Integrating Cognitive Architectures with Foundation Models: Cognitively-Guided Few-Shot Learning to Support Trusted Artificial Intelligence

Robert H. Thomson, Nathaniel D. Bastian
Army Cyber Institute
United States Military Academy
2101 New South Post Road
West Point, NY, 10996 USA
robert.thomson@westpoint.edu, nathaniel.bastian@westpoint.edu

Abstract

We present an updated position integrating cognitive architectures into workflow by utilizing the architecture for what it does most effectively: human-like few-shot learning integrating the vast amount of data stored by foundation models. By supplementing the language-generation capabilities with the constraints of cognitive-architectures guiding prompts, it should be possible to generate more relevant output and possibly even predict when the foundation model is hallucinating. Recent advances in few-shot learning capabilities of cognitive architectures in applied domains will be discussed with some parallel capabilities described by foundation models. Just as we use research from social psychology to 'nudge' people into making informed decisions, we should be able to use cognitive architectures to 'nudge' foundation models into developing more human-relevant content.

Background

With the advent of foundation models (Bommasani et al. 2022) having caused the degree of 'strategic surprise' not seen since Russia first launched Sputnik in 1957, the research community has rushed to understand how to best apply their *prima facie* meaningful responses while understanding their limitations (e.g., hallucinating responses that are factually incorrect yet sound plausible). It is contested the degree to which foundation models actually comprehend the underlying relational structure of their inputs as opposed to just replicating the most statistically-likely response (Ma, Zhang, and Zhu 2023). What is clear, however, is that this technology is here, widely available, and has changed the trajectory of scientific inquiry into re-usable models supporting the development of generalizable artificial intelligence .

Throughout the past decade, we have argued that cognitive architectures should serve as a bridge providing common ground between human users and artificial intelligence (AI) algorithms to maximize what each does best: humans operating over well-structured knowledge and imparting wisdom, while AI finds patterns sifting through the large amounts of unstructured data (see Figure 1; Thomson, Lebiere, and Bennati 2014; Mitsopoulos et al. 2022; Somers et al. 2019; Mitsopoulos et al. 2020).

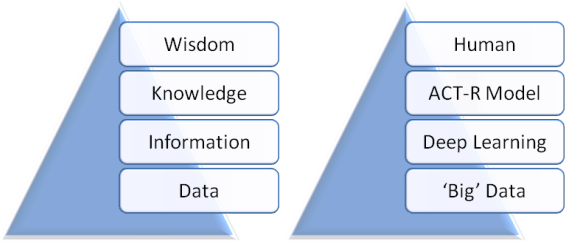

Figure 1: The original link between cognitive architectures and the DIKW pyramid described in (Thomson, Lebiere, and Bennati 2014).

With foundation models capable of generalized problem-solving, question-answering, and apparent creativity, our previous conception of the role of a cognitive architecture in this pyramid is now being challenged. Foundation models are built with sufficiently-broad information to converse in human-like language and seemingly bridge the gap directly between humans and 'big' data. They are able to complete many cognitive assessments better than undergraduate students, and they provide reasonable answers to many compare-and-contrast and synthesis-style questions in brief essay form (Ma, Zhang, and Zhu 2023).

This said, one of the challenges facing foundation models is their tendency to hallucinate responses (Bang et al. 2023), which at best make it difficult to adequately trust their responses, and at worst can lead to grave consequences for those who do over-trust their responses (Weiser and Schweber 2023). They also require substantial fine-tuning to ensure that they do not fall prey to the biases of their underlying data distributions (e.g., racism, sexism, hate-speech). There is evidence that they exhibit some cognitive biases in their patterns of responses (Abramski et al. 2023) although the newest version of foundation models, ChatGPT-4, may not to the same extent (Hagendorff and Fabi 2023).

It is unclear whether purely transformer-based foundation models sufficiently reason over their knowledge, or serve as the real-world embodiment of Searle's Chinese Room (Preston and Bishop 2002), or less charitably as the overseer of an infinite army of monkeys. Perhaps one role for cognitive architectures remains as the homuncular arbiter over foundation models' output, best understanding human operators'

intent and providing cognitively-grounded symbolic-hybrid and goal-directed oversight to provide only the right kind of information to human operators of varying skills and backgrounds?

For the remainder of this paper, we identify a concern with how fundamental research into trust in AI is being overshadowed by technical research into trustworthy AI, which limits the practical implications of integrating foundation models into user workflow (the *Clippy* issue). We then will identify several few-shot learning capabilities of both extant cognitive architectures and foundation models, arguing how each should use the other's capabilities to develop trusted AI.

Trust in AI vs Trustworthy AI

While similar-sounding, trust and trustworthiness are not identical concepts. Trustworthy AI is a feature of the AI algorithm itself, for instance being reliable, robust, accurate, fair, and explainable (Bellamy et al. 2018). Being trustworthy is not sufficient to be actually trusted, although it should ideally be a necessary component. Trust, on the other hand, is a value judgment that a person is willing to accept risk by giving up some agency (i.e., decision-making) to the AI. In prior literature, much of human-AI trust has focused on AI competency (a feature of trustworthiness) while much of inter-human trust has focused on the broader set of ability, benevolence, and integrity (Mayer, Davis, and Schoorman 1995). The differences are clear in practice: human-human trust requires an estimation of trustee's capabilities along with a moral judgment that the trustee is 'trustworthy'. In the case of the AI, however, the focus has been more on understanding the underlying competencies of a given context, while the moral and psychological predispositions get put in the broad 'human factors' box.

We argue that more in-depth analysis of the human component of human-centered AI teaming is required. Prior research has shown that humans are biased to initially undertrust AI and over-trust with more experience with the AI and need to be explicitly calibrated to the competencies of the AI (Zhang, Liao, and Bellamy 2020). We have further shown that human attentional requirements change between human-in-the-loop and human-on-the-loop scenarios (Cassenti et al. 2022) and require their own kinds of explanation (calibrated systems require only functional explanations while errors require mechanistic; (Schoenherr and Thomson 2023)). We need to not forget to apply resources devoted to the human-aspect of human-AI teaming to understand their condition for trust under various task, risk, and automation conditions.

A rule of thumb in human-factors literature is that it costs 20x more to shoehorn in human-factors after product development, but this is exactly what is happening in the AI domain right now. Proportionately, too much effort is going into developing explanation' without understanding what the human users actually require. The explanation needs to be interpretable (i.e., consumable and understandable) to the correct end-user. This goes beyond interface design principles and requires better understanding of the nature of human-AI trust and how to ensure that a system maintains appropriate trust. Extant research in explainable AI and uncertainty quantification should provide the basis for trustworthiness, but we further need to understand how the presentation of this information impacts - and possibly biases - humans to respond (*inappropriately*).

A first effort would be apply Gricean maxims to AI communication with the human. Grice proposed four maxims to understand the implication of what was said. These maxims are: quality of information, quantity of information, the manner that it is communicated, and the relevance to the question (Liang et al. 2019). In practice, this means getting the right information and right amount of information to the operator, and only when they require it. This is simpler said than done. Microsoft's great failure was the integration of 'Clippy' into Office, and an inappropriately-targeted AI will just be seen as an ineffective Clippy 2.0 (Gruber 2018).

Few-Shot Learning

Few-shot learning (FSL) defines the concept of training a machine learning model where there is limited labeled data available for training. The goal of FSL for classification is to train a model that can accurately classify observations for both classes present in the labeled training data and unseen classes present in an unlabeled support set. To achieve this, FSL typically utilizes a meta-learning framework where the model learns to adapt quickly to new classes based on a limited number of labeled support instances. The model's adaptation is guided by the knowledge gained from the seen classes in the labeled training dataset.

Few-Shot Learning in Cognitive Architectures

One of the primary capabilities that sets humans apart is our ability to learn new domains and pick up new skills with few-shot or even zero-shot examples. Based on a combination of innate capabilities and broad knowledge, we are able to reason inductively and deductively. Cognitive architectures are also able to reason successfully with limited training exemplar via a combination of experience (e.g., instance-based learning; Gonzalez, Lerch, and Lebiere 2003) and architectural structure (e.g., base-level learning and similarity; Anderson et al. 2004). Cognitive architectures have been successful at achieving human-level performance via experience at complex games such as backgammon (Sanner et al. 2000) playing only hundreds of games (compared with deep learning techniques which generally require thousands to millions of games).

For modeling human decision-making, (Lebiere et al. 2013) were able to predict the performance and expression of trial-by-trial expression of eight cognitive biases from intelligence analysts sense-making over geospatial intelligence data across a range of tasks. By learning from feedback on previous trials, the model were able to predict whether the analyst would be risky or risk averse and the consequent biases which arose.

Adapting the few-shot learning technique from (Lebiere et al. 2013), (Nunes et al. 2015) was able to rapidly develop a model of malware identification achieving high performance (unbiased F1 score $\geq$.9) even when trained on only 1/10th of

the data. More recently, (Thomson, Cranford, and Lebiere 2021) adapted this framework to network security, specifically intrusion detection. Using the UNSW-NB15 dataset training data, their instance-based model was able to achieve comparable multi-class performance to standard ML techniques (e.g., decision tree, random forest) learning as few as one instance of each of nine attack categories and nine examples of normal computer network traffic. For reference, the training dataset contains 175k instances.

The strength of cognitive architectures is that they use approximations of human memory and reasoning capabilities to enable few-shot learning across a range of possibly tasks/applications, and they can be designed such that their output is cognitively-plausible and can be tuned to individual capabilities and/or preferences. A limitation to architectures using instance-based learning as a framework is that they are not scalable without some extension to keep the number of instances manageable, requiring some mechanism for discarding, joining, and/or ignoring instances when searching memory.

Few-Shot Capabilities of Foundation Models

One of the benefits of foundation models is that they come pre-trained on vast amount of text (GPT-3) and multi-modal (GPT-4) data. This large knowledge background and statistical prediction capability give them the ability to create plausible sounding documents, computer code, and images from only a few prompts (Kojima et al. 2022; Baldazzi et al. 2023; Ahmed and Devanbu 2022; Li et al. 2023), or even play a Minecraft agent learning new rules via prompting (Wang et al. 2023). In terms of several recent successes, (Buchholz 2023) was able to use prompting to conduct near state-of-the-art few-shot detection of false political statements on the LIAR dataset. (Roy et al. under review) has used foundation models to do few-shot detection of out-of-context images.

This said, there are limits as these models are non-deterministic and may not be sufficiently reliable for human-on-the-loop or automated tasks (Reiss 2023) where consistent responses are required. As previously mentioned, foundation models also have a tendency to hallucinate and while there are some efforts to remediate this (Jha et al. 2023), this remains a concern.

Current Limits of Foundation Models

There is a practical limit to how much more foundation models can grow. At present levels of growth, it is forecasted that the cost to train foundation models will exceed 2.2% of the United States' GDP between 2025-2032 and exceed GDP between 2027-2036 (Li 2023). There is thus a renewed focus on training relatively smaller models supplemented by external knowledge bases, which are capable of comparable performance to larger models in particular domains. For instance, the retrieval-enhanced transformer (RETRO) leverages an external corpus of documents to train a model with performance comparable to much larger models (Jurassic-1 178B, Gopher, 280B) using 25x fewer parameters (RETRO has 7B; Borgeaud et al. 2022; Bommasani et al. 2022). This does come at relatively higher training requirements, requiring upwards of 10x the training data similarly-sized smaller models, however it can be leveraged to limit hallucination and tying to preexisting corpi or ontologies can make the models more explainable as it is possible for them to point to their source material (e.g., what document(s) fed into their response).

Extending Cognitive Architectures

There have been numerous techniques to improve the scalability of cognitive architectures to support more complex decision-making and run stably over longer time-courses. (Oltramari and Lebiere 2012a,b) have utilized ontologies to structure larger off-line memory stores as a readily-searched database, while the ACT-R/E architecture (Trafton et al. 2013) has a built-in mechanism to move memory between an activated declarative memory and 'cold storage' memories which could never be retrieved without seeing them again, to keep memory search scalable with millions of instances in memory. (Rutledge-Taylor et al. 2014) developed a holographic memory system that replaces the symbolic declarative memory in ACT-R into a vector-based approach more compatible with deep learning models. Along a similar trajectory, (Vinokurov et al. 2011, 2012, 2013) integrated ACT-R as a hybrid-symbolic component with the Leabra connectionist cognitive-architecture to support metacognitive reasoning and higher-level visual processing (called SAL; the *Synthesis of ACT-R and Leabra*), and was further able to play games in the Unreal engine (Jilk et al. 2008).

Integrating Foundation Models and Cognitive Architectures

As we have seen, both foundation models and cognitive architectures have exhibited success at few-shot learning, but they do so with different strengths. Foundation models have processed large amounts of unstructured data and are amazing statistical generators/predictors using this data. Models built from cognitive architectures utilize cognitively-inspired mechanisms (e.g., base-level learning, similarity) to drive decision-making without the need for large amounts of training data.

One way where foundation models may support cognitive architectures is in their ability to synthesize large amount of information and act as a knowledge store, solving many of the scalability issues in declarative memory. For instance, (Trajanoska, Stojanov, and Trajanov 2023) were able to enhance traditional knowledge graphs via foundation models, and were able to use foundation models to conduct automated ontology creation. These ontologies would distill *data* into *information* under DIKW pyramid and make it more accessible for the kinds of higher-order reasoning that cognitive architectures excel at performing.

Cognitive architectures may support foundation models by providing a cognitive grounding which we argue may catch hallucinating foundation models, although this is still primarily speculation. Cognitive models, by virtue of their ability to predict human performance (and preferences) with few-shot learning may be an ideal source to assist foundation models in further personalizing their responses to the human user. This personalization would go a long way to

support trust in AI-based models.

Concluding Thoughts

Both cognitive architectures and foundation models are capable of few-shot learning, but they do so by distinct mechanisms. With further investigation into how these mechanisms could be integrated together, we argue that it is possible to get the best of both worlds: models that come pretrained with large amounts of background information that allow for scalable and adaptive cognitive models capable of personalizing content and prediction a broader range of human behaviors.

Future work needs to further explore this integration in the form of neuro-symbolic AI, where learning and inference exploit symbolic knowledge and reasoning. For example, (Abdelzaher et al. 2022) propose a multi-layered neuro-symbolic architecture inspired by Predictive Processing (PP) - a theory of mind, will enable context-aware, few-shot AI models with tight integration between symbolic reasoning (via cognitive architectures) and deep learning (via foundation models). The PP-inspired architecture relies on building a "world model" that captures context and uses this context to hypothesize and confirm predictions over the data. Together, they form the preconditions to achieve trusted AI.

Acknowledgments

This research was supported in part by the Office of Naval Research FY21 Multi-University Research Initiative Award N0001422MP00465 and N0001423MP00219, as well the Defense Advanced Research Projects Agency under Support Agreement No. USMA 23004. The views expressed in this work are those of the authors and do not necessarily reflect the official policy or position of the United States Military Academy, Department of the Army, Office of Naval Research, Department of Defense, or U.S. Government.

References

Abdelzaher, T.; Bastian, N. D.; Jha, S.; Kaplan, L.; Srivastava, M.; and Veeravalli, V. V. 2022. Context-aware Collaborative Neuro-Symbolic Inference in IoBTs. In *MILCOM 2022 - 2022 IEEE Military Communications Conference (MILCOM)*, 1053–1058.

Abramski, K.; Citraro, S.; Lombardi, L.; Rossetti, G.; and Stella, M. 2023. Cognitive Network Science Reveals Bias in GPT-3, GPT-3.5 Turbo, and GPT-4 Mirroring Math Anxiety in High-School Students. *Big Data and Cognitive Computing*, 7(3): 124.

Ahmed, T.; and Devanbu, P. 2022. Few-shot training LLMs for project-specific code-summarization. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, 1–5.

Anderson, J. R.; Bothell, D.; Byrne, M. D.; Douglass, S.; Lebiere, C.; and Qin, Y. 2004. An integrated theory of the mind. *Psychological review*, 111(4): 1036.

Baldazzi, T.; Bellomarini, L.; Ceri, S.; Colombo, A.; Gentili, A.; and Sallinger, E. 2023. Fine-tuning Large Enterprise Language Models via Ontological Reasoning. *arXiv preprint arXiv:2306.10723*.

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Bellamy, R. K.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N.; Chen, A.; Creel, K.; Davis, J. Q.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Fei-Fei, L.; Finn, C.; Gale, T.; Gillespie, L.; Goel, K.; Goodman, N.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D. E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P. W.; Krass, M.; Krishna, R.; Kuditipudi, R.; Kumar, A.; Ladhak, F.; Lee, M.; Lee, T.; Leskovec, J.; Levent, I.; Li, X. L.; Li, X.; Ma, T.; Malik, A.; Manning, C. D.; Mirchandani, S.; Mitchell, E.; Munyikwa, Z.; Nair, S.; Narayan, A.; Narayanan, D.; Newman, B.; Nie, A.; Niebles, J. C.; Nilforoshan, H.; Nyarko, J.; Ogut, G.; Orr, L.; Papadimitriou, I.; Park, J. S.; Piech, C.; Portelance, E.; Potts, C.; Raghunathan, A.; Reich, R.; Ren, H.; Rong, F.; Roohani, Y.; Ruiz, C.; Ryan, J.; Ré, C.; Sadigh, D.; Sagawa, S.; Santhanam, K.; Shih, A.; Srinivasan, K.; Tamkin, A.; Taori, R.; Thomas, A. W.; Tramèr, F.; Wang, R. E.; Wang, W.; Wu, B.; Wu, J.; Wu, Y.; Xie, S. M.; Yasunaga, M.; You, J.; Zaharia, M.; Zhang, M.; Zhang, T.; Zhang, X.; Zhang, Y.; Zheng, L.; Zhou, K.; and Liang, P. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258.

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G. B.; Lespiau, J.-B.; Damoc, B.; Clark, A.; et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, 2206–2240. PMLR.

Buchholz, M. G. 2023. Assessing the Effectiveness of GPT-3 in Detecting False Political Statements: A Case Study on the LIAR Dataset. *arXiv preprint arXiv:2306.08190*.

Cassenti, D.; Roy, A.; Hawkins, T.; and Thomson, R. 2022. The Effect of Varying Levels of Automation during Initial Triage of Intrusion Detection. In *Artificial Intelligence and Social Computing*.

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

Gruber, D. 2018. *The effects of mid-range visual anthropomorphism on human trust and performance using a navigation-based automated decision aid*. Ph.D. thesis, University of Minnesota.

Hagendorff, T.; and Fabi, S. 2023. Human-Like Intuitive Behavior and Reasoning Biases Emerged in Language Models–and Disappeared in GPT-4. *arXiv preprint arXiv:2306.07622*.

Jha, S.; Jha, S.; Lincoln, P.; Bastian, N.; Velasquez, A.; and Neema, S. 2023. Dehallucinating Large Language Models Using Formal Methods Guided Iterative Prompting. In *Proceedings of IEEE International Conference on Assured Autonomy (ICAA) 2023*.

Jilk, D. J.; Lebiere, C.; O'Reilly, R. C.; and Anderson, J. R. 2008. SAL: An explicitly pluralistic cognitive architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, 20(3): 197–218.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Lebiere, C.; Pirolli, P.; Thomson, R.; Paik, J.; Rutledge-Taylor, M.; Staszewski, J.; and Anderson, J. R. 2013. A functional model of sensemaking in a neurocognitive architecture. *Computational intelligence and neuroscience*, 2013: 5–5.

Li, T.; Shetty, S.; Kamath, A.; Jaiswal, A.; Jiang, X.; Ding, Y.; and Kim, Y. 2023. Cancergpt: Few-shot drug pair synergy prediction using large pre-trained language models. *arXiv preprint arXiv:2304.10946*.

Li, Y. 2023. Business Analysis — AI Computational Cost. Https://medium.com/geekculture/business-analysis-ai-computational-cost-67a136957c95.

Liang, C.; Proft, J.; Andersen, E.; and Knepper, R. A. 2019. Implicit communication of actionable information in human-ai teams. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.

Ma, Y.; Zhang, C.; and Zhu, S.-C. 2023. Brain in a Vat: On Missing Pieces Towards Artificial General Intelligence in Large Language Models. *arXiv preprint arXiv:2307.03762*.

Mayer, R. C.; Davis, J. H.; and Schoorman, F. D. 1995. An integrative model of organizational trust. *Academy of management review*, 20(3): 709–734.

Mitsopoulos, K.; Somers, S.; Schooler, J.; Lebiere, C.; Pirolli, P.; and Thomson, R. 2022. Toward a psychology of deep reinforcement learning agents using a cognitive architecture. *Topics in Cognitive Science*, 14(4): 756–779.

Mitsopoulos, K.; Somers, S.; Thomson, R.; and Lebiere, C. 2020. Cognitive architectures for introspecting deep reinforcement learning agents. In *Workshop on Bridging AI and Cognitive Science, at the 8th International Conference on Learning Representations*. ICLR.

Nunes, E.; Buto, C.; Shakarian, P.; Lebiere, C.; Bennati, S.; Thomson, R.; and Jaenisch, H. 2015. Malware task identification: A data driven approach. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, 978–985.

Oltramari, A.; and Lebiere, C. 2012a. Knowledge in action: Integrating cognitive architectures and ontologies. In *New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems*, 135–154. Springer.

Oltramari, A.; and Lebiere, C. 2012b. Using ontologies in a cognitive-grounded system: automatic action recognition in video surveillance. In *Proceedings of the 7th International Conference on Semantic Technology for Intelligence, Defense, and Security*. Citeseer.

Preston, J.; and Bishop, M. 2002. *Views into the Chinese room: New essays on Searle and artificial intelligence*. Oxford University Press.

Reiss, M. V. 2023. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085*.

Roy, A.; Cobb, A.; Jha, S.; Bastian, N.; Cruickshank, I.; Berenbeim, A.; Thomson, R.; Velasquez, A.; and Jha, S. under review. Zero-shot Detection of Out-of-Context Objects Using Foundation Models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Rutledge-Taylor, M. F.; Kelly, M. A.; West, R. L.; and Pyke, A. A. 2014. Dynamically structured holographic memory. *Biologically Inspired Cognitive Architectures*, 9: 9–32.

Sanner, S.; Anderson, J. R.; Lebiere, C.; and Lovett, M. C. 2000. Achieving Efficient and Cognitively Plausible Learning in Backgammon. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 823–830.

Schoenherr, J.; and Thomson, R. 2023. When AIs Fail, Who Do We Blame? Attributing Responsibility in Human-AI Interactions. Submitted to IEEE Transactions on Technology and Society.

Somers, S.; Mitsopoulos, K.; Lebiere, C.; and Thomson, R. 2019. Cognitive-level salience for explainable artificial intelligence. In *Proceedings of the 17th Annual Meeting of the International conference on Cognitive Modeling*.

Thomson, R.; Cranford, E.; and Lebiere, C. 2021. Achieving active cybersecurity through agent-based cognitive models for detection and defense. In *Proceedings of the Autonomous Intelligent Cyber-defence Agent Workshop*.

Thomson, R.; Lebiere, C.; and Bennati, S. 2014. Human, model and machine: a complementary approach to big data. In *Proceedings of the 2014 Workshop on Human Centered Big Data Research*, 27–31.

Trafton, J. G.; Hiatt, L. M.; Harrison, A. M.; Tamborello, F. P.; Khemlani, S. S.; and Schultz, A. C. 2013. Act-r/e: An embodied cognitive architecture for human-robot interaction. *Journal of Human-Robot Interaction*, 2(1): 30–55.

Trajanoska, M.; Stojanov, R.; and Trajanov, D. 2023. Enhancing Knowledge Graph Construction Using Large Language Models. *arXiv preprint arXiv:2305.04676*.

Vinokurov, Y.; Lebiere, C.; Herd, S.; and O'Reilly, R. 2011. A metacognitive classifier using a hybrid ACT-R/Leabra architecture. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Vinokurov, Y.; Lebiere, C.; Szabados, A.; Herd, S.; and O'Reilly, R. 2013. Integrating top-down expectations with bottom-up perceptual processing in a hybrid neural-symbolic architecture. *Biologically Inspired Cognitive Architectures*, 6: 140–146.

Vinokurov, Y.; Lebiere, C.; Wyatte, D.; Herd, S.; and O'Reilly, R. 2012. Unsurpervised learning in hybrid cognitive architectures. In *Workshops at the twenty-sixth AAAI conference on artificial intelligence*.

Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Weiser, B.; and Schweber, N. 2023. The ChatGPT Lawyer Explains Himself. *The Guardian*, Retrieved 8/4/23. Https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html.

Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 295–305.