Extending Cognitive Architectures

Alexei V. Samsonovich

Krasnow Institute for Advanced Study, George Mason University, Fairfax, VA 22030, USA
asamsono@gmu.edu

Abstract. New powerful approach in cognitive modeling and intelligent agent design, known as biologically inspired cognitive architectures (BICA), allows us to create in the near future general-purpose, real-life computational equivalents of the human mind, that can be used for a broad variety of practical applications. As a first step toward this goal, state-of-the-art BICA need to be extended to enable advanced (meta-)cognitive capabilities, including social and emotional intelligence, human-like episodic memory, imagery, self-awareness, teleological capabilities, to name just a few. Recent extensions of mainstream cognitive architectures claim having many of these features. Yet, their implementation remains limited, compared to the human mind. This work analyzes limitations of existing extensions of popular cognitive architectures, identifies specific challenges, and outlines an approach that allows achieving a "critical mass" of a human-level learner.

Keywords: BICA Challenge, human-level AI, learner critical mass, episodic memory, goal generation.

1 Introduction

Emergent new field of BICA[1] research brings together artificial intelligence, cognitive and neural modeling under a new umbrella: the overarching BICA Challenge to create a computational equivalent of the human mind [1, 2]. The challenge calls for an extension of cognitive architectures with new features that should bring them to the human level of cognition and learning. The list of these features includes episodic memory, theory-of-mind, a sense of self, autonomous goal setting, various forms of metacognition, self-regulated and meta-learning, emotional[2] and social intelligence, and more. Many recently extended popular cognitive architectures are claimed to have some or most of these features and capabilities. However, a critical question is whether the level of their implementation and usage is adequate to requirements set

[1] BICA stands for "biologically inspired cognitive architectures". The acronym was coined by DARPA in 2005 as the name of a program intended to develop psychologically and neurobiologically based computational models of human cognition.

[2] While the terms "emotional cognition" and "emotional intelligence" are highly overloaded in the literature with controversial semantics, they are used here generically to refer to cognitive representation and processing of emotions, moods, feelings, affects, appraisals, etc.

A. Chella et al. (Eds.): Biologically Inspired Cognitive Architectures 2012, AISC 196, pp. 41–49.
springerlink.com

by the challenge [2]. The present work addresses this question by examining particular examples, pointing to problems with existing implementations and setting specific challenges for future research.

Since the onset of cognitive modeling as a research paradigm, attempts are made to implement and study complete cognitive agents embedded in virtual or physical environments [3]. Computational frameworks used for designing these agents are known as cognitive architectures [4-8]. A cognitive architecture is considered "biologically inspired" when it is motivated by the organization and principles of biological intelligent systems, primarily, the human brain-mind. From this point of view, the majority of modern cognitive architectures belong to the BICA category. E.g., the most popular cognitive architectures, including ACT-R [9, 10] and Soar [11-14], originated from the Allen Newell's goal to model principles and mechanisms of human cognition [3], as opposed to the original goal of reproducing human intelligent capabilities in artificial intelligence [15] without necessarily replicating their mechanisms, or the goal in neuroscience – to understand how the brain works at the neuronal level.

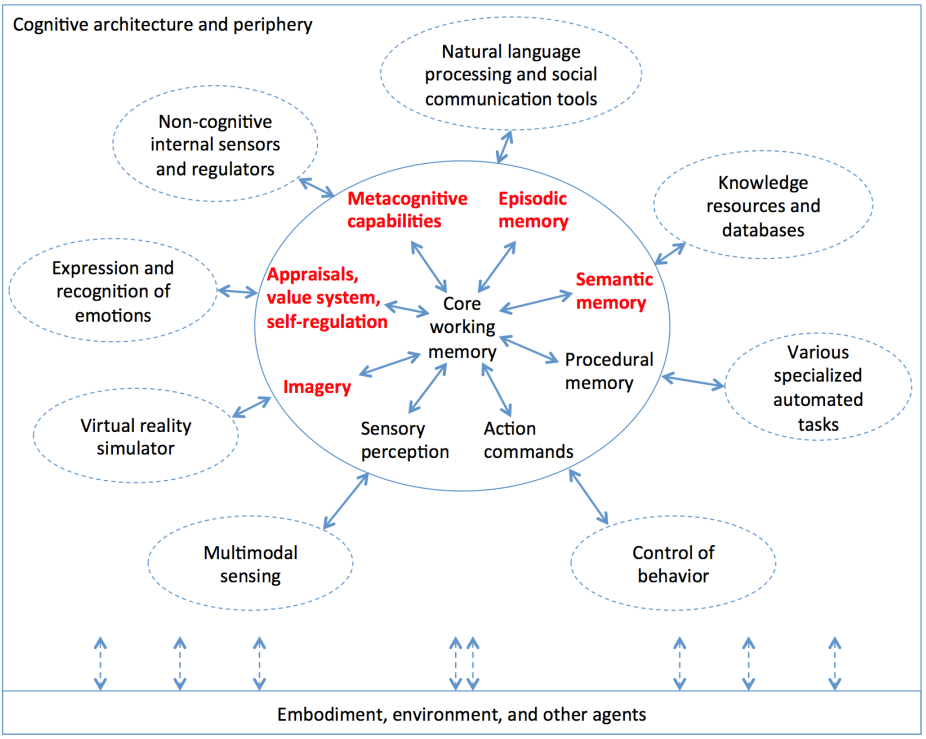

Fig. 1. Template for comparison of extensions of cognitive architectures. Only components within the solid circle belong to the cognitive architecture proper, of which the components shown in bold red are typically considered extensions. Virtually all circles may have direct connections to the environment (dashed vertical arrows).

A certain minimal set of elements including perception, cognition based on stored procedures, and action control, is common for all cognitive architectures due to the requirement of completeness. In these sense, other features can be regarded as extensions. The focus here is on extensions that are necessary for solving the BICA Challenge (Figure 1, red; [2]).

2 State of the Art and Limitations of Extensions: Examples

2.1 Limited Episodic Memory

As an example, let us consider the state of the art of episodic memory implementation and usage in cognitive architectures represented by the extended Soar [13, 14]. Episodic memory in Soar is stored as snapshots of contents of working memory taken together with contextual metadata. In principle, forms of episodic memory in Soar also include prospective memories of intents and plans. Retrieval is possible by activation of cues or by context. Usage may include many functions [14]: e.g., analysis of past episodes and retroactive learning, prediction and guidance in action selection, repetition avoidance.

This broad spectrum of functionality and usage of episodic memory looks much better than early implementations; yet many limitations still remain. Unlike human episodic memory, episodic memory in Soar does not remain plastic after its formation, and it is not modified or replicated every time when it is accessed (cf. [16]). Also, it mostly represents experiences of the actual past situations of the agent. The uniqueness of remembered episodes in many cases appears not critical for their usage: merging similar episodic memories may be allowed, which in psychological terms means mixing the notions of episodic and semantic memory.

The rich system of relations among remembered episodes characteristic of human memory is missing in these implementation, and as a consequence, strategic retrieval mechanisms with step-by-step contextual reinstatement [25, 26] are not implementable. Episodic memory of imagery is also missing, but is on the list in [14].

2.2 Limited Metacognition

Metacognition is a very broad notion, which in various forms is intimately interleaved virtually in all human cognition. It is impossible to address this topic here in detail. First, it is important to separate from metacognition anything that is not cognition on its own: e.g., internal sensing and autoregulation of the computational process (Figure 1, upper-left circle).

Speaking of metacognition as cognition about cognition, of particular interest are functions with known implementations in cognitive architectures, for example, Theory-of-Mind reasoning and autonomous goal generation [17, 18]. One general limitation of these implementations is that reasoning about goals is driven by a persistent meta-goal, and in this sense amounts to a sub-goal reasoning or planning.

Similarly, Theory-of-Mind in artificial intelligence is traditionally understood as a system of beliefs about beliefs of others processed from the same first-person mental perspective of the agent, in contrast with more rich mental simulations of others' mental perspectives performed by humans [20, 21].

2.3 Limited Affective Cognition

Extended Soar [13, 14] implements emotional intelligence by adding an appraisal detector as a separate module, which implements the theory of Scherer. The usage of tis module is that it generates one global characteristic (appraisal) of the current state of the agent, which can be used as a reward signal in reinforcement learning.

Limitations compared to human emotional cognition are innumerous; only a few examples can be named here. First and foremost, this approach does not allow for representation of social (complex) emotions. Secondly, it does not allow for simultaneous processing of multiple appraisals and appraisals of mental states of others.

3 Specific Challenges for BICA Designs: Examples

The above limitations suggest challenges for new cognitive architecture designs. The subset of challenges selected as examples below is not random. They contribute to an emergent coherent story that addresses the critical question of the BICA Challenge [2]: how to achieve the human learner critical mass? One specific approach is outlined in Section 4.

3.1 Plastic Prospective Episodic Memory

Episodic memory in humans is not limited to static snapshots of past experiences: it also stores imagined future or abstract situations, imagined experiences of others, and it is changed every time when it is accessed by mechanisms like reconsolidation and multiple trace formation. Many of these features will be critical for the believability of the agent and for its autonomous cognitive growth up to a human level. For example, remembered dreams of the future may help to generate new goals (see below); re-evaluating the past based on corrected beliefs may improve self-consistency of the agent cognition, and so on.

One specific challenge for future implementations is to have plastic prospective episodic memory of the imagined future scenario, in which not only the plan of achieving the goal, but also the understanding of the goal itself, as well as sub-goals, may change in response to new information in a more natural, human-like way. As a prerequisite for doing this, a significant first step, e.g., in Soar would be getting a goal situation represented in episodic memory as an imagined experience of the agent, thereby giving the agent new reasoning capabilities. This format of goal representation is already the standard for some existing frameworks: e.g., GMU BICA [24].

3.2 Creative Autonomous Goal Generation

The extension of episodic memory discussed above will allow the agent to reason about the goal as a perceived state of the world, questioning own beliefs, applying new knowledge and performing mental simulations in that state. The challenge is then to enable bootstrapped generation of higher and more complex goals that make sense in a given world, starting from a minimal set of innate primitive drives. A successful approach will combine many cognitive capabilities discussed in this article.

With multiple potential goal situations represented as plastic prospective episodic memories that are subject to metacognition, the agent will have possibilities of engineering and selection of goals. This process also requires metacognitive reasoning about goals. Processes of goal reasoning and goal selection can be automated using various approaches [17, 27], and in general rely on a system of values.

3.3 Human-Level Emotional Intelligence

Thus, a system of values appears necessary for the agent to be able to generate new goals. In order to be human-like, the goal selection process in an agent must be guided by a human-level system of values. This is only one aspect of the challenge of achieving human-level emotional intelligence in artifacts. Another aspect is the necessity for an agent to be integrated with human partners in a team, implying the ability to develop mutual relationships of trust, respect, subordination, etc. Complex social emotions are inevitably involved in the formation and maintenance of such relationships, which means that artifacts should be able to understand, generate, recognize and express social emotions. The "understand" part appears to be the hardest at present.

3.4 Human Learner Critical Mass

The human learner critical mass challenge is to identify, design and create a minimal cognitive architecture with minimal initial knowledge and skills, capable of human-like learning up to a level that a typical human can achieve in similar settings. The learning process does not need to be autonomous and may involve human instruction, mentoring, scaffolding, learning from examples, interaction with resources of various nature available to humans. The challenge can be further divided into a set of domain-specific human learner critical mass challenges and a general-purpose human learner critical mass challenge, the former being a precondition for the latter. The hypothesis is that a solution can be found in a simple form, as opposed to manual implementation and integration of all human intelligent capabilities. In this case, interactive learning, self-organization and evolution is expected to be among the main techniques used for creation of intelligent agents.

4 Toward Human-Like Intelligent Artifacts

The above consideration suggests that the challenge of cognitive growth of an artifact up to a human level of general intelligence (the human learner critical mass challenge)

impinges on the evolvability of goals and values in an artificial cognitive system, which in turn requires artificial emotional intelligence. This section introduces a new approach to addressing the challenge, based on developing emotional intelligence in artifacts [22] and using it together with potential goal enumeration in order to generate a system of goals.

4.1 Emotional Extension of the Mental State Framework

The mental state framework [19] essentially relies on two building blocks: a mental state, that attributes specific content of awareness to a specific mental perspective of an agent, and a "schema": the term in this case refers to a specific structure that can be used to represent any concept or category. Instances of schemas populate mental states. Each instance has a standard set of attributes [24].

"Emotional" extension of this framework is based on the introduction of three new elements [22]: (i) an emotional state (an attribute of a mental state), (ii) an appraisal (an attribute of a schema), and (iii) higher-order appraisal schemas, that can be also called "moral schemas". As the name suggests, these schemas recognize patterns of appraisals and emotional states, and are intended to represent complex or social emotions and relationships, including pride, shame, trust, resentment, compassion, jealousy, sense of humor, etc. Available phenomenological data [28] can be used to define these schemas – or to map naturally emerging in the architecture new schemas onto familiar concepts.

4.2 Enumeration of Potential Goals

A useful enumeration of possible, virtually relevant, or potential goals in a given world or situation could be the key to goal generation. In order to enumerate possible goals in a useful way, one can use a semantic metalanguage [23, 29]: specifically, the shared by all languages lexical-conceptual core of semantic primes and their associated grammar. Examples of semantic primitives include very basic notions like "above", "big", "more", "have", "inside", "move", "see", "want", etc. From these fundamental notions, generic goal-like notions can be formed, e.g.: "survive", "satisfy desire", "possess", "secure", "dominate", "have freedom", "explore", etc. that can be applied to specific objects and situations in various combinations. Therefore, in a given setup, a conceptual lattice [31] of potential goals can be generated using the fundamental primitives. Then, classification and selection among potential goals should be done using a system of values organized in a Maslow hierarchy [30].

4.3 Generation of New Values

Thus, goal selection guided by the system of values requires a human-like system of values, and its natural development depends on the agent's ability to generate new values, which can be done by moral schemas, which therefore play the role of a "critical element" of the critical mass of a human-level learner. Moral schemas can be

innate or emergent. Future studies will estimate this component of the critical mass in terms of a minimal subset of moral schemas that enable autonomous development of a human-compatible system of values and goals in a given environment.

5 Discussion

This paper presented a brief overview of cognitive architecture extensions with advanced, human-inspired cognitive capabilities, and pointed to the wide gap between existing implementations and the human mind. Several examples of specific challenges in bridging the gap were outlined, that allow us to decompose the BICA Challenge. Possible approaches to solving some of these challenges were discussed.

The key question is, which of these biologically inspired advanced features are critical, and which may be optional? "Critical" here means critical for acceptance as cognitively "equal" minds by humans, and for achieving a human-level learner critical mass. The analysis of the latter challenge presented here suggests that the critical set should include the above examples described as specific challenges, and more. Specifically, human-level emotional intelligence appears to be a necessary feature for the agent believability and for the sense of co-presence associated with the agent. It is also an essential component in self-regulated learning, which is one of the mechanisms required for achieving the critical mass: this aspect will be discussed elsewhere. As the consideration presented here illustrates, a general approach to solving the outlined challenges can be based on the formalism of multiple mental states simultaneously present in working memory [19], which therefore appears to be promising.

5.1 Conclusions

As a first step toward solving the BICA Challenge, state-of-the-art BICA need to be extended to enable advanced cognitive capabilities, including emotional intelligence, human-like episodic memory, and the ability to generate new goals and values. Existing extensions of mainstream cognitive architectures remain limited, compared to the human mind. Based on the analysis of these limitations and challenges, it is found here that human-level emotional intelligence is a critical component in the human-level learner critical mass. A specific approach to achieving the critical mass outlined here implies that more complex goals can be generated automatically based on an enumerated set of potential goals generated using universal cognitive elements like semantic primes, and the rules of goal selection can be based on an evolving system of values generated by moral schemas. These findings suggest tasks for future research.

References

1. Chella, A., Lebiere, C., Noelle, D.C., Samsonovich, A.V.: On a roadmap to biologically inspired cognitive agents. In: Samsonovich, A.V., Johannsdottir, K.R. (eds.) Biologically Inspired Cognitive Architectures 2011: Proceedings of the Second Annual Meeting of the BICA Society. Frontiers in Artificial Intelligence and Applications, vol. 233, pp. 453–460. IOS Press, Amsterdam (2011)

2. Samsonovich, A.V.: On the roadmap for the BICA Challenge. Biologically Inspired Cognitive Architectures 1(1), 100–107 (2012)
3. Newell, A.: Unified Theories of Cognition. Harward University Press, Cambridge (1990)
4. SIGArt, Special section on integrated cognitive architectures. Sigart Bulletin 2(4) (1991)
5. Pew, R.W., Mavor, A.S. (eds.): Modeling Human and Organizational Behavior: Application to Military Simulations. National Academy Press, Washington, DC (1998), http://books.nap.edu/catalog/6173.html
6. Ritter, F.E., Shadbolt, N.R., Elliman, D., Young, R.M., Gobet, F., Baxter, G.D.: Techniques for Modeling Human Performance in Synthetic Environments: A Supplementary Review. Human Systems Information Analysis Center (HSIAC), Wright-Patterson Air Force Base (2003)
7. Gluck, K.A., Pew, R.W. (eds.): Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation. Erlbaum, Mahwah (2005)
8. Gray, W.D. (ed.): Integrated Models of Cognitive Systems. Series on Cognitive Models and Architectures. Oxford University Press, Oxford (2007)
9. Anderson, J.R., Lebiere, C.: The Atomic Components of Thought. Lawrence Erlbaum Associates, Mahwah (1998)
10. Anderson, J.R.: How Can the Human Mind Occur in the Physical Universe? Oxford University Press, New York (2007)
11. Laird, J.E., Rosenbloom, P.S., Newell, A.: Universal Subgoaling and Chunking: The Automatic Generation and Learning of Goal Hierarchies. Kluwer, Boston (1986)
12. Laird, J.E., Newell, A., Rosenbloom, P.S.: SOAR: An architecture for general intelligence. Artificial Intelligence 33, 1–64 (1987)
13. Laird, J.E.: Extending the Soar cognitive architecture. In: Wang, P., Goertzel, B., Franklin, S. (eds.) Artificial General Intelligence 2008: Proceedings of the First AGI Conference, pp. 224–235. IOS Press, Amsterdam (2008)
14. Laird, J.E.: The Soar Cognitive Architecture. MIT Press, Cambridge (2012)
15. McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E.: A proposal for the Dartmouth summer research project on artificial intelligence. In: Chrisley, R., Begeer, S. (eds.) Artificial Intelligence: Critical Concepts, vol. 2, pp. 44–53. Routledge, London (1955)
16. Nadel, L., Samsonovich, A., Ryan, L., Moscovitch, M.: Multiple trace theory of human memory: Computational, neuroimaging, and neuropsychological results. Hippocampus 10(4), 352–368 (2000)
17. Molineaux, M., Klenk, M., Aha, D.W.: Goal-driven autonomy in a Navy strategy simulation. In: Proceedings of the National Conference on Artificial Intelligence, vol. 3, pp. 1548–1554 (2010)
18. Hiatt, L.M., Khemlani, S.S., Trafton, J.G.: An explanatory reasoning framework for embodied agents. Biologically Inspired Cognitive Architectures 1, 23–31 (2012)
19. Samsonovich, A.V., De Jong, K.A., Kitsantas, A.: The mental state formalism of GMU-BICA. International Journal of Machine Consciousness 1(1), 111–130 (2009)
20. Gallese, V., Goldman, A.: Mirror neurons and the simulation theory of mind-reading. Trends in Cognitive Science 2, 493–501 (1998)
21. Nichols, S., Stich, S.: Mindreading: An Intergrated Account of Pretence, Self-Awareness, and Understanding Other Minds. Oxford University Press, Oxford (2003)
22. Samsonovich, A.V.: An approach to building emotional intelligence in artifacts. In: Burgard, W., Konolige, K., Pagnucco, M., Vassos, S. (eds.) Cognitive Robotics: AAAI Technical Report WS-12-06, pp. 109–116. The AAAI Press, Menlo Park (2012)
23. Wierzbicka, A.: Semantic complexity: conceptual primitives and the principle of substitutability. Theoretical Linguistics 17(1-3), 75–97 (1991)

24. Samsonovich, A.V., De Jong, K.A.: Designing a self-aware neuromorphic hybrid. In: Thorisson, K.R., Vilhjalmsson, H., Marsela, S. (eds.) AAAI 2005 Workshop on Modular Construction of Human-Like Intelligence: AAAI Technical Report, pp. 71–78. AAAI Press, Menlo Park (2005)
25. Becker, S., Lim, J.: A computational model of prefrontal control in free recall: Strategic memory use in the California Verbal Learning task. Journal of Cognitive Neuroscience 15, 821–832 (2003)
26. Fletcher, P.C., Henson, R.N.A.: Frontal lobes and human memory—Insights from functional neuroimaging. Brain 124, 849–881 (2001)
27. Jaidee, U., Muñoz-Avila, H., Aha, D.W.: Integrated learning for goal-driven autonomy. In: Proceedings of IJCAI 2011, pp. 2450–2455 (2011)
28. Ortony, A., Clore, G., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge (1988)
29. Goddard, C.: Semantic primes, semantic molecules, semantic templates: Key concepts in the NSM approach to lexical typology. Linguistics 50(3), 711–743 (2012)
30. Maslow, A.H.: A theory of human motivation. Psychological Review 50(4), 370–396 (1943)
31. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Berlin (1999)