Growing An Embodied Generative Cognitive Agent

Spencer K. Lynn, Bryan Loyall, James Niehaus
Charles River Analytics
slynn@cra.com, bloyall@cra.com, jniehaus@cra.com

Abstract

An evolutionary perspective on embodiment puts maintenance of physiology within a functional envelope as the brain's base goal, with all other goals as refinements. Thus, all goals have physiological perturbation for their motivation and allostatic recovery as their signal of fulfillment. From this account, two entailments emerge. First, an object's properties are not intrinsic to the object but a situated function of the morphology of the object and the affordances required by the goal. Second, categories do not exist without reference to some goal; they are constructed at the time of perception by blending prior conceptual knowledge to create an understanding of the perception with respect to the goal. Our thesis is that generative large language model (LLM) architectures are part of the solution to creating artificial organic-like cognitive architectures, but that LLMs as currently trained are generative only at a surface-level of behavior rather than deeper levels of cognition and, furthermore, that generative architectures must be coupled with an embodied cognitive agent architecture, which suggests both the additional levels at which generativity must operate and capabilities that the combined architecture must support.

Introduction

The performance of generative large language models (LLMs) is by turns extraordinary and extraordinarily untrustworthy. It is itself generating extraordinary interest throughout artistic, social, educational, commercial, and governmental, including military, stakeholders. Within the cognitive sciences, discussion about LLMs is the current manifestation of the on-going conversation about the nature and distinction of organic intelligence and general artificial intelligence (AI). LLMs are useful now, will be more useful (e.g., highly performant within well-bounded domains), and will be made to be more trustworthy (perhaps, e.g., as the user interface to other, more transparent, systems). Here, however, we describe critical barriers to the general or common-sense intelligence that many in the field of AI desire (e.g., Marge et al. 2022, Blaha et al. 2022) that we feel LLMs, alone, will not bridge.

An evolutionary biology perspective on embodiment puts physiological homeostasis at the center of brain function, with a ca. 600-million-year history (Shaffer et al. 2022). If we adopt the proposal that the base goals of an organism involve maintenance of physiology within a functional envelope that facilitates the "four Fs" – feeding, fleeing, fighting, and reproduction (Churchland 1987), then let us adopt a strong position for the sake of argument: all goals are refinements of this base goal. This is a strong sense in which cognition is embodied – even abstract goals (e.g., math, economic transactions) have physiological perturbation for their motivation and allostatic recovery as the signal of their fulfillment (Barrett 2017).

On this account, two requirements emerge, which have been explored in psychology and linguistics but seldom addressed in cognitive architectures. First, perception serves goals – its function is to detect affordances in the environment that help the organism achieve its goals. Properties of an object, then, are not intrinsic to the object but temporary, situated functions of the morphology of the object and the affordances required by current goals. Second, concepts are blended – categories do not exist without reference to some goal or function and are instead constructed at the time of perception by blending prior conceptual knowledge to create an understanding of the perception with respect to the goal.

Our thesis is that generative LLM architectures are part of the solution to creating artificial organic-like intelligence, but that LLMs as currently trained are generative only at a surface-level of behavior (language production) rather than deeper levels of cognition (e.g., models of entities in the world and their relations). Furthermore, the generative architectures must be coupled with an embodied (sensu Lakoff 2008) cognitive agent architecture. Embodiment suggests both the additional levels at which generativity

must operate and capabilities that the combined architecture must support.

To develop this thesis, our starting point is the problem of LLM trust as an illustration of misaligned goals and affordances. That discussion leads us to the critical importance of embodied goals, active perception, and constructivist categorization, and eventually to hypotheses about the capabilities an integrated embodied, generative, cognitive architecture will need.

Designed versus Embodied Goals

LLMs, like any engineered system, become untrustworthy when people use them to achieve goals for which they were not designed (i.e., goals the LLMs themselves do not share, because they were not designed with that goal in mind). For example, using an LLM as an internet search engine or a calculator can produce wrong answers. It should not be surprising when the answers are wrong, because LLMs were designed (i.e., their goal is) to produce naturalistic language output, not factual output. LLMs, alone, are ill suited as search engines – companies are working to put guardrails in place to make workable search engines of LLMs (e.g., Microsoft, for Bing with ChatGPT). So, trust requires shared goals. Where do human goals come from? They come from our bodies, of course, and every act is a kind of goal (*problem* or *drive* or *behavior* may also be suitable words). Goals decompose into subgoals, such that the goal of, e.g., changing a light bulb, has subgoals that include resolving contrast gradients to detect edges, which construct objects, such as a stepstool, and raising your blood pressure sufficiently to support climbing the stepstool to change the lightbulb.

Affordances

The way humans and other animals solve a problem is by detecting the affordances in the environment that address the problem. We perceive the environment according to how it can fulfill our needs. If your goal is to grasp an object, you detect features of the object that afford grasping (Gibson 1977; Hedblom et al. 2015). LLMs are generative at levels of word, phrase, and longer language structures. However, mismatch of user goals versus LLM design causes perceived useability failures because the affordances LLMs detect (e.g., associations developed during training) do not match those that the humans need to solve the problem that humans bring to the LLM in the form of a prompt. For example, LLMs can learn the affordances of writing in a particular linguistic style (where such affordances might correspond to phrase structures and word choices characteristic of the style). However, mismatch can occur when factual accuracy of the content is an important part of the goal (see, e.g., Cheng 2023), because LLMs are not designed to detect the affordances of producing accurate output (where such affordances might be source reputability or returning a verbatim quote rather than a paraphrase). Our point is that generativity needs to operate on affordances because it is affordances that suggest solutions to achieving goals. AI is called brittle and untrustworthy when it is not capable of detecting the affordances that address the goals that the user has brought to the AI. The generativity of LLMs is effective but applied to a limited set of goals – specifically, natural-seeming written language.

Goal-Directed Perception

Perception serves goals – it detects features, affordances, in the environment that can be used to achieve a goal. An agent perceives affordances that are suited to what the agent can do in the world, the goals it has, and problems it can solve. Because perception is about affordances with respect to current goals, perception does not discover intrinsic properties of things in the world. It discovers properties that are suitable for addressing the current goal. What is intrinsic to this process is in the agent, not in the world. The agent's goal and affordances have intrinsic meaning for the agent because they are grounded in their functional relevance to the agent – what the agent needs to be successful, such as the four Fs for an organic agent.

Concept Blending and Category Construction

We have said that organic intelligence perceives the world through the lens of its goals. Here, "perceives the world" means apply knowledge about a concept (e.g., one's goal) to one's current perception. It is this act of categorization of sensory input that creates understanding; it imputes meaning to sensory input. Meaning is always with respect to a goal, otherwise there is no grounding to the body's physiology. From the embodied cognition perspective, categorization is understood to be constructed (Barrett 2017), blended (Turner 2019), or situated (Barsalou 2015), all of which are descriptions that capture the notion that categories do not exist without reference to some goal or function and are instead *constructed* at the time they are needed by *blending* prior conceptual knowledge, a process influenced by the exigencies of the current *situation* (i.e., internal and external context, including current goals and past history). This process is what enables a person to who needs to change a light bulb (the goal) to understand that a chair (the sensor input) can be used as a stepstool (a prior concept believed to have affordances suited to the goal). Blending the desired affordances from the stepstool concept with the inferred attributes of the current chair percept constructs a new, situated category: chair-as-stepstool. This ad hoc category

gives meaning to the perception (the chair) with respect to the goal (changing a light bulb).

Predictive Coding

LLMs exemplify a performant, if shallow, generative architecture. Agent cognitive architectures exemplify attempts to capture features of embodiment and biologically inspired cognitive constructs, including goals, affordances, perception, decision, memory, and action selection. In contrast, organic intelligence appears generative at every level of organization, an idea captured by the neural theory of predictive coding (Allen and Friston 2018). Like LLMs, predictive coding is generative, but it is also deeply hierarchically predictive, implementing causal models at all levels of cognition. Notably, this organization may not align with traditional cognitive constructs and psychological faculties embedded in cognitive architectures (e.g., a predictive coding hierarchy blurs the distinction between perception, decision, and action; Chanes and Barrett 2016, Lynn 2018). Predictive coding is congruent with the constructivist approach described here and deserves consideration as a key feature of agent cognitive architectures.

In the traditional stimulus-response model of cognition, cognitive processes categorize perceived stimuli to build models from which to derive meaning and make decisions. Beliefs and concepts are the result of perception. In contrast, from the predictive coding perspective, beliefs about the world yield predictions about sensory inputs. Beliefs and concepts are the beginning of perception, not the result. The upper levels of a predictive coding hierarchy are about physiological allostasis, which motivates behavior (Barrett & Simmons 2015). The predictions become increasingly specific down the hierarchy towards sensory and motor neurons. Prediction errors, not stimulus properties, as passed upward in the hierarchy. Predictive coding offers a model of how a generative architecture can integrate with conventional cognitive architectures. Parent nodes in the hierarchy are more functionally oriented (what the system needs to do to: goals), child nodes are more mechanistically oriented (how the system might do it: affordances).

Embodied Generative Agents

What must the architecture look like that can support this dynamic concept blending in the service of deriving meaning from perceptions over hierarchies of goals and subgoals? Our consideration of how embodied, organic intelligence derives meaning from sensory inputs suggests seven characteristics that a cognitive architecture must implement to escape brittleness and engender trustful human-AI teaming.

1. Embodied: As an agent-based cognitive architecture, the agent's goals must motivate categorization of its perceptions in support of its possible actions.
2. Hierarchical: The architecture must have data structures that support what we have called goals and subgoals, from top-level functions down toward raw sensor processing to some arbitrary depth that is sufficient to detect the affordances by which the system can achieve its goals.
3. Constructive: The architecture must blend or construct new conceptual structures from prior knowledge to explain inputs.
4. Situated: The concept blending must be influenced by multiple conceptual structures – some of which supply inputs to the blend (sources), but some of which constrain or select which elements of prior knowledge are inherited by the new blended concept (contexts).
5. Generative: The architecture must be generative at many levels, not merely behavior (e.g., LLM output) but also generating collections of affordances hypothesized to solve a problem so that the system can understand the environment with respect to its goals before candidate behaviors can be generated.
6. Sub/Symbolic: The architecture must combine non-symbolic, data-driven approaches with symbolic, model-like knowledge representations. LLM narratives are often nonsensical or "imaginative" because there is no underlying model of relationships or processes beyond mere associations learned from the text-based training material. It is therefore common for this generated narrative to violate human expectations, which themselves appear model-like, with causes that we use to explain our observations.
7. Learnable: Learning must, to some extent, be able to occur during operation, reflecting the refining, blending, and generation of new concepts as the system acquires new goals or executes familiar goals in new situations.

These characteristics intersect with and expand on LLMs and current cognitive architectures. For example, agents can provide embodied goals, sensing, and behavior to semantically ground LLM associations (e.g., Kirk et al. 2023). Additionally, the reasoning LLMs perform might benefit from fine-tuning with structured agent experience (Xiang et al. 2023), which can provide cognitive models; hierarchically organized knowledge can provide that structure. As well, learnability suggests an important role for cognitive architecture integration; traditional faculties, such as learning and memory, symbolic representation, temporal reasoning, chunking, and concurrent goal execution would be required for autonomous fine-tuning of LLM-equipped agents.

We have investigated preliminary computational feasibility of some of the ideas described here (e.g., Pfeffer and Lynn 2019, Hyland et al. forthcoming), and developed an open-source probabilistic and composable AI framework, Scruff, capable of integrating the components

expressing the desired characteristics (Pfeffer et al. 2021, github.com/charles-river-analytics/Scruff.jl).

Conclusion

The approach described here grows an embodied deep generative agent. The agent grows because it can start with a shallow hierarchy, limited sensoria, few goals, and simple actions. Initially, its goals and actions are necessarily simple by dint of limited depth and sensors. It is embodied because it the meaning of sensory and motor processes are only defined with respect to its goals, which are intrinsically motivating. It is deep because it spawns new computational nodes in a hierarchy as it learns. The hierarchical architecture efficiently provides richness of representation as the system grows. It is generative because the motivation to perceive and act is causal. The same generative mechanism can efficiently motivate both perception and action.

The outline provided here of the theory of embodied cognition as developed in psychology, linguistics, and biology over the last three decades suggests that category construction in the service of perceiving affordances that can be used to satisfy goals is a critical perspective missing from many current approaches in AI. Integrating generative and embodied approaches should be part of the solution to AI brittleness and trustworthiness because it is through embodied generative hierarchies, spanning perception to concept, that meaning is grounded – the environment is understood with respect to goals, which serve to maintain the agent's functions.

To increase autonomy and human-AI teaming, the challenge is to provide the AI with human-like conceptual structure. "Common ground" refers to congruent knowledge, beliefs, and assumptions among a team about their objectives, context, and capabilities (Clark and Wilkes-Gibbs 1986). Common ground is essential to human-AI teaming and trusted autonomy (Dafoe et al. 2021). A cognitive architecture that provides the AI with representational capacity and algorithms that mimic features of human conceptual structure and flexibility by integrating deep generativity and constructive processes can shift human-AI common ground from mere user interface transparency to concept congruency, where it resides for trusted human-human interactions.

References

Allen, M., and Friston, K. J. 2018. From Cognitivism to Autopoiesis: Towards a Computational Framework for the Embodied Mind. *Synthese*, 195(6), 2459-2482.

Barrett, L. F. 2017. The Theory of Constructed Emotion: An Active Inference Account of Interoception and Categorization. *Social Cognitive and Affective Neuroscience,* 12(1), 1-23.

Barrett, L. F. 2017. *How Emotions Are Made: The Secret Life of the Brain*. London: Pan Macmillan.

Barrett, L. F. and Simmons, W. K. 2015. Interoceptive Predictions in the Brain. *Nature Reviews Neuroscience* 16(7), 419-29.

Barsalou, L. W. 2015. Situated Conceptualization. In *Perceptual and Emotional Embodiment: Foundations of Embodied Cognition* (pp. 1–11). Routledge.

Blaha, L. M.; Abrams, M.; Bibyk, S. A.; Bonial, C.; Hartzler, B. M.; Hsu, C. D.; Khemlani, S.; King, J.; St. Amant, R.; Trafton, J. G.; and Wong, R. 2022. Understanding is a Process. *Frontiers in Systems Neuroscience* 16:800280.

Chanes, L., and Barrett, L. F. 2016. Redefining the Role of Limbic Areas in Cortical Processing. *Trends in Cognitive Sciences*, 20(2), 96-106.

Chiang, T. 2023. ChatGPT is a Blurry JPEG of the Web. *The New Yorker*, 9 February, https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web.

Churchland, P. S. 1989. *Neurophilosophy: Toward a unified Science of the Mind-Brain*. Cambridge: MIT press.

Clark, H. H.; and Wilkes-Gibbs, D. 1986. Referring as a Collaborative Process. *Cognition*, 22(1), 1–39.

Dafoe, A.; Bachrach, Y.; Hadfield, G.; Horvitz, E.; Larson, K.; and Graepel, T. 2021. Cooperative AI: Machines Must Learn to Find Common Ground. *Nature*, 593(7857), 33–36.

Gibson, J. J. 1977. The Theory of Affordances. In *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, edited by Shaw, R., and Bransford, J., 67–82, Hillsdale: Lawrence Erlbaum.

Hedblom, M. M.; Kutz, O.; and Neuhaus, F. 2015. Choosing the Right Path: Image Schema Theory as a Foundation for Concept Invention. *Journal of Artificial General Intelligence*, 6(1), 21-54.

Hyland, R.; Lu, K.; Lynn, S. K.; Marotta, S. J.; Niehaus, J.; Norsworthy, W.; Pfeffer, A.; Wu, C.; and Loyall, B. Forthcoming. AI Inference of Team Effectiveness for Training and Operations. Interservice/Industry Training, Simulation and Education Conference, 27 November-1 December, Orlando, Florida

Kirk, J. R.; Wray, R. E.; Lindes, P.; and Laird, J. E. 2022. Evaluating Diverse Knowledge Sources for Online One-Shot Learning of Novel Tasks. arXiv preprint arXiv:2208.09554.

Lakoff, G. 2008. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. Chicago: University of Chicago Press.

Lynn, S. K. 2018. A Predictive Processing Model of Categorical Perception. In Proceedings of the International Conference on Social Computing--Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation, edited by H. Bisgin, A. Hyder, C. Dancy, and R. Thomson. 10-13 July, Washington, DC, Springer. (On-line version). Retrieved from sbp-brims.org/2018/proceedings/.

Marge, M.; Espy-Wilson, C.; Ward, N.G.; Alwan, A.; Artzi, Y.; Bansal, M.; Blankenship, G.; Chai, J.; Daumé III, H.; Dey, D.; and Harper, M. 2022. Spoken Language Interaction with Robots: Recommendations for Future Research. *Computer Speech & Language*, 71, 101255.

Pfeffer, A.; Harradon, M.; Campolongo, J.; and Cvijic, S. 2021. Unifying Al Algorithms with Probabilistic Programming Using Implicitly Defined Representations. arXiv preprint arXiv:2110.02325.

Pfeffer, A., & Lynn, S. K. 2019. Scruff: A Deep Probabilistic Cognitive Architecture for Predictive Processing. In Biologically Inspired Cognitive Architectures 2018: Proceedings of the Ninth Annual Meeting of the BICA Society (pp. 245-259). New York: Springer International Publishing.

Shaffer, C., Westlin, C., Quigley, K. S., Whitfield-Gabrieli, S., Barrett, L. F. 2022. Allostasis, Action, and Affect in Depression: Insights from the Theory of Constructed Emotion. *Annual Review of Clinical Psychology* 9(18), 553-580.

Turner, M. 2019. Blending in Language and Communication. In *Cognitive Linguistics–Foundations of Language*, edited by E. Dąbrowska and D. Divjak, 245-270. Berlin: De Gruyter.

Xiang, J., Tao, T., Gu, Y., Shu, T., Wang, Z., Yang, Z., & Hu, Z. (2023). Language Models Meet World Models: Embodied Experiences Enhance Language Models. arXiv preprint arXiv:2305.10626.