Dual-Process consideration for decision making and multi-agent architecture of AICA

Bryan Fruchart and Benoit Le Blanc

ENSC, IMS laboratory, Bordeaux INP, CNRS, Talence, France
bfruchart@ensc.fr

Abstract. In this position paper we briefly present research in cognitive architectures and results of the Dual-Process Theory as valuable inputs for AICA decision making specifications. We introduce the proposition of an IPSEL cognitive architecture that integrates Dual-Processes and Constructivism theory of psychology. Considering previous works done for AICA decision making abilities, we also provide a variation of the IPSEL architecture for a multi-agent system.

Keywords: Cognitive Architecture, Dual-Process Theory, Multi-Agent System, Artificial Decision Making.

1 Introduction

Autonomous Intelligent Cyber-Defense Agent is required to monitor and defend a perimeter of a host systems. It should detect signs of cyber-attacks, devises plan for countermeasures in real-time, executes tactically such plans in real-time and reports on their doings to human operators. [Kot18] enumerates three different ways of implementing the desired AICA agent: (i) A society of specialized agents, (ii) a multi-agent system and (iii) an autonomous collaborative agent. In the case of a *society of specialized agents*, the question is to define what would be the cognitive architecture of each specialization; alternatively, if only one typical cognitive architecture is used, the question becomes to differentiate each specialization as agent's functions. For the two other cases (a *multi-agent system* or an *autonomous collaborative agent*), we suppose that implementation used a unique cognitive architecture. Nonetheless, in all cases, it seems that the whole system will be led to be a kind of multi-agent implementation; each part of the system coding its own decision process and its own capacities to communicate.

Cognitive architecture deals with specifying the organizational principles of artificial agent architecture. It aims to provide a mean to the relation between the structure and functions of complex systems designed to achieve human-level capacities of information processes. This kind of system can be considered as Artificial General Intelligence (AGI), even if it is rarely explicitly formulated by cognitive architects. AGI term is sometimes associated with popular and science-fiction representations, making its use hazardous. By AGI, we mean an artificial agent capable of performing information processing functions that are not domain-specific, hence endorsing the ability to adapt to complexes and real-life environments. An AGI system does not need to be an artificial Human as some ideologist communities portray it. AICA research should investigate AGI communities' works to theorize and program artificial agent displaying behaviors that cyber-defense requires. It concerns the capacity to adapt to unseen situations, represent a world model, perform skillful sensory-motor actions, elaborate plan and implement it while communicating and arguing its logic.

Cognitive architects have always been inspired by Human cognition. One of the most prominent psychologic theory that is discussed by systems engineer is the *Dual Process theory*. Initially formulated by one of the fathers of psychology, William James, the theory proposes a paradigm in which Human psyche is divided into two sub-systems. One operates with automatisms, without conscious control and manipulates implicit knowledge; the other operates in a controlled manner using logical rules applied on explicit knowledge. We will use the terminology of *system 1*, for the automatic subsystem, and *system 2* for the logical one. This terminology has been proposed by Richard West and Keith Stanovich [Sta00] and popularized by the recent work of Daniel Kahneman [Kah03]. This system 1 and 2 considerations have gained much attention from the AI community during the last years. For example, Daniel Kahneman has been invited to talk with AI actors [Aaa20] [Lex20], some researchers proposed research perspectives to advance AI from this inspiration [Boo20] while others discuss the analogy between man and machine systems 1 and 2 [Bon20]. These efforts are beneficial, but the investigation should not be confined to Daniel Kahneman works. His contribution concerns many empirical observations conducted under the scope of decision making in an economic context. The Dual Process theory model he has popularized come from previous studies (consider [Eps94]) that may present other details that are not restricted to economic decision making. For a complete review of Dual-Process Theories see [Gaw13].

The basic analogy between Dual Process Theory and machine processing is usually to associate system 1 processes with Machine Learning and System 2 processes with logic-based systems. This comparison sounds natural and obvious but should not be considered as unflawed; many questions remain [Boo20]. Nonetheless, the cooperation between Machine Learning and rules-based programs is a popular direction taken by engineers to design complex systems, such as AlphaGo. This kind of connectionist cross symbolic systems are sometimes named Hybrid systems, and many cognitive architectures are based on this point. Some of them are explicitly proposed to integrate the Dual Process Theory like the CLARION architecture from Ron Sun [Sun03]. For a review and comparison of existing cognitive architectures, see [Kot16-1] [Kot16-2].

In this position paper, we present a cognitive architecture we have created as a design specification for AICA cognitive functions. The model is named IPSEL for *Information processing system with emerging logics*. It shares assumptions with other architectures while differing on other points. It integrates the Dual Process Theory by explicitly naming system 1 and 2 as entities of the model. The originality of this approach is to let the system elaborates its own logical rules from its “experience”. We then propose a multi-agent specification of the IPSEL model that could be used as an entry point for discussing multi-agent or individual autonomous agent implementations.

2 IPSEL Cognitive Architecture

IPSEL is a model [Fru20] based on the psychologic theories of Dual Process [Kah03] [Jam84] and Constructivism [Pia25]. It aims to describe the functional components of an information processing system capable of cognitive behaviors at the Human-level. It could be easily viewed as a Hybrid or Neuro-Symbolic proposition, but IPSEL theory is not limited to one technologic implementation. A fully integrated variation of artificial neural network or a complete expert system may also be paths to explore. Overall, IPSEL is about the structural distribution of functional capacities.

The main characteristics of IPSEL model shared with some other architectures are described in this paragraph. IPSEL distinguishes a fast, massively parallel, sub-symbolic set of processes that are called and executed as automatisms. This set of processes is System 1 named *Intuitive System* and manipulates implicit knowledge using an associative memory named *common-sense* in IPSEL model. The associative memory stores association between perception, system state and actions. It is also distinguishing a slow, sequential, symbolic set of processes that are executed using logical rules. This set of processes is System 2 named *Declarative System* and manipulates explicit knowledge using rules depicted in a memory representing *conceptual entities* and their relations as a world model. It also includes a self-generated evaluation signal which aims to monitor the system's states concerning its integrity and objectives. This evaluation signal is called *Emotional Responses* in IPSEL model.

Assumptions taken by the IPSEL proposition which differ from other cognitive architecture are resume as follow. All high cognitive functions that imply explicit knowledge (argued decision making, planification, deliberated communication, the theory of the mind, introspection, symbolic computation) are supported by a unique mechanism: the production of conceptual speeches. Explicit pieces of knowledge describing the world model are stored in a memory which takes the form of a graph of concepts connected by weighted edges representing probabilistic rules of their causality relations. Producing conceptual speeches consists to starts with an activated concept and then finds the following concepts with respect to the probabilities and constraints that apply on the graph. Speech production is finding a path in the *space of concepts* that represent explicit knowledge.

The space of concepts and probabilistic causal rules that connect them is not described *a priori* during the conception but is an emerging construction of the system. Concepts are constructed from reinforcement of recurrent combinations of simultaneously activated associative memory patterns, hence grounding them in sensory, system's state and motor correspondences. Syntax rules are constructed and calibrated from reinforcement of recurrent order of concepts appearances. These inductive knowledge construction mechanisms are seen as the system 2 emergence from system 1 activities in a bottom-up fashion. While system 2 produces conceptual speeches, concepts are activated in a specific order. Since these concepts have connections in terms of sensor/state/motor associations from which they have been constructed, the newly produced conceptual speech act as a new source of information for system 1 in a top-down fashion. The challenge for Human-Machine communication is not for the machine to correctly use human concept but is for humans to understand what machine's concepts mean.

Figure 1 summarizes IPSEL's organisation for information process. Perceptive messages from the external feed both a *direct system* (to process reflexes and activate direct actions) and an *intuitive system* (the presented above S1). *Intuitive system* (S1) and *Deliberative system* (S2) constitute a processing loop in which *conceptual sequences* are in a central place. For more details see [Fru20].

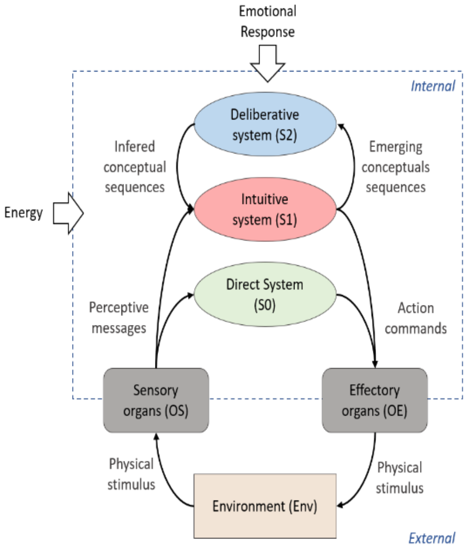

Fig. 1. IPSEL model.

3 IPSEL Multi-Agent Architecture for AICA

The Dual Process theory's central precept is to distinguish between two sets of processes, the automatic ones operating on implicit knowledge and the logical one operating on explicit knowledge. In an approach of a society of specialized agents (case (i) exposed in introduction), it would require distinguishing between two types of specializations with a common communication process between those two types to achieve system 1 and 2 collaboration. For the two other approaches explained in introduction (cases (ii) for a multi-agent system and (iii) for an autonomous collaborative agent), agents are required to perform all the AICA's functions and therefore, from a dual-process theory point of view, having both system 1 and 2 integrated.

As the dual-process theorists have suggested, systems 2 functions are slow to process and consume a vast amount of energy which can engender severe limitation for numerous agents' efficient cooperation. From IPSEL theory, we also add that systems 2 manipulate explicit pieces of knowledge that are constructed from agent experience and are subjective. Agent collaboration at the system 2 level would require an initial phase of knowledge and representations alignment.

This part advocates that a hybrid solution composed of a multi-agent swarm having automatic coordinated behaviours and taking orders from a holistic master agent should be explored. The difference with approaches cited earlier is that agents of the swarm would only be system 1 based, all being able to communicate as automatic behaviours with one another and with a higher-level module responsible for system 2 functions.

In this proposition, AICA micro agents would be autonomous software applications operating at various positions of the environment. They perform different tasks of information gathering, basics security checks and actions. These processes are referred as (A) and (B) on the figure 2. They may all have a learning kernel to improve their operations successes or download up-to-date models from a unique learning centre. They can also do direct communication with each other to coordinate operations that require multiple actions. This sort of communication is referred as (C) on the figure 2 and allows intelligent collective behaviours. An example of multi-agent capacities emerging from inductive and reinforcement learning on an adversarial context can be found in [Bak19].

Altogether, theses micro agents form the body of the whole system that communicate relevant pieces of information concerning their operation to a higher-level structure: the AICA macro agent, viewed as the "mind" of the system.

The AICA macro agent gathers monitoring pieces of information about the environment states. Having its own learning capacities based on logical inferences; it can discover causal relations between situations occurring on different parts of the controlled environment. Its use of rules allows it to perform deliberative functions such as argued decision making or planification. It is not directly related to the environment, which is only modified by micro agents that can keep operating while the macro agent is "thinking".

This conception integrates the Dual-Process theory on a multi-agent perspective. Micro agents are system 1 based agents operating in an autonomous and parallel fashion or as an automatic response to the macro agent commands. The macro agent has system 2 based processes and uses a world model filled with micro agents information. If human intervention is needed; the communication between operators and the system is handled by exchanges with the macro agent ((G) and (F) on the figure 2). By this way, only the part of the system with a full understanding of the environment is communicating with humans. Also, parts of the system that perform actions on the environment are isolated from the human, allowing them to keep the advantage of fast computing and autonomy. The whole mechanism mimics the way that humans collaborate. Each one have the full control of his own muscles. If someone wants that another people performs an action, he does not address the direct command to his colleague’s muscles but sends to his colleague’s mind a message and let him accomplishes the action if he well understand the message.

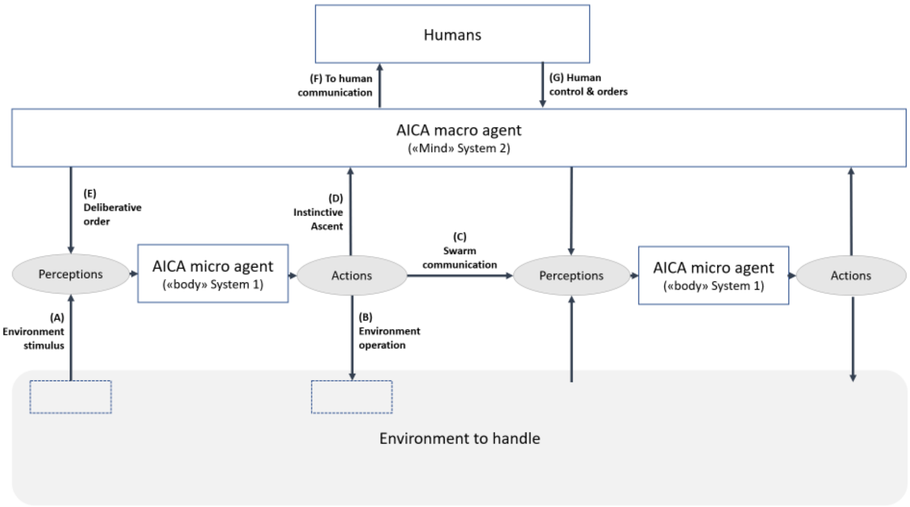

Fig. 2. An illustration of a multi-agent architecture integrating the Dual-Process Theory.

4 Conclusion

Capacities required for an AICA are similar of those one would expect from an AGI system. To design such a system, it is necessary to think about its global cognitive architecture. Many architectures have been proposed but only some of them clearly integrate a Dual-Process theory approach whereas its consideration is gaining lot of attention from the AI community. As an example of a Dual Process Theory cognitive architecture, we briefly introduce IPSEL model and reference resources to explore more deeply this paradigm.

AICA implementation may be considered as one autonomous agent or as a multi-agent system. As a starting point for discussing the mode of implementation we have presented a multi-agent specification of the IPSEL architecture. It differs with other Dual-Process cognitive architecture by the fact that the swarm of agent is considered as only system 1 based agents while other proposition considers all the agent of the swarm capable of performing system 1 and 2 processes. This proposition can also fit the unique autonomous agent approach considering that the swarm of system 1 based agents is its “body” and system 2 based master program its “mind”.

References

1. [Aaa20] AAAI-20 Talk with Daniel Kahneman, Geoffrey Hinton, Yann LeCun and Yoshua Bengio. https://vimeo.com/390814190, last accessed 2021/01/14.

2. [Bak19] Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2019). Emergent tool use from multi-agent autocurricula. arXiv preprint arXiv:1909.07528.
3. [Bon20] Bonnefon, J. F., & Rahwan, I. (2020). Machine Thinking, Fast and Slow. Trends in Cognitive Sciences.
4. [Boo20] Booch, G., Fabiano, F., Horesh, L., Kate, K., Lenchner, J., Linck, N., ... & Srivastava, B. (2020). Thinking fast and slow in ai. arXiv preprint arXiv:2010.06002.
5. [Eps94] Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. American Psychologist , 49 , 709–724 .
6. [Fru20] Fruchart, B., & Le Blanc, B. (2020, September). Cognitive Machinery and Behaviours. In International Conference on Artificial General Intelligence (pp. 121-130). Springer, Cham.
7. [Gaw13] Gawronski, B., & Creighton, L. A. (2013). Dual process theories.
8. [Jam84] James, W. (1984). Psychology, briefer course (Vol. 14). Harvard University Press.
9. [Kah03] Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. American psychologist, 58(9), 697.
10. [Kot16-1] Kotseruba, I., & Tsotsos, J. K. (2016). A review of 40 years of cognitive architecture research: Core cognitive abilities and practical applications. arXiv preprint arXiv:1610.08602.
11. [Kot16-2] Kotseruba, I., Gonzalez, O. J. A., & Tsotsos, J. K. (2016). A review of 40 years of cognitive architecture research: Focus on perception, attention, learning and applications. arXiv preprint arXiv:1610.08602, 1-74.
12. [Kot18] Kott, A., Théron, P., Drašar, M., Dushku, E., LeBlanc, B., Losiewicz, P., ... & Rzadca, K. (2018). Autonomous Intelligent Cyber-defense Agent (AICA) Reference Architecture. Release 2.0. arXiv preprint arXiv:1803.10664.
13. [Lex20] Lex Fridman AI podcast #65 with Daniel Kahneman 2020. https://youtu.be/UwwBG-MbniY, last accessed 2021/01/14.
14. [Pia25] Piaget, J., Rousseau, J., Piaget, M., Deslex, M., & Claparéde, E. (1925). Le langage et la pensée chez l'enfant.
15. [Sta00] Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate?. Behavioral and brain sciences, 23(5), 645-665.
16. [Sun03] R. Sun, (2003). A Tutorial on CLARION. Technical report, Cognitive Science Department, Rensselaer Polytechnic Institute.