The Observable Mind: Enabling an Autonomous Agent Sharing Its Conscious Contents Using a Cognitive Architecture

Daqi Dong
The University of Memphis, Memphis, TN, USA
daqi.dong@gmail.com

Abstract

We enable an autonomous agent sharing its artificial mind to its audiences like humans. This supports the autonomous human robot interactions relying on a cognitive architecture, LIDA, which explains and predicts how minds work and is used as the controllers of intelligent autonomous agents.

We argue that LIDA's cognitive representations and processes may serve as the source of the mind content its agent shares out, autonomously.

We proposed a new description (sub) model into LIDA, letting its agent describing its conscious contents. Through this description, the agent's mind is more observable so we can understand the agent's entity and intelligence more directly. Also, this helps the agent explains its behaviors to its audiences so engage into its living society better.

We built a LIDA software agent embedding with this description model. The agent shares its conscious content autonomously, reasonably explaining its behaviors.

Introduction

We humans are recognized by our appearance, behaviors, and maybe most importantly, minds. The minds decide what we do next and form who we are eventually. Thus, proactively sharing about our minds to others is an effective way helping other people understanding who we are and our behaviors.

We propose to give an autonomous agent (Franklin & Graesser, 1997) the similar ability, sharing its own control-structure, a kind[1] of artificial mind (Franklin, 1995), to others, such as humans or other agents. As Franklin and Graesser defined (1997), an autonomous agent is a system situated within an environment where it interacts with the world and other agents, in pursuit of its own agenda overtime, and affecting what it senses in future. The capability of sharing the mind helps the agent pursue its agenda through the communications.

Toward this sharing mind modeling work, we present a description cognitive model that supports the agent to describe its conscious contents, grounding upon a systems-level cognitive model LIDA (Learning Intelligent Decision Agent) (Franklin et al., 2016).

The LIDA model hypothesizes and predicts how minds work. It provides an architecture integrating multiple cognitive modules, and each of which has different cognitive representations and processes. We argue that these cognitive components may naturally serve as the source of the mind content an agent shares out.

We apply a LIDA model into an agent to control it, as its artificial mind. Borrowed from Global Workspace Theory (GWT) (Baars, 1988, 2002), a LIDA-based agent's conscious content is (1) formed from its understanding of the situation, of both internal and external, (2) chosen as the most salient attention content, and (3) further used in action and learning parts (Franklin et al., 2016).

We can infer both what the agent's mind is and how it works from its conscious. Knowing the conscious content helps determine what (who) the agent is from its inside.

Also, giving an agent the ability to describe its conscious allows it to illustrate what its attentions were, from where these attentions come, and why it acted certain behaviors. This helps the agent engage its audiences and earn more understanding from them (Chin-Parker & Bradner, 2010; Lombrozo, 2006; Matarese, Rea, & Sciutti, 2021), so engage to its living society better (Umbrico et al., 2022).

Cognitive HRI and Cognitive Architectures

From a recent survey of the cognitive HRI, "[it] is a research area that seeks to improve interactions between robots and their users by developing cognitive models for robots and understanding human mental models of robots." (Mutlu, Roy, & Šabanović, 2016).

"[A] systems-level [cognitive] model (cognitive architecture) attempts the full range of activities from incoming stimuli to outgoing actions, together with the full range of cognitive processes in between." (Franklin et al., 2016). It models not only some separate functions of cognition, but also the relationships between them. The necessity of systems-level cognitive modeling has been argued from different disciplines such as artificial intelligence (Newell, 1973), cognitive modeling (Langley, Laird, & Rogers, 2009), and neuroscience (Bullock, 1993).

ACT-R is a cognitive architecture, providing a theory for simulating and understanding human cognition. ACT-R was applied in building an autonomous agent as a human-like collaborator, providing a more efficient interface to the HRI tasks (Sofge et al., 2004). Also, a story-telling social robot was built to represent the story characters, through the definition of appropriate cognitive models replying on the ACT-R (Bono et al., 2020).

Soar is a cognitive architecture developing functional capabilities to tasks such as natural language processing, control of intelligent agents in simulations, virtual humans, and embodied robots (Laird, Lebiere, & Rosenbloom, 2017) and it had been regularly used in robot interactive task learning (Laird, Gluck, et al., 2017). Soar was used to build a robot that leverages humans' natural teaching skills by understanding her teaching intentions in HRI (Ramaraj, Klenk, & Mohan, 2020). Also, an interactive system was built within the Soar which provides the grounding language supporting interactive tasks (Lindes et al., 2017), where a common model of cognition and humanlike language processing had been introduced (Lindes, 2018).

LIDA and Its Agent

As a systems-level cognitive model, LIDA attempts to model minds, which is taken to be the control structures for autonomous agent (Franklin & Graesser, 1997). Regarding the robot models of interaction, Khayi and Franklin (2018) have proposed and implemented a perceptual learning mechanism, to simulate how an infant vervet monkey learns the meanings of vervet monkey alarm calls.

The LIDA model has integrated three phases: perception and understanding, attention, and action and learning (Fig. 1). These phases are functioning continually in a cognitive cycle (~10 Hz) and may (partially) overlap among multiple cycles (Madl, Baars, & Franklin, 2011).

The fundamental data type in each of LIDA modules is a digraph consisting of nodes and links. More complex structures are built from these (Franklin et al., 2016). We may represent objects using nodes and the relationships between them the links. Each of these entities has activation variables attached to it to represent its salience.

A LIDA agent may sense a thirsty feeling and a glass of water on the table, so she chooses to reach out to consume the water. The agent keeps both a *thirsty* and a *water* node structures in its Current Situational Model (CSM) (Fig. 1). The CSM continually updates itself so keep tracks and represents the agent's current situation, by taking the sensations from Sensory Memory, percepts the Perception Associative Memory (PAM), and local associations the other long term memories. Then one of its Attention Codelets chooses these two structures and forms them into a *coalition* sending to the Global Workspace (GW) to compete to be the conscious content. (Franklin et al., 2016).

The conscious content recruits some relevant schemes from the Procedural Memory and instantiates them to the behaviors. Through the Action Selection module, one behavior such as the *grasp* may be selected based on its activation, which will be used to choose a Motor Plan Template (MPT) in the Sensory Motor Memory, to be specified to a concrete Motor Plan to generate a sequence of motor commands onto the agent actuators, such as its simulated hands to execute the *grasp*.

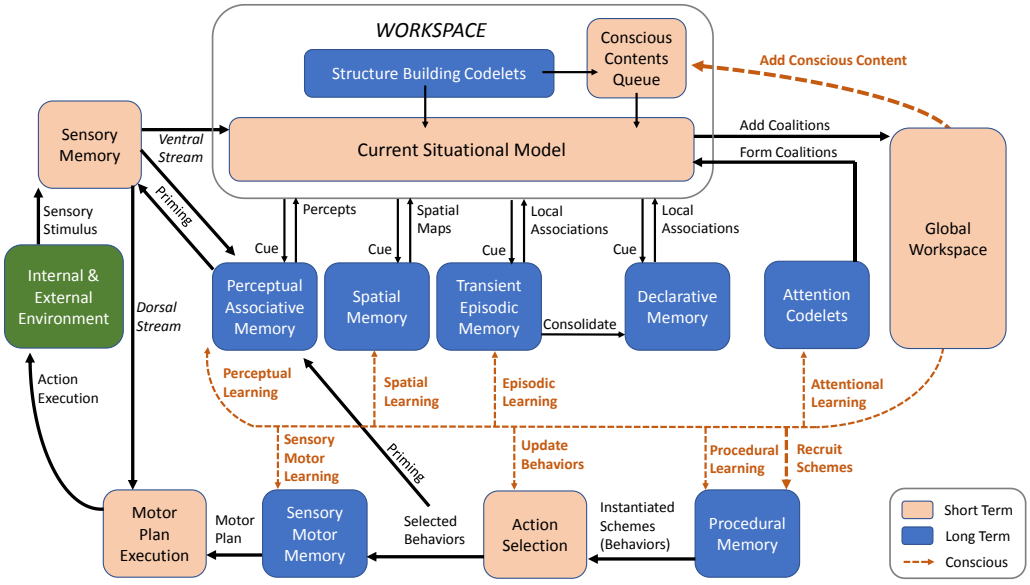

Figure 1: The LIDA model diagram

The Design of the Description Model

We added a new description cognitive model into the agent's controller, the LIDA model, to take its conscious content as the describing contents, and carry a describing behavior to share the contents.

Describing Contents

Similar to having a thirsty node, we build in a *description* node into the agent's (internal) environment which will be sensed and perceived into the agent's CSM, representing that the agent understands it has the description need internally. We add a new *description attention codelet* concerning the availability of the *description* node and its relationships with other structures such as the *thirsty* node, then to form a new coalition to represent the concept of *needing to describe the thirsty (feeling)*.

Based on this description attention codelet, multiple description kind of coalitions may be formed, such as that of needing to describe thirsty, or a glass of water. Within the Global Workspace, one of these description coalitions will win from the competition to be the conscious content.

Furthermore, we proposed two other types of description coalitions. First, when the thirsty agent was grasping a cup of water, an expectation (attention) codelet had been created in its CSM to monitor the grasping result and form a coalition to bring this result into its consciousness. The new description attention codelet may combine forces with this expectation codelet to create a joint coalition, to represent the concept of *needing to describe how well/much the grasp has been done*.

Second, in LIDA (Fig. 1) a Structure Building Codelet (SBC) monitors the Conscious Contents Queue (CCQ) to build an event relating to the time concept (Snaider, McCall, & Franklin, 2010). The above agent may have attended on its *thirsty* feeling a while, so some near past conscious contents listed in its CCQ contain the *thirsty* node. A SBC will count these thirsty-node-involved conscious contents, to build a new structure in the CSM to represent the duration of being thirsty. The new description attention codelet may choose this *thirsty duration* structure to form a coalition to represent the concept of *needing to describe how long the thirsty has been available*.

Describing Behaviors

When a description coalition became the conscious content and arrives to Procedural Memory (Fig. 1), certain *description schemes* will be recruited if their contexts are matching to the conscious contents and their actions are capable of accomplishing the describing, such as in the type of *draw*, *speak*, or *write*. (Franklin et al., 2016)

These schemes are instantiated to behaviors with their variables are specified according to the conscious contents. In the Action Selection module, one of these description-capable behaviors is selected depending on their contexts and activations.

In Sensory Motor Memory, a Motor Plan Template such as *speak* is selected based on the selected behavior, and will be specified to a Motor Plan to run where the template's variables are initiated using the contexts. For example, when the agent chose to speak about the thirsty, the context of the *thirsty duration* helps determine the value of a variable as the term of "very much" if longer duration, or "a little bit" if shorter. Here the variable value is exampled as human language terms, while broader types may apply such as the voice tones, etc. (Dong & Franklin, 2015)

The Sensory Memory collects the environmental stimuli online and feed it to the Motor Plan Execution through the dorsal stream directly (Fig. 1). The internal environmental data such as its describing contents will be sent to the current description motor plan to support its running. When the describing contents changes, for example the duration of being thirsty grows, the running motor plan will update its corresponding variable to strengthen the thirsty degree. This online control forms a dynamic pattern to control the agent's actuators during the description execution.

So far, we illustrated how our description model works with examples, showing that an agent describes its current conscious content as its mind activities at the moment. This description gives insights to explain the agent's behaviors: the agent describes about its *thirsty* when it *grasps* a glass of water, which jointly tells a potentially reasonable causal relationship between the agent's mind activities and its behaviors.

The Initial LIDA Agent

The LIDA Framework is an underlying computational software framework (Snaider, McCall, & Franklin, 2011) which provides the default implementations of main LIDA modules and processes. It supports generic and configurable design principles, helping the developers build their customized LIDA agents productively. It also provides an experimental GUI tool, displaying the cognitive representations and processes of the agent's modules (Fig. 2).

We implemented the new description model into a LIDA agent using this Framework. It senses an object of water from the external environment and has both a *thirsty* and *description* nodes built-in internally. The agent attends on some of these as its conscious content, and chooses to execute certain actions to meet its agenda, such as *grasp, speak,* or *draw*.

In detail, we implemented the describing contents and behaviors among different LIDA modules as listed below in next page. Fig. 1 illustrates the relationship of these modules.

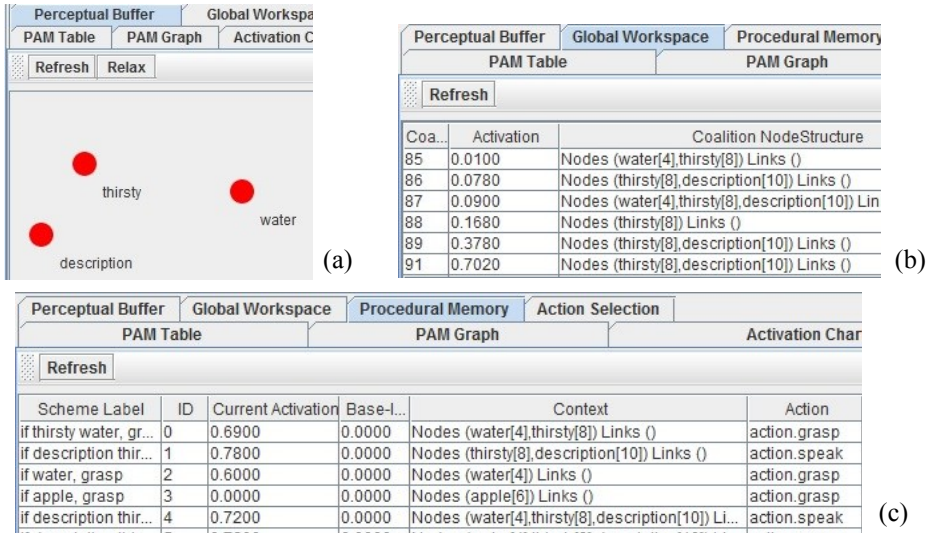

Figure 2: The snapshots of the LIDA agent's inner modules: (a) Percepts of the Current Situational Model, (b) Attention coalitions in Global Workspace, and (c) Schemes in Procedural Memory

The Environment Module

Besides a water object randomly appearing in the external environment, we configured an internal environment module where the *thirsty* and *description need* are built in.

Sensory Memory (SM)

We added the sensing processes to get sensory data from the Environment Modules, both externally and internally.

Perceptual Associate Memory (PAM) and Feature Detectors (FDs)

PAM stores a set of nodes, each of them representing a specific aspect of an environmental state. These nodes are the object water and the inner *thirsty* and *description*. FDs constantly obtain the current state from the SM, activating relevant nodes in PAM.

The Current Situational Model (CSM)

The CSM receives currently activated nodes from PAM, and builds the agent's understanding of the current situation. It understands that the water is out there and it is both *thirsty* and has *description need*, as represented in the CSM through its perceptual buffer (Fig. 2 (a)).

Attention Codelets and the Global Workspace (GW)

Besides the water object and internal *thirsty* attention codelets, we added a set of *description* attention codelets concerning the node structures which have the *description* node involved. These attention codelets form their concerns to coalitions and bringing it to the GW (Fig. 2 (b)), where a description coalition may win the competition to become the agent's conscious content.

Procedural Memory (PM)

We added two types of schemes: 1) *grasp* it if any objects out there like water, and setting activations to them depending on other structures like having *inner thirsty* and 2) *speak* or *draw* the current conscious contents if the agent has the *description need* internally. Being thirsty and sensing water are the possible conscious contents (Fig. 2 (c)).

Conclusions and Next-Steps

The proposed LIDA agent describes its conscious content autonomously without any supervising assistances. The describing contents are supplied from the agent's artificial mind, and the describing behaviors are carried by the agent itself. Also from its mind sharing, it is reasonable to infer that being thirsty may cause the agent to grasp the water.

We will build more detailed *description* schemes needing to have extensive communication context to support Action Selection better. For example, whether it is a short elevator speech thus needing to *speak*, or is in a workshop so allowing a longer expression through the *draw*.

A feeling-based motivation system had been studied in LIDA (Franklin et al., 2016; McCall et al., 2020), which provided bridges between LIDA and other existing motivation related concepts. We plan to continue this motivation study on the mind sharing part.

Further, we plan to study this mind sharing capabilities upon among different cognitive architectures such as ACT-R, Soar, DIARC (Scheutz et al., 2019), etc., to build more general intelligent agents sharing their minds.

Acknowledgments

The authors would like to truly thank Stan Franklin for his ideas, questions, and guidance during initial discussions. Specific thanks go to Steve Strain and Sean Kugele who gave helpful suggestions.

References

Baars, B. J. 1988. *A cognitive theory of consciousness*. New York: Cambridge University Press.

Baars, B. J. 2002. The conscious access hypothesis: origins and recent evidence. *Trends in cognitive sciences, 6*(1), 47-52.

Bono, A., Augello, A., Pilato, G., Vella, F., & Gaglio, S. 2020. An ACT-R based humanoid social robot to manage storytelling activities. *Robotics, 9*(2), 25.

Bullock, T. H. 1993. Goals and Strategies in Brain Research: The Place of Comparative Neurology *How do Brains Work?* (pp. 1-8): Springer.

Chin-Parker, S., & Bradner, A. 2010. Background shifts affect explanatory style: How a pragmatic theory of explanation accounts for background effects in the generation of explanations. *Cognitive processing, 11*(3), 227-249.

Dong, D., & Franklin, S. 2015. A New Action Execution Module for the Learning Intelligent Distribution Agent (LIDA): The Sensory Motor System. *Cognitive Computation*, 1-17. doi: 10.1007/s12559-015-9322-3

Franklin, S. 1995. *Artificial Minds*. Cambridge, MA: MIT Press.

Franklin, S., & Graesser, A. 1997. Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents *Intelligent agents III agent theories, architectures, and languages* (pp. 21-35). London, UK: Springer-Verlag.

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*, 105-130. doi: 10.1016/j.bica.2016.04.003

Khayi, N. A., & Franklin, S. 2018. Initiating language in LIDA: learning the meaning of vervet alarm calls. *Biologically Inspired Cognitive Architectures, 23*, 7-18.

Laird, J. E., Gluck, K., Anderson, J., Forbus, K. D., Jenkins, O. C., Lebiere, C., Salvucci, D., Scheutz, M., Thomaz, A., & Trafton, G. 2017. Interactive task learning. *IEEE Intelligent Systems, 32*(4), 6-21.

Laird, J. E., Lebiere, C., & Rosenbloom, P. S. 2017. A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *Ai Magazine, 38*(4), 13-26.

Langley, P., Laird, J. E., & Rogers, S. 2009. Cognitive architectures: Research issues and challenges. *Cognitive Systems Research, 10*(2), 141-160.

Lindes, P., Mininger, A., Kirk, J. R., & Laird, J. E. 2017. *Grounding language for interactive task learning.* Paper presented at the Proceedings of the First Workshop on Language Grounding for Robotics (1-9).

Lindes, P. 2018. The Common Model of Cognition and humanlike language comprehension. *Procedia Computer Science, 145*, 765-772.

Lombrozo, T. 2006. The structure and function of explanations. *Trends in cognitive sciences, 10*(10), 464-470.

Madl, T., Baars, B. J., & Franklin, S. 2011. The timing of the cognitive cycle. *PloS one, 6*(4), e14803.

Matarese, M., Rea, F., & Sciutti, A. 2021. A user-centred framework for explainable artificial intelligence in human-robot interaction. *arXiv preprint arXiv:2109.12912*.

McCall, R. J., Franklin, S., Faghihi, U., Snaider, J., & Kugele, S. 2020. Artificial motivation for cognitive software agents. *Journal of Artificial General Intelligence, 11*(1), 38-69.

Mutlu, B., Roy, N., & Šabanović, S. 2016. Cognitive human–robot interaction. *Springer handbook of robotics*, 1907-1934.

Newell, A. 1973. You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing*: New York: Academic Press.

Ramaraj, P., Klenk, M., & Mohan, S. 2020. *Understanding intentions in human teaching to design interactive task learning robots.* Paper presented at the RSS 2020 Workshop: AI & Its Alternatives in Assistive & Collaborative Robotics: Decoding Intent.

Scheutz, M., Williams, T., Krause, E., Oosterveld, B., Sarathy, V., & Frasca, T. 2019. An overview of the distributed integrated cognition affect and reflection diarc architecture. *Cognitive architectures*, 165-193.

Snaider, J., McCall, R., & Franklin, S. 2010. *The immediate present train model time production and representation for cognitive agents.* Paper presented at the 2010 AAAI Spring Symposium Series.

Snaider, J., McCall, R., & Franklin, S. 2011. The LIDA framework as a general tool for AGI *Artificial General Intelligence* (pp. 133-142). Berlin Heidelberg: Springer

Sofge, D., Trafton, J. G., Cassimatis, N., Perzanowski, D., Bugajska, M., Adams, W., & Schultz, A. 2004. *Human-robot collaboration and cognition with an autonomous mobile robot.* Paper presented at the In Proceedings of the 8th Conference on Intelligent Autonomous Systems (IAS-8) (80-87).

Umbrico, A., De Benedictis, R., Fracasso, F., Cesta, A., Orlandini, A., & Cortellessa, G. 2022. A Mind-inspired Architecture for Adaptive HRI. *International Journal of Social Robotics*, 1-21.