On integrating generative models into cognitive architectures for improved computational sociocultural representations

Christopher L. Dancy, Deja Workman
The Pennsylvania State University, University Park
cdancy@psu.edu, dqw5409@psu.edu

Abstract

What might the integration of cognitive architectures and generative models mean for sociocultural representations within both systems? Beyond just integration, we see this question as paramount to understanding the potential wider impact of integrations between these two types of computational systems. Generative models, though an imperfect representation of the world and various contexts, nonetheless may be useful as general world knowledge with careful considerations of sociocultural representations provided therein, including the represented sociocultural systems or, as we explain, *genres of the Human*. Thus, such an integration gives an opportunity to develop cognitive models that represent from the physiological/biological time scale to the social timescale and that more accurately represent the effects of ongoing sociocultural systems and structures on behavior. In addition, integrating these systems should prove useful to audit and test many generative models under more realistic cognitive uses and conditions. That is, we can ask what it means that people will likely be using knowledge from such models as knowledge for their own behavior and actions. We further discuss these perspectives and focus these perspectives using ongoing and potential work with (primarily) the ACT-R cognitive architecture. We also discuss issues with using generative models as a system for integration.

Introduction

Cognitive architectures have long presented an opportunity to consider complex and contextual human behavior from the perspective of a unified theory of cognition implemented in an a physical (software) system. More recent generative models (e.g., transformer-based large language models as the one presented by Scao et al., 2022) have offered an interesting opportunity to computationally represent environments (represented digitally through data collected) and the information exchanged within those environments. Thus, these generative models provide a model of (digitally represented) worlds, which are places that hold and exchange the knowledge humans use to make decisions and solve problems (e.g., see Dancy, 2022; Dancy & Saucier, 2022 for related contextualized discussion).

Though these lines of research development and inquiry have had (mostly) different goals, there is an opportunity for integration from the perspective of modeling and simulating behavior, particularly (for this paper) from the perspective of sociocultural systems and social impact of cognitive models built to operate within a cognitive architecture. An integrated cognitive architecture with a generative model may give cognitive modelers opportunities to develop more realistic cognitive models that represent the ongoing sociocultural worlds that humans must operate within. Such systems could more readily give the opportunity to develop cognitive models that are important to *all of* humanity (Prather et al., 2022) or at least more of it. This, of course, will not be without the need to realistically assess and understand the contexts and environments in which such more realistic models might be applied and even the effects of creating those models (e.g., Bender et al., 2021; Birhane et al., 2022).

We lead with critical discussion of potential issues and pitfalls of integrating generative models, many of which have been discussed within (what might be broadly seen as) the AI ethics literature. Following this needed early gesture towards and discussion of such issues, we then discuss the benefits of this integration of cognitive architectures and generative models, both from a perspective of how cognitive architectures may be helpful for generative models and how generative models may be useful for cognitive architectures.

On Issues with Generative Models

Though we come into this paper with optimism for integration of these systems, we see it as a benefit to note and discuss (some of the) potential issues and pitfalls with connecting cognitive architectures and generative models. Furthermore, given the seriousness with which one should take these issues within an environment of techno-optimism, which is at times bolstered large corporation influence (e.g., see Whittaker, 2021, and Young et al., 2022 for related discussions on *corporate capture*), we lead with some discussion on these issues to nod to the importance and impact of these issues given the task at hand.

The taking of data in any form is extractive, and when these data are used to power generative models in the way

that they are currently being made, this extraction can easily become exploitive. Many of the unresolved risks from generative models can be derived from the sourcing of the information, which brings us to the problem of consent. When tools like web scrapers are used to gather the information that will be used by a generative model, not only are the creators or sharers of that information not required to consent to the scraping, but they are also largely unlikely to be aware of it. Users of various sites may arrive with the intention of sharing information with particular groups of people or for specific purposes, but consent and privacy can be disregarded in the massive acquisition of information for these generative models; consequentially, people go from being humans to being a means of production.

More concerns arise in tandem with using people as producers of information, particularly surrounding truth and representation. Truth is a slippery concept for people to grasp, and unsurprisingly even more elusive to be presented by generative technology. There is not a way to ensure that the reality portrayed by these generative technologies is completely accurate—and they are known to not be—so, one is left to consider how to handle the misrepresentations of the world that are known to exist within these models (Mitchell & Krakauer, 2023). These misrepresentations can come in the form of biases and lack of contextualization. As improvement of these models continues and more stake and trust is put into generative models, the significance of these misrepresentations continues to grow.

The creation of generative models is not an entirely transparent process and as a result it is not widespread knowledge how much manual labor often goes into processing this information. OpenAI, creators of ChatGPT and the other GPT models, has been in the headlines as their labor processes were exposed. OpenAI contracted with an organization that paid global workers as low as a few dollars an hour to process large amounts of information, some of which was reported as causing work-related PTSD (Perrigo, 2023). As seen in the case of GPT3, most of the automated processes currently in use do not have the ability to unilaterally filter out all unwanted content. There are certain things that cannot be automatically contextualized well enough to be filtrated without human help, and historically, that human help has not been adequately compensated. With the insistence on supporting further generative technology development needs to come the ability and willingness to responsibly support all working members of the development process.

The increase in generative technologies poses risk not only to people socially but to our physical environment as well. Bender et al. (2021) brought attention to the implications that this level of computational power has for our environment due to high energy requirements. Large models of any kind have the ability to cause a noticeable increase in power consumption and thus energy production which plays a role in further progressing climate change. Not only is this a worldwide consequence, but certain groups of people are disproportionately impacted by climate change, and non-coincidentally this includes marginalized and exploited communities who are already unlikely to benefit as much from these models as other parties.

Taking these factors into consideration, there is a need for greater collective responsibility to address the consequences of large models. While there may not be definitive remedies to some of these problems, it is imperative that awareness that they are present is increased and there continues to be reflection on the social impacts of large generative models in the future.

Generative Models as Sociocultural Symbol (Knowledge) Generators for Cognitive Architectures

Generative models provide a unique opportunity for cognitive architectures at the knowledge level. They can serve both as a tool for translation between representations and a dynamic database of knowledge based on those data a generative system might be trained on. This database of knowledge will also replicate sociocultural representations in relations and structure, including potentially replicating existing systems and structures that have resulted in the oppression of certain groups of people (e.g., Caliskan et al., 2017), thereby certain genres of *the Human* (Wynter, 2003; Wynter & McKittrick, 2015).

Generators of symbol structures for models

Generative models, naturally, may be thought of as *generators* for symbol structures that can solve satisfy tests as specified for a problem space (Newell & Simon, 1976). Though it is now more of the norm for cognitive architectures to have some form of (symbolic/subsymbolic) hybrid representation (e.g., Kelly et al., 2020; Laird et al., 2017), these representations have been limited in their ability to *generate* differing symbol structures and representations from a large number of possible structures. That is to say, approaches in symbolic systems have struggled to scale in the problem of matching a given query of knowledge to reasonable key-value pairs within a *database* of knowledge to accomplish tasks/solve problems. In addition to the problem of scaling, many approaches have shown to be relatively limited, partially because of the issue with being unable to scale well.

Generative models represent an opportunity for representing larger and more flexible databases of knowledge that can be queried for use in cognitive models. While the previously discussed limitations and drawbacks must be kept in mind, the flexibility to represent a wider range of information (in complexity, modality, etc.) presents an opportunity for cognitive architectures to integrate these models to explore

knowledge-level effects across modalities and time scales (Newell, 1990). Thus, we can think through a problem such as how sociocultural knowledge may affect interaction with AI systems (e.g., Atkins et al., 2021; Dancy & Saucier, 2022) while also keeping in mind how lower-level *physiological* and *affective* systems (e.g., Dancy, 2021; Larue et al., 2018) may interact across time to affect behavior. Cognitive architectures need approaches that can pull in larger or more varied databases of knowledge to effectively represent many of the symbolic and subsymbolic structures of sociocultural systems; representation of sociocultural systems is needed for more realistic applied models in many contexts (e.g., those applications requiring dynamic decisions, Gonzalez et al., 2003).

Models of *the Human*

Given arguments for the importance of representations within language for understanding effects of sociocultural systems (e.g., discussion of blackness/antiblackness and how language informs such structural antagonisms, Costa Vargas & Jung, 2021; Fanon, 2008) these generative models will be useful for understanding how *genres of the Human* (Wynter, 2003; Wynter & McKittrick, 2015) may be translated as knowledge and structure within existing digital environments. Here, *genres of the Human* points to the ways in which humanity is (or is not) ascribed to certain people based on an ideal representation and how particular *genres* are grounded by existing systems of oppression. Thus, one may (for example) understand how antiblackness (anti-black oppression) may be represented and enacted within digital environments by thinking through the ways such environments may assume and police for White, Western, heteropatriarchal, cis-gendered (etc.) norms. This might be represented both directly in the text/data that is pulled from those environments, and the structures that allow those data to take certain configurations, or symbol structures.

Given that the knowledge transmitted by and within these digital environments holds an important place in cognitive memory and action (Sparrow et al., 2011) these generative models give a potentially useful opportunity to computationally explore how such digital environments represent a particular genre of the Human and how this may play out at the knowledge level to effect cognitive behavior across time. The importance becomes especially apparent when we consider the ways important and common sources of knowledge (such as search engines and social media sites), enforce at scale (whether intentional or not) certain genres of the Human (e.g., see Noble, 2018, for discussion of a particular search engine and its representation for girls/women, and especially black girls/women). Given their sources of knowledge (and that some are already populating search engines themselves Mehdi, 2023) these systems will give a unique opportunity to develop cognitive models with knowledge similar to the *everyday* (even if problematic) knowledge many are using for their cognitive behavior and actions; models that can explore what specific genres of the Human mean for behavior in various contexts.

Related work on world models and expanded knowledge for ACT-R

Though there has been less work on using generative models as world models for cognitive architectures, there has been related work on using larger knowledge sources (e.g., Salvucci, 2014; Workman & Dancy, 2023), distributed representations (e.g., Arora et al., 2018; Kelly et al., 2020), and thinking through reasons for and ways to add ontological representation to cognitive architectures for improved knowledge level representation and representative action (Halbrügge et al., 2015; Lieto et al., 2018).

Salvucci (2014) and Halbrügge et al. (2015) point to the potential usefulness of these expanded knowledge-level representation for contextual cognitive models, while the alternative distributed representation approach taken by Kelly et al. (2020) lead towards a way forward with distributed semantic symbolic representations. These vector-space embeddings are a good first step towards exploring integration of generative models, which (in the case of language models) will use embeddings as a part of the overall (transformer) architecture. Dancy (2022) and Workman and Dancy (2023) point to important potential applications of using such a cognitive architecture within ongoing (problematic) sociocultural contexts that those models/systems themselves represent. Notably Dancy (2022) doesn't pick a task which might normally be considered in a *sociocultural context*, but instead argues for the use of (ACT-R) cognitive models to begin to understand the always present effects of particular sociocultural structures on the design, development, and deployment of AI systems; such an application would certainly apply to the development process of generative models as well.

Cognitive Architectures for Generative Models

Though we've led our discussion with the ways integration of generative models would benefit cognitive architectures, there are benefits to generative models from this integration as well. Cognitive architectures tend to have good explanation and tracing capabilities given the (historical) purpose of such systems, which is useful for generative models that can have issues with explanability and having explanations that one can trust and trace. Given the issues with generative models such as large language models, enacting and mimicking problematic social structures (Bender et al., 2021),

cognitive architectures may also be used to audit these systems, especially in cases where one might be trying to understand the implications of people interacting with them.

On Improving Explainability, Traceability, and Trust

The ability to explain and trace what leads to output of larger "opaque" discriminate systems has remained important problem. Generative models present a different problem in the sense that the output tends to differ from previous, opaque deep-learning systems, but the issue of explaining and tracing process still must be addressed. There have been some recent work with cognitive modeling as a tool to help explain certain features of decisions/actions output by a deep-learning (discriminate) system (Somers et al., 2019) that could give potential pathways for using (cognitive architecture-based) cognitive models with generative models for explanability.

Somers et al. (2019) developed system that incorporated a cognitive model into a deep RL agent pipeline, so that the cognitive model might be used to provide explanations of actions by the Deep RL agent. The system worked by tracing actions/decisions by the deep-RL agent and representing those symbolic traces within the (ACT-R) cognitive model. Those representations are then used in a cognitive model, which gives an idea of *salience* of features used for given actions/decisions.

Given the ability to *generate* symbol structures themselves, generative models may see the most benefit from cognitive architecture's potential to be used for model-based knowledge tracing. When queried for information, an understanding of potential user knowledge states could help generative models produce symbols structures that would best meet users at their current state of knowledge (and potentially even affective/physiological state depending on the architecture used). Cognitive models may also be used to trace aspects of the generative model itself, potentially giving a more explainable, traceable representation to be presented to the user, similar to the approach taken Somers et al. (2019), though the ability to accomplish this would depend on expertise and the task itself and finding a suitable input of features into the cognitive models (e.g., see Somers et al., 2018, for an explanation of their approach for connecting CNN output to a cognitive model).

Recent work exploring explainable AI-related needs for generative AI systems (Sun et al., 2022) points to some of the expanded of these systems (even if this is for a particular task). Beyond just *why*, study participants wanted to understand many of the *how*'s of generative models (e.g., *how does the system work* or *how can I improve accomplish X*). Cognitive models built within cognitive architectures should prove useful in allowing a system to have a better understanding of the knowledge-level details and features important to provide for better user interaction.

Auditing Generative models.

Ethics-based AI audits (Mökander & Floridi, 2021) have become one of the typical tools to test AI systems within various ethical guidelines and to encourage trust in AI systems. The evolution of large, opaque AI systems has resulted in continued evolution of these auditing processes, including using human-AI systems to audit the more recent, very complex generative models (Rastogi et al., 2023). The use of human-AI auditing systems presents another powerful use for cognitive models built within cognitive architectures - cognitive models as humans for initial large-scale audit simulations.

This use allows for systematic exploration of existing and possible audits on a system with the simulated humans being grounded in a unified theory of cognition. Though one would likely still want actual humans farther down the testing pipeline, and there remains the risk of using such systems for further *ethics washing* (Floridi, 2019), such an application should prove to be very useful in a tool-kit of possible audits. Those simulations would be useful for exploring the increasingly complex problem spaces presented by these generative models before using the resources needed for (actual) human-AI auditing. This application would need to be applied while keeping in mind the various contexts of potential generative models use, especially the ways institutions may use generative models in ways that enact and entrench existing problematic sociocultural systems (e.g., Birhane et al., 2022).

Conclusion

The integration of generative models into cognitive architectures may provide a system that is useful both from the perspective of an improved, more representative cognitive architecture and from the perspective of improving generative models. Despite this promise, one should take pause on rushing to integrate these generative systems without understanding the related issues of these systems. Though integrations may improve sociocultural representations for cognitive models built to operate within cognitive architectures, the applications and contexts in which to use such new cognitive models should be approached with a critical lens and an understanding of historical and ongoing sociocultural structures and systems that impact the design, development, and use of generative models.

Acknowledgments

This work was supported by the National Science Foundation under grant No. 2144887.

References

Arora, N., West, R. L., Brook, A., & Kelly, M. A. (2018). Why the Common Model of the mind needs holographic a-priori categories. *Procedia Computer Science, 145*, 680-690.

Atkins, A. A., Brown, M. S., & Dancy, C. L. (2021). Examining the Effects of Race on Human-AI Cooperation. In R. Thomson, M. N. Hussain, C. L. Dancy, & A. Pyke (Ed.). *Proceedings of the 14th International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation*. Virtual, 279-288.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big. In (Ed.). *Proceedings of the 4th Conference on Fairness, Accountability, and Transparency*. Virtual Event, Canada.

Birhane, A., Ruane, E., Laurent, T., Brown, M. S., Flowers, J., Ventresque, A., & Dancy, C. L. (2022). The forgotten margins of AI ethics. In (Ed.). *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Seoul, Republic of Korea, 948–958.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183-186.

Costa Vargas, J. H., & Jung, M.-K. (2021). Antiblackness of the Social and the Human. In M.-K. Jung & J. H. Costa Vargas (Eds.), *Antiblackness*. Durham, NC: Duke University Press.

Dancy, C. L. (2021). A Hybrid Cognitive Architecture with Primal Affect and Physiology. *IEEE Transactions on Affective Computing, 12*(2), 318-328.

Dancy, C. L. (2022). Using a Cognitive Architecture to consider antiblackness in design and development of AI systems. In T. C. Stewart (Ed.). *Proceedings of the 20th International Conference on Cognitive Modeling*. Toronto, Ontario, CA, 65-72.

Dancy, C. L., & Saucier, P. K. (2022). AI and Blackness: Towards moving beyond bias and representation. *IEEE Transactions on Technology and Society, 3*(1), 31-40.

Fanon, F. (2008). *Black skin, white masks*. New York, NY, USA: Grove Press.

Floridi, L. (2019). Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy & Technology, 32*(2), 185-193.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science, 27*(4), 591-635.

Halbrügge, M., Quade, M., & Engelbrecht, K.-P. (2015). How Can Cognitive Modeling Benefit from Ontologies? Evidence from the HCI Domain. In (Ed.). *Proceedings of the 8th International Conference on Artificial General Intelligence*. Berlin, DE, 261-271.

Kelly, M. A., Arora, N., West, R. L., & Reitter, D. (2020). Holographic Declarative Memory: Distributional Semantics as the Architecture of Memory. *Cognitive Science, 44*(11), e12904.

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.

Larue, O., West, R., Rosenbloom, P. S., Dancy, C. L., Samsonovich, A. V., Petters, D., & Juvina, I. (2018). Emotion in the Common Model of Cognition. *Procedia Computer Science, 145*, 740-746.

Lieto, A., Kennedy, W. G., Lebiere, C., Romero, O. J., Taatgen, N., & West, R. L. (2018). Higher-level knowledge, rational and social levels constraints of the common model of the mind. *Procedia Computer Science, 145*, 757-764.

Mehdi, Y. (2023). https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/. Retrieved from https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences, 120*(13), e2215907120.

Mökander, J., & Floridi, L. (2021). Ethics-Based Auditing to Develop Trustworthy AI. *Minds and Machines, 31*(2), 323-327.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA, USA: Harvard University Press.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM, 19*(3), 113–126.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York, NY, USA: NYU Press.

Perrigo, B. (2023). Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. *Time*.

Prather, R. W., Benitez, V. L., Brooks, L. K. l., Dancy, C. L., Dilworth-Bart, J., Dutra, N. B., . . . Thomas, A. K. (2022). What Can Cognitive Science Do for People? *Cognitive Science, 46*(6), e13167.

Rastogi, C., Ribeiro, M. T., King, N., & Amershi, S. (2023). *Supporting Human-AI Collaboration in Auditing LLMs with LLMs*. arXiv preprint arXiv:2304.09991.

Salvucci, D. D. (2014). Endowing a Cognitive Architecture with World Knowledge. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Ed.). *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Austin, TX, USA, 1353-1358.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., . . . Others, a. (2022). *Bloom: A 176b-parameter open-access multilingual language model*. arXiv preprint arXiv:2211.05100.

Somers, S., Mistupoulos, K., Lebiere, C., & Thomson, R. (2018). Explaining Decisions of a Deep Reinforcement Learner with a Cognitive Architecture. *ACI Journal Articles, 124*.

Somers, S., Mistupoulos, K., Lebiere, C., & Thomson, R. (2019). Cognitive-Level Salience for Explainable Artificial Intelligence. In I. Juvina, J. Houpt, & C. Meyers (Ed.). *Proceedings of the 17th International Conference on Cognitive Modeling*. Montreal, QC, Canada, 235-240.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science, 333*(6043), 776-778.

Sun, J., Liao, Q. V., Muller, M., Agarwal, M., Houde, S., Talamadupula, K., & Weisz, J. D. (2022). Investigating Explainability of Generative AI for Code through Scenario-based Design. *Paper presented at the 27th International Conference on Intelligent User Interfaces*, Helsinki, Finland. https://doi.org/10.1145/3490099.3511119

Whittaker, M. (2021). The Steep Cost of Capture. *interactions, 28*(6), 50–55.

Workman, D., & Dancy, C. L. (2023). Using ACT-R to model racial biases in a semantic knowledge graph. In (Ed.). *Proceedings of the 21st International Conference on Cognitive Modeling*. Amsterdam, NL.

Wynter, S. (2003). Unsettling the Coloniality of Being/Power/Truth/Freedom: Towards the Human, After Man, Its Overrepresentation - An Argument. *CR: The New Centennial Review, 3*(3), 257-337.

Wynter, S., & McKittrick, K. (2015). Unparalleled Catastrophe for Our Species? Or, to Give Humanness a Different Future: Conversations. In K. McKittrick (Ed.), *Sylvia Wynter: On being human as praxis*. Durham, NC, USA: Duke University Press.

Young, M., Katell, M., & Krafft, P. M. (2022). Confronting Power and Corporate Capture at the FAccT Conference. In (Ed.). *Proceedings of the the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1375–1386 , numpages = 1312.