Metacognition in Content-Centric Computational Cognitive (C^4) Modeling

Sergei Nirenburg, Marjorie McShane and Sanjay Oruganti

The Language-Endowed Intelligent Agents Lab
Rensselaer Polytechnic Institute

Abstract. For AI agents to emulate human behavior, they must be able to perceive, meaningfully interpret, store, and use large amounts of information about the world, themselves, and other agents. Metacognition is a necessary component of all of these processes. In this paper we briefly a) introduce content-centric computational cognitive (C^4) modeling for next-generation AI agents; b) review the long history of developing C^4 agents at RPI's LEIA (Language-Endowed Intelligent Agents) Lab; c) discuss our current work on extending LEIAs' cognitive capabilities to cognitive robotic applications developed using a neurosymbolic processing model; and d) sketch plans for future developments in this paradigm that aim to overcome underappreciated limitations of currently popular, LLM-driven methods in AI.

Metacognitive abilities are a key prerequisite for making AI agents full-fledged members of human-AI teams. AI agents must use metacognition both for *introspection* and for *mindreading*—that is, understanding the knowledge, reasoning, intentions, skills, personality traits, and preferences of themselves and their teammates. Core prerequisites for introspection and mindreading are maintaining and dynamically enhancing a) the agent's ontological model of the world and the agents in it; b) resources that link elements of perception with the agent's mental models (e.g., a lexicon that links words and phrases to ontological concepts); and c) the agent's memories of past experiences of perception interpretation, reasoning, and action. All this content supplies essential metacognitive heuristics for the agent's decisions.

It cannot be overemphasized that semantically interpretable knowledge resources are essential to an agent's ability to select appropriate actions and explaining why those actions were chosen. This is the *content-centric* aspect of C^4 modeling. Moreover, interpreted knowledge facilitates an agent's instructing and being instructed by other agents through show and tell, the way people are taught in everyday situations and in all manner of training environments.

Crucially, interpreted knowledge resources support *a variety* of computational approaches to realizing metacognitively endowed AI agents – rule-based and machine learning-based ones as well as hybrid, so-called neurosymbolic approaches.

A Brief Survey of Metacognition in Agents Developed Using C^4 Modeling

Wei et al. [3] characterize metacognition as supporting the following four capabilities (the definitions are ours): **transparency**, which involves an agent's explaining its reasoning and decision-making; **adaptability** to novel situations and in support of lifelong learning; **reasoning**, including its self-aware aspects; and **perception**, which requires interpreting the output of perception-oriented technologies. This taxonomy is incomplete, especially if we take into account cognitive robotics. For this, a fifth capability must be added: **action,** both physical and verbal, which agents must carry out within their teams.

In the LEIA lab, we have been developing all of the abovementioned capabilities within the C^4 modeling framework. Sample prototype applications are the Maryland Virtual Patient (MVP)

system for training medical students [4: Chapter 8]; a virtual vehicle agent [5: section 7.1.5]; and several simulated human-robot team applications based on the HARMONIC cognitive-robotic architecture [6].

Transparency. In all our systems, (a) the output of all system modules is available, in human-legible form, for inspection, and (b) a special module is devoted to generating explanations, in plain English, of the reasons for agent decisions.

Adaptability. When virtual patients in MVP engage in dialog with human users, they mindread them, taking into account their personality traits, physical and mental states, and levels of domain knowledge. In all of our systems, agents engage in lifelong learning of new ontological concepts and new lexical material through dialog with teammates. This is made possible by our extensive work on deep natural language understanding that uses stored knowledge resources for bootstrapping (see, e.g., [5: chapter 7]).

Reasoning. LEIAs engage in reasoning when interpreting input, deciding on instantiating and prioritizing goals, selecting plans, carrying out plans, dealing with disturbances, and choosing how to implement individual actions within the plans. All these tasks involve heuristic decision functions whose argument sets include values of a number of metacognitively-related metaparameters. For example, if the computational *cost* of determining a parameter value in a decision function is too high, then the function can be run without that feature, albeit with a lower *confidence* in the resulting decision. Confidence is, in turn, computed using metaparameters including *vagueness* and *incompleteness* of sensory input. Confidence is among the determinants of *actionability* – that is, whether the agent believes it is licensed to act on an incomplete understanding of an input, given an application's requirements (for detailed discussions of actionability, see [4-6]).

Perception. In all our systems, results of perception are interpreted in terms of the system's knowledge resources. Interpretation routinely takes into account metacognitive aspects, such as the agent's history, its mindreading of other agents, etc. A good example of the use of metacognition in perception is the LEIAs' ability to recover from ill-formed language utterances and detect cognitive biases in others (see relevant discussions throughout [4], especially chapters 3 and 4, and section 8.2]).

Action. In all our systems, LEIAs generate verbal actions to communicate with teammates. LEIAs not only produce an English rendering of the underlying thought but also select a style and word choice that reflects mindreading of teammates' beliefs, intentions, emotions and personality traits as well as their shared history. Thus, when a virtual patient in MVP comes to a repeat visit to a particular doctor and the doctor asks, "How are you?" the LEIA judges it appropriate to respond with the comparative "I'm feeling better."

Evolution of the Computational Infrastructure for C^4 modeling

Originally we implemented LEIAs as predominantly rule-based systems. But in light of the technological leap offered by LLMs, we recently switched to a hybrid, neurosymbolic infrastructure. Howeover, our approach to hybridization differs from most current integration proposals (see [7] for a survey), which focus on LLMs and use limited knowledge-based support in an effort to boost performance. Our approach is the opposite: We focus on C^4 modeling with the goal of

producing trustworthy agents and integrate LLMs as means of improving system performance. To date, we have incorporated LLMs in two components of LEIAs – language generation and lifelong learning through understanding.

Unlike LLMs, C^4 agents generate text intentionally, as a step in consciously pursuing a goal. This process involves both the selection of the content to be conveyed and the choice of how to actually say it in English. We use knowledge-based methods to select the content and generate multiple candidate sentences to convey it. Then we use an LLM to decide which of those sentences is best in the context. This is precisely the kind of task that LLMs are good for because it requires a mastery of how the surface level of language works without the need to take responsibility for its content (see [5], section 4.3). We have tested a variation on the above capability in a system for automatic authorship anonymization [8] in which LLMs helped to filter out atypical textual formulations and offered additional text paraphrase solutions when the knowledge-based engine failed to adequately anonymize a text.

To implement lifelong learning through understanding, LEIAs use their available resources and processors to learn new, and improve existing, lexicon entries and ontological concepts by understanding natural language texts or inputs from human or robotic instructors. Our team's early implementations of this process [9-13] were rule-based. The approach we are now working on incorporates LLMs to enhance the efficiency of the learning process by filtering the lexical material for LEIAs to interpret during the learning process. The algorithm for this process, described in detail in [5: Chapter 7], is currently being implemented in an application of the HARMONIC cognitive-robotic architecture [6].

Conclusion

This paper argues that content-centric computational cognitive (C^4) modeling is the most promising methodology for building trustworthy AI agents that are self-aware and capable of human-level explanations. Only such agents are fit for truly critical applications in defense, health, finance, etc. Metacognition is an integral feature of C^4 modeling, as illustrated by the above examples of C^4-based systems the RPI LEIA lab has built. We have also demonstrated that C^4 modeling can be implemented in a variety of computational infrastructures, including the novel neurosymbolic one we are implementing. In the immediate future we intend to demonstrate that our approach to lifelong learning through understanding will remove the so-called "knowledge bottleneck" and will facilitate the development of flexible and reliable agents and robots that can become full-fledged members of human-AI teams.

References.

1. Gunning, D. 2017. Explainable artificial intelligence (XAI), Tech. Rep., Defense Advanced Research Projects Agency (DARPA)].
2. Babic, B., Gerke, S., Evgeniou, T, & Cohen, I. G. 2021. Beware explanations from AI in health care. *Science,* 373:6552. July.
3. Wei, H., Shakarian, P., Lebiere, C., Draper, B., Krishnaswamy, N., & Nirenburg, S. 2024. Metacognitive AI: Framework and the case for a neurosymbolic approach. *Proceedings of NeSy 2024*, Bacelona, Spain, September.
4. McShane, M., & Nirenburg, S. 2021. *Linguistics for the Age of AI*. MIT Press.
5. McShane, M., Nirenburg, S., & English, J. 2024. *Agents in the Long Game of AI: Computational cognitive modeling for trustworthy, hybrid AI***.** MIT Press.

6. Oruganti, S., Nirenburg, S., McShane, M., English, J., Roberts, M., & Arndt, C. 2024. HARMONIC: A framework for explanatory cognitive robots. *Proceedings of ICRA@40*. Rotterdam, The Netherlands, September.
7. Besold, T., A. d. Garcez, S. Bader, H. Bowman, L. C. Lamb, L. de Penning, B. Illuminoo, H. Poon, C. Gerson Zaverucha. 2022. Neural-symbolic learning and reasoning: A survey and interpretation. arXiv:1711.03902**.**
8. McShane, M., Nirenburg, S., Arndt, C., Oruganti, S., English, J. Forthcoming. A neurosymbolic approach to authorship anonymization. *Cognitive Systems Research.*
9. Nirenburg, S., Oates, T., & English, J. 2007. Learning by reading by learning to read. *Proceedings of the International Conference on Semantic Computing*. San Jose, August.
10. McShane, M., Nirenburg, S., Jarrell, B., and Fantry, G. 2015. Learning components of computational models from texts. In Mark A. Finlayson, Ben Miller, Antonio Lieto, and Remi Ronfard (Eds.), *Proceedings of the 6th Workshop on Computational Models of Narrative* (CMN'15), pp. 108-123.
11. Nirenburg, S., & Wood, P. 2017. Toward human-style learning in robots. *AAAI Fall Symposium on Natural Communication with Robots*.
12. Nirenburg, S., McShane, M., & English, J. 2018. Toward life-long human-like learning by intelligent agents. *Proceedings of the Annual Conference on Advances in Cognitive Systems*. Stanford. May.
13. Nirenburg, S., McShane, M., & English, J. 2021. Overcoming the knowledge bottleneck using lifelong learning by social agents. *Proceedings of NLDB-21*, Saarbrücken, Germany, June.