include *arbitrating* among systems (i.e., choosing among mutually exclusive ones) and *multitasking,* including such sub-processes as *scheduling* and *task switching*. Control by Category-1 processes is normally expected to either *modulate* or *bias* the behavioral-control systems it addresses.

This list of components is merely a starting point. Other functionality expected to be added includes those related to self-reflection and self-improvement (e.g., understanding, awareness, generating, organizing, maintaining, modifying, debugging, healing, configuring, adapting) (Project CogX, http://cogx.eu/; Lee et al., 2018b; Sampson, Khan, Nisenbaum, & Kralik, 2018). We turn next to perhaps the most quintessentially 'meta' cognition: conscious processing and consciousness.

5. Consciousness

Consciousness involves perceiving, thinking about, and experiencing elements derived from other decision processes (e.g., our concept of 'self'). Thus, consciousness is also a form of metacognition. The 'Hard Problem' (Chalmers, 1995) is explaining the first-person, subjective experience of human consciousness that goes from self-concept to the interpretation of our experiences as sentient. How and why, for example, people are able to experience things like love, the color red, self-doubt (Damasio, 1996; Dennett, 1991; Gazzaniga et al., 2013). To date, there is no philosophical or scientific consensus on this, but there are notable, important developments, which we outline here.

Tononi (2008) and Tononi and Koch (2015) described and refined integrated information theory (IIT) as a theoretical framework to describe and measure consciousness. IIT argues that a theory of consciousness must begin from a set of axioms based on the phenomena to be explained, and then derive a set of postulates from those axioms. Central to IIT is the notion that a proper theory of consciousness must first consider the essential properties of the phenomenon that the conscious being has had, that is, its own experience of the phenomenon. McGreggor (2017) established a theoretical framework that allows such experiences to be considered as proper knowledge representations, a crucial connection between the various theories of consciousness and the analytical techniques of cognitive science and AI.

Jackson (2014) discussed how computers could potentially obtain enough self-awareness to achieve human-level AI by adapting the 'axioms of being conscious' proposed by Aleksander and Morton (2007) for research on artificial consciousness. For a system to approach artificial consciousness, there are a set of metacognitive "observations" it must achieve:

Observation of an external environment.
Observation of itself in relation to the external environment.
Observation of internal thoughts.
Observation of time: the present, the past, and potential futures.
Observation of hypothetical or imaginative thoughts.
Reflective observation: Observation of observations.

To attain these observational abilities, an AI system would need to create and process data structures that represent them. Indeed, there appears to be nothing inherently impossible about creating such data structures. Jackson (2014, p.245) discussed how the potential to support artificial consciousness is illustrated by the TalaMind prototype demonstration system.

Dehaene (2014) described consciousness as "the mind's virtual reality simulator." The functions of consciousness, he argued, are the stable retention of information (as opposed to the fleeting signals of perception), the compression of information to facilitate routing and further processing, and the ability to broadcast information through

language. These capabilities should also be considered in a cognitive model of consciousness.

An interpreter and inner speech are also critical features of human consciousness. The former is a unified cognitive process that mediates the sense of "I" or "me", and the control we believe we have over our decisions (Gazzaniga, 2011). Functionally, this interpreter is an overarching cognitive system that organizes the "findings" of the multiple other lower-level processes to produce one coherent story (and sense of self). This interpreter appears to seek a consistent narrative that makes sense of the world causally, with this story considered as a set of higher-level beliefs. The interpreter can then use these beliefs to manipulate and affect the agent's goal-directed behavior. Consciousness thus appears to have aspects drawn from Types 1, 2 and 3 metacognition — to orchestrate other systems, and to follow its own muse and decision policies. Representation of the interpreter and inner speech in metacognition would support several of the axioms of consciousness: observation of internal thoughts, observation of hypothetical or imaginative thoughts, and reflective observation.

Johnson-Laird (1983, pp. 448-477) discussed how a computational system could approach artificial consciousness. He reasoned that such a system must process in parallel, and that a form of self-awareness could result if the system could recursively represent mental models within mental models and have a higher-level model of its own operating system.

Beyond the emulation of human consciousness, Sanz (2007) offered an alternative perspective to Aleksander and Morton's axioms. Gamez (2017) offered an approach to neutralize philosophical conundrums around consciousness to ground a scientific, measurable theory that can be used in the analysis of consciousness in metacognition for both humans and machines.

Finally, in neuroscience, *affective processing* has been identified as playing a key role in conscious experience (Damasio, 1996; Gazzaniga et al., 2013). In short, the sentience that humans feel derives at least in part from a highly integrated (and likely "resonant loop") signal of both deeply bottom-up body-state signals with those from the highest top-down conceptual understanding. Progression towards a greater understanding will therefore require focus on the processes involved in the large-scale integration of cognitive processing. Metacognition should play a prominent role, since it focuses on the mechanisms of system-wide integration.

6. Design Issues and Case Studies

To help provide a roadmap for constructing models of metacognition, this section presents general design issues, followed by case studies where metacognitive components have been successfully added to working computational models of cognition.

6.1 Design Issues for Metacognition

The homunculus fallacy A persistent issue for metacognition concerns whether to consider metacognition as a central module for executive control. This can give rise to the homunculus fallacy, where a little person inside one's brain reasons like an intelligent being to deal with the situation it observes. Such reasoning leads to an infinite regress: to explain how the homunculus functions, one must assume that it has a mind, which itself implies another homunculus inside it, which must contain yet another homunculus, and so on. One possible solution to this problem is to provide *a priori* constraints on what the highest-level executive system should entail, both in what it can achieve and how it is constructed. From a functional and architectural perspective, it may be best to conceive of metacognition not as a collection of parts but holistically, as a whole distributed over many components. Ultimate understanding is not located in any one of the components, but in the network of their relations or interconnections, and in their inter-processing.

Internal languages of thought. In principle, we can generate arbitrarily long recursive metacognitive processes and their output (e.g., long, embedded sentences), In practice, however, probably only a few combinations require execution in a metacognitive system. One general approach would be to build additional components for the model once sufficient evidence supports them. Another design possibility, however, would make the system itself able to add a metacognitive process when it is needed. This could, for example, be prompted by a decision point relative to an ongoing process. For example, "I don't know how to do X, so I can try to learn how to do X." Systems could also be designed to halt metacognitive processes when they are no longer deemed worthwhile. This might use mechanisms corresponding to an 'economy of mind' (Wright, 2000). In addition, recursively nested mental models (Johnson-Laird, 1983 et seq.), and a 'natural language of thought' (Jackson, 2018) can be tools for representation and implementation of metacognition.

A multi-language approach to an internal language of thought proposed a cognitive hierarchy of logical languages. Other logical language families (e.g., AI's description logics or classical mathematical logic with its propositional logic, first-order and higher-order predicate logics) emphasize cognitive adequacy as measured by time complexity, following Newell (1990). In particular, the continuous domains of perception require high-complexity description logics built on taxonomic reasoning. Decision processes, however, must handle perceptual input much faster than taxonomic inference. Accordingly, a logical hierarchy was developed that is based on a minimalistic language for reasoning about continuous domains equivalent to a

fragment of propositional logic, that is, fast enough for real-time processing at the time scale higher end of perception and the lower end of reasoning (see Schmidtke, 2018). A recent result has shown that, through a strictly logical, ontology-free, and particularly simple reasoning mechanism, this primitive language's formulae give rise to graphical representations analogous to the content expressed in the formulae (Schmidtke, 2018). This result opens new ways to connect the lower end of decision with the higher end of perception. It is also fundamental for higher cognition because it provides a simple mechanism for the construction of mental images from logical representations, which can be employed for the construction or reconstruction of remembered, inferred, or communicated contents. That is, the hierarchy can provide a primitive step that facilitates metacognitive tasks.

Limitations of human cognition Finally, an important design consideration is the extent to which any model of cognition — including metacognition — should mirror human abilities. On the one hand, human cognition provides an important existence proof of some of the highest cognitive and metacognitive abilities known to exist, yet it often makes simple logical errors (Johnson-Laird, 1983; Kahneman, 2011). Thus, there may be a downside to building cognitive systems that fall into the same logical traps as people. Perhaps we should not be guided solely by human models.

6.2 Metacognition in Cognitive Architectures

Computational frameworks for cognition have tackled metacognition from different perspectives depending on the underlying neuropsychological and psychological theories used for their construction. Next, we briefly present three case studies for metacognition: the ACT-R (Anderson, 2009), CLARION (Sun, 2007), and LIDA (Franklin, 2016) cognitive architectures.

While a metacognitive module is not included in the core ACT-R architecture, recent work has developed a metacognitive module (Anderson & Fincham, 2014) that consciously assesses what one knows and how to extend it to solve a problem. This activity is associated with the rostrolateral prefrontal cortex (RLPFC), which has been linked to reflective functions (e.g., prospective memory, reasoning about analogies). The metacognitive module implements this activity by reflecting on declarative representations of cognitive procedures. Use of the metacognitive module was illustrated with Exception and Regular (simpler) mathematical problem solving. Solution of Exception problems required the modification or replacement of elements in procedures for solving Regular problems. The metacognitive module builds a representation of the required elements. The working memory holds the problem representation, while the metacognitive module holds declarative representations of procedures to be modified and "rehearses" the modified procedure.

CLARION's hybrid architecture comprises two representation levels; symbolic and subsymbolic. It is based on Flavell's notion of metacognition as the active monitoring and consequent regulation and orchestration of cognitive processes in relation to the cognitive objects and data of which they bear) (Flavell, 1976). CLARION uses multiple metacognitive criteria to decide when and how to use symbolic or sub-symbolic processing; a particular learning method (e.g., reinforcement, supervised, unsupervised) or combination of them; and a specific reasoning mechanism (e.g., rule-based, similarity-based).

CLARION also provides cross-layer learning mechanisms to synchronize (accommodate and assimilate knowledge) in both symbolic and subsymbolic layers, to enable both top-down and bottom-up learning. Moreover, CLARION includes a variety of metacognitive processes to set parameters (e.g., learning rates, thresholds, temperature in stochastic decision making, action costs); set dynamic goals driven by competition; and set reinforcement functions to measure the agent's degree of satisfaction. CLARION's metacognition thereby depends heavily upon interaction with a motivational subsystem concerned with drives and their interactions.

Unlike CLARION, LIDA does not define a specific module or subsystem for metacognition; instead, metacognition emerges from the interaction of cascading sequences of cognitive cycles corresponding to action-perception loops. Metacognition in LIDA is based on Sloman's classification of levels of control (Sloman, 1999). These include reactive (for agents requiring little flexibility in their action selection), deliberative (higher-level cognitive processes as planning, scheduling and problem solving), and metacognitive levels (monitoring deliberative processes, allocating cognitive resources, and regulating cognitive strategies). Metacognition in LIDA is implemented by a collection of appropriate behavior streams, each with its own metacognitive task. Metacognitive control adds yet another level of flexibility to an agent's decision making, allowing it to function effectively in an even more complex and dynamically changing environmental niche. Additionally, LIDA defines an artificial consciousness mechanism based on the Global Workspace theory, a neuropsychological theory of consciousness and cognition (Baars, 2007). Attention 'codelets' are little processes that bring items of interest to consciousness, gather current information from the workspace, and compete to see which can bring its information to consciousness. The winner's information is broadcast widely throughout the cognitive apparatus. The purpose of the conscious broadcast is to recruit appropriate resources with which to deal with the current situation. Though various types of resources can, theoretically, be recruited, the conscious broadcast is mostly aimed at procedural memory, where it can directly bring to bear the

information in the contents of consciousness so as to affect the next action to be chosen.

7. Discussion and Conclusions

The first steps towards an architecture for metacognition are to develop a common language and outline the main concepts and research across the relevant fields, which the current paper has begun. The next steps should elaborate on every section, and begin piecing them together to construct a consensus model of metacognition. In 1973, Allen Newell challenged scientists to achieve "a science of [humans] adequate in power and commensurate with [their] complexity". The endeavor to include metacognition in a Common Model of Cognition is one way to accept his challenge.

Acknowledgements

Contributions by JDK, JHL and SWL were supported by the ICT R&D program of MSIP/IITP. [2016-0-00563, Research on Adaptive Machine Learning Technology Development for Intelligent Autonomous Digital Companion] and Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-TC1603-06. Contributions by OJR were sponsored by the U.S. Army, Verizon-CMU InMind project. Contributions by PSR were supported by the U.S. Army under contract W911NF-14-D-0005. Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

Alderson-Day, B., Weis, S., McCarthy-Jones, S., Moseley, P., Smailes, D. & Fernyhough, C. (2016). The brain's conversation with itself: neural substrates of dialogic inner speech. *Social Cognitive and Affective Neuroscience*, 2016, 110–120.

Alechina, N., Dastani, M., & Logan, B. (2012). Programming norm-aware agents. *Autonomous Agents and Multiagent Systems (AAMAS '12)*, 1057–64.

Aleksander, I., & Morton, H. (2007). Depictive Architectures for Synthetic Phenomenology. *Artificial Consciousness*, 67-81, ed. Chella, A. and Manzotti, R. Imprint Academic.

Anderson, J. R. (2009). How can the human mind occur in the physical universe? Oxford University Press.

Anderson, J. R., & Fincham, J. M. (2014). Extending problem-solving procedures through reflection. *Cognitive psych.*, *74*, 1-34.

Aroor, A., S. L. Epstein, S.L. & Korpan, R. (2018). Online learning for crowd-sensitive planning. *Proceedings of AAMAS-2018*, Stockholm.

Baars, B. J. (2007). The global workspace theory of consciousness. The Blackwell companion to consciousness, 236-246.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. Annual Review of Psychology, 52(1):1–26.

Blair, G., Bencomo, N., and France, R. B. Models@run.time. Computer, 42(10):22–27, 2009.

Bonasso, R. P., Firby, R. J., Gat, E., Kortenkamp, D., Miller, D. P. & Slack, M. G. (1997_. Experiences with an Architecture for Intelligent Reactive Agent. *Journal of Experimental and Theoretical Artificial Intelligence*, 9, 237-256.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2, 3, 200-219.

Coward, L.A. & Sun, R. (2004). Criteria for an effective theory of consciousness and some preliminary attempts. *Consciousness and Cognition*, *13*, 268-301.

Cox, M. T., Alavi, Z., Dannenhauer, D., Eyorokon, V., Muñoz-Avila, H. and Perlis, D. (2016). MIDCA: A metacognitive, integrated dual-cycle architecture for self-regulated autonomy. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (pp. 3712-3718).

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. Philosophical Transactions of the Royal Society B: Biological Sciences, 351(1346), 1413–1420.

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, vol. 8, no. 12, pp. 1704–1711, 2005.

Dennett, D. C. (1991). *Consciousness Explained.* New York: Little, Brown and Company.

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes our Thoughts.* Penguin Books, NY.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural networks*, vol. 12, no. 7, pp. 961–974, 1999.

Doyle, J. (1983). A Society of Mind – multiple perspectives, reasoned assumptions, and virtual copies. *Proceedings 1983 International Joint Conference on Artificial Intelligence*, 309-314.

Epstein, S. L. (1994). For the Right Reasons: The FORR Architecture for Learning in a Skill Domain. *Cognitive Science* 18 (3): 479-511.

Epstein, S. L., and Petrovic, S. (2011). Learning a Mixture of Search Heuristics. In *Metareasoning: Thinking about thinking*: MIT Press.

Epstein, S. L., E. C. Freuder & M. Wallace (2005). Learning to Support Constraint Programmers. *Computational Intelligence* **21**(4): 337-371.

Fernyhough, C. (2016). *The Voices Within – The History and Science of How We Talk to Ourselves*. Basic Books.

Fernyhough, C. (2017). Talking to ourselves. *Scientific American*, August 2017, 317, 2, 76–79.

J. Flavell, (1976). Metacognitive aspects of problem solving. In: B. Resnick (ed.), The Nature of Intelligence. Erlbaum, NJ.

Franklin, S., Madl, T., Strain, S., Faghihi, U., Dong, D., Kugele, S., Snaider, J., Agrawal, P., Chen, S. (2016). A LIDA cognitive model tutorial. Biologically Inspired Cognitive Architectures.

Gamez, D.. Human and Machine Consciousness. Open Book Publishers, Oxford, 2018.

Gazzaniga, M. S. (2011). Who's in charge? NY: Harper Collins.

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2013). Cognitive neuroscience: the biology of the mind. WW Norton.

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-Control in

Decision making Involves Modulation of the vmPFC Valuation System. Science, 324(5927), 646–648.

Hernández, C., Bermejo-Alonso, J., and Sanz, R. A self-adaptation framework based on functional knowledge for augmented autonomy in robots. Integrated Computer-Aided Engineering, 25:157–172, 2018.

Holyoak, K. J., & Morrison, R. G. (Eds.). 2012. The Oxford Handbook of Thinking and Reasoning. Oxford University Press.

Jackson, P. C. (2014). *Toward Human-Level Artificial Intelligence–Representation and Computation of Meaning in Natural Language.* Ph.D. Thesis, Tilburg Univ., The Netherlands.

Jackson, P. C. (2017). Toward human-level models of minds. *AAAI Fall Symposium Series Tech. Reports*, FS-17-05, 371-375.

Jackson, P. C. (2018). Natural language in the Common Model of Cognition. *BICA 2018 Post-Proceedings*, to appear.

Johnson-Laird, P. N. (1983). *Mental Models – Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press.

Kahneman, D. (2011). *Thinking Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Kowaguchi, M., Patel, N. P., Bunnell, M. E., & Kralik, J. D. (2016). Competitive control of cognition in rhesus monkeys. Cognition, 157, 146–155.

Korpan, R., S. L. Epstein, A. Aroor and G. Dekel 2017. WHY: Natural Explanations from a Robot Navigator. In *Proceedings of AAAI Fall Symposium on Natural Communication for Human-Robot Collaboration*, Arlington, VA.

Kralik, J. D. (2017). Architectural design of mind & brain from an evolutionary perspective. Proc. AAAI 2017 Fall Symposium: A Standard Model of the Mind.

Laird, JE, Lebiere, C, & Rosenbloom, PS. in press. AI Magazine.

Lee, J., Kralik, J. D., & Jeong, J. (2018a). A Sociocognitive-Neuroeconomic Model of Social Information Communication: To Speak Directly or To Gossip. CogSci 2018

Lee, J., Kralik, J. D., & Jeong, J. (2018b). A General Architecture for Social Intelligence in the Human Mind & Brain. To appear in: Procedia Proc. 2018 Fall Symposium: Common Model of Cognition.

Lee, J., Padget, J., Logan, B., Dybalova, D., and Alechina, N. (2014a). Run-Time Norm Compliance in BDI Agents. *The international conference on Autonomous agents and multi-agent systems (AAMAS '14)*. pages 1581-1582.

Lee, J., Padget, J., Logan, B., Dybalova, D., and Alechina, N. (2014b). ***N-Jason***: Run-Time Norm Compliance in AgentSpeak (L). *Engineering Multi-Agent Systems*, pages 367-387.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3), 687–699.

McGreggor, Keith. (2017). Experience is a Knowledge Representation. AAAI Fall Symposium Tech. Reports, FS-17-05.

Minsky, M. L. (1986). *The Society of Mind*. Simon & Schuster.

Nelson, T. O. (Ed.). (1992). Metacognition: Core readings. Needham Heights, MA, US: Allyn & Bacon.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In *Visual Information Processin*g, 283-310, ed. Chase, W. G. Academic Press, New York.

Newell, A. (1990). *Unified Theories of Cognition*. Harvard University Press.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, 14, 11-28.

Ortony, A., Norman, D. A. & Revelle, W., 2005. Affect and Proto-affect in effective functioning. In J. M. Fellous and M. A. Arbib (eds.) *Who Needs Emotions? The Brain Meets the Machine*. New York, NY: Oxford University Press.

Pynadath, D.V. & Marsella, S. C. (2005). PsychSim: Modeling theory of mind with decision-theoretic agents. In *Proceedings of the International Joint Conf. on Artificial Intelligence*, 1181–86.

Pynadath, D. V., Rosenbloom, P. S. & Marsella, S. C. (2014). Reinforcement learning for adaptive Theory of Mind in the Sigma cognitive architecture. *Proceedings of the 7th Annual Conference on Artificial General Intelligence* (pp. 143-154).

Rosenbloom, P. S., Demski, A. & Ustun, V. (2016). The Sigma cognitive architecture and system: Towards functionally elegant grand unification. *J. of Artificial General Intelligence*, *7*, 1-103.

Rosenbloom, P. S., Laird, J. E. & Newell, A. (1988). Meta-levels in Soar. In P. Maes & D. Nardi (Eds.), *Meta-Level Architectures and Reflection* (pp. 227-240). Amsterdam, NL: North Holland.

Sacks, O. (1989). *Seeing Voices - A Journey into the World of the Deaf*. Vintage Books.

Sampson, W. W. L., Khan, S. A., Nisenbaum, E. J., & Kralik, J. D. (2018). Abstraction promotes creative problem-solving in rhesus monkeys. Cognition, 176, 53–64.

Sanz, R., López, I., Rodríguez, M., and Hernández, C.. Principles for consciousness in integrated cognitive control. Neural Networks, 20(9):938–946, 2007.

Schmidtke, H. R. (2018). Logical lateration – a cognitive systems experiment towards a new approach to the grounding problem. Cognitive Systems Research, in press.

Schneider, W. & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, 84, 1-66.

Sloman, A. (1999). What Sort of Architecture is Required for a Human-like Agent? In Foundations of Rational Agency, ed. M. Wooldridge, and A. S. Rao.

R. Sun. 2007. The motivational and metacognitive control in CLARION. In: W. Gray (ed.), Modeling Integrated Cognitive Systems. Oxford University Press, New York.

Tononi G (2008). Consciousness as integrated information: a provisional manifesto. The Biological Bull., 215 (3), pp. 216–242.

Tononi G, Koch C (2015). Consciousness: here, there and everywhere? Philosophical Transactions of the Royal Society of London B: Biological Sciences, 370 (1668), pp. 20140167.

Wright, I. (2000) The society of mind requires an economy of mind. *Proceedings AISB'00 Symposium Starting from Society - the Application of Social Analogies to Computational Systems*, Birmingham, UK: AISB, 113-124.