Toward Artificial Metacognition

Paulo Shakarian

Syracuse University
Syracuse, NY
pashakar@syr.edu

Abstract

The research trend of metacognitive AI deals with the study of artificial intelligence systems that can self-monitor and/or regulate resources. This concept has its roots in cognitive psychology studies on human metacognition. It has led to the understanding of how people monitor, control, and communicate their cognitive processes. An emerging research trend in artificial intelligence is to build systems that possess these capabilities. This paper summarizes the key ideas about metacognition from cognitive psychology, describes recent attempts to instantiate these concepts in AI systems, and discusses metacognitive capabilities observed in humans that are not thoroughly explored in AI research.

Resource page — https://metacognition.syracuse.edu/

1 Introduction

The research trend of *metacognitive artificial intelligence* deals with the study of AI systems that can self-monitor and/or regulate resources. This line of work is inspired by ideas from cognitive psychology on metacognition (Henmon 1911; Flavell 1979; Nelson 1990; Metcalfe and Shimamura 1994) which in is defined as "the processes that monitor our ongoing thought processes and control of the allocation of mental resources"(Ackerman and Thompson 2017). Metacognition is a powerful concept regarded by some as a self-monitoring process that is integral to the functioning of the human mind (Demetriou et al. 1993). It has influenced thinking regarding two philosophical challenges closely related to artificial intelligence - namely consciousness (Nelson 1996) and the symbol grounding problem (Taddeo and Floridi 2007).

While the study of artificial metacognition from a computational perspective is not new (e.g., see (Cox 2005; Cox and Raja 2011) for a summary of earlier work) it has received renewed attention in light of the recent advances in AI. This paper describes and synthesizes this emerging trend. In particular, in late 2023, an interdisciplinary group of researchers were brought together in the *Workshop on Metacognitive Prediction of AI Behavior*, the results of which are reported in (Shakarian and Wei 2025). In mid 2025, a second iteration of the workshop was held in conjunction with the SIAM Data Mining conference (Shakarian et al. 2025). In addition to the interdisciplinary perspectives of these workshops, there has been also exploration of the topics at several AI conferences, relating to concepts such as language (Shinn et al. 2023), vision (Zhang et al. 2025), and data mining (Kricheli et al. 2024).

In this paper, we provide a synthesis of this emerging trend, inspired by the interdisciplinary nature of the aforementioned workshops and the ensuing discussions. In Section 2 we review the requisite concepts from cognitive psychology. Using these concepts as a guide, we will then discuss recent trends in metacognitive monitoring (Section 3) and metacognitive architectures (Section 4). We will conclude the paper by discussing avenues for future inquiry in Section 5.

2 Cognitive Psychology Concepts

Here we provide a brief synopsis of some relevant frameworks for metacognition from the field of First, consider the bi-level paradigm of Nelson and Narens (Nelson and Narens 1994), where cognition consists of *object-level* and *meta-level* processes. Object-level processes include tasks such as perception, learning, reasoning, and planning while meta-level processes monitor and assess the object-level processes. Most of the work done by AI systems today as focused on the object-level. Augmenting AI systems with by working at the meta-level is our focus.

There are several ways to think about meta-level processes, but perhaps the most natural way stems from the widely-used dual-process theory used to classify more general cognitive processes (Wason and Evans 1974) later popularized as "System 1 - System 2"[1] or (colloquially) as "thinking fast-thinking slow" (Kahneman 2012). In AI, we tend to think of the System 1 - System 2 dichotomy as the divide between fast perception/System 1 (e.g., a vision model recognizing mathematical symbols) and slower reasoning/System 2 (e.g., solving a calculus problem). In seminal paper on dual process theory (Evans and Stanovich 2013), the authors identify metacognition as the mecha-

[1]Note that later work uses the term 'Type 1 - Type 2' (Evans and Stanovich 2013) as there can be more than one system at each level. We use "system" as it is the more common term.

nism by which incorrect perceptions (obtained quickly by System 1) are overridden by a "reflexive mind" - a System 2 process for deeper reasoning. This aligns with a dichotomy in metacognition that is similar to dual process theory as originally noted by (Flavell 1979) where metacognitive processes are identified as "automatic" and "deliberate". Here "automatic" metacognition entails the emergence of metacognitive "cues"(Ackerman 2019) – heuristics that indicate provide information about the quality of the cognitive action. For example, Feeling of Rightness (FoR) (Thompson, Prowse Turner, and Pennycook 2011) refers to a person's intuition of the answer correctness. Other examples include Feeling of Knowing (FoK) (Reder and Ritter 1992) and Expectation Violation (Anderson and Fincham 2014), among others. These cues, which occur quickly in the subject with no immediate explanation differ from deliberate metacognition (i.e., the "reflexive" processes noted by (Evans and Stanovich 2013)) that serves various purposes with two principle activities including the communication of cognitive state (Shea et al. 2014), seeking help (Undorf, Livneh, and Ackerman 2021), and regulation of time investment (Ackerman 2013; Ackerman and Undorf 2017; Toplak, West, and and 2014).

Another important dichotomy is metacognitive monitoring and metacognitive control (Ackerman and Thompson 2017). Monitoring deals with the assessment of object-level tasks while control deals with with the allocation of cognitive resources for a given task. Various researchers have identified how these two types of metacognition work together (Thompson 2009; Ackerman 2013; Shea et al. 2014). A common feature in these paradigms is the triggering of a metacognitive process. Here is where cues can often play a role. However, metacognitive monitoring can involve more than the cues of automatic (System 1) metacognition. For example it may require some deliberate (System 2) metacognition to transform cues into a representation to communicate with other individuals (Shea et al. 2014). Once the performance is assessed, further reasoning relating to metacognitive control can be performed which can be thought of a "System 2 Intervention" with respect to perceptual tasks (Thompson 2009) or an evolution of a solution strategy for reasoning tasks (Ackerman 2013).

In what follows, we first look at aspects of metacognitive monitoring realized in computational systems, then describe several recent attempts at architectural approaches. This is followed by examining less-studied areas such as metacognitive control and reasoning - which are important areas for future work.

3 Metacognitive Monitoring

Metacognitive monitoring - the use of cues on some System 1 output to trigger a reflection process - is perhaps the most natural form of metacognition to implement computationally. Here the System 1 is often a machine learning model, the cue is some type of detector, and the reflection process is an attempt to update the model output to a more correct response. Neural attention for symbolic reasoning (NASR) (Cornelio et al. 2022) is an example of this paradigm. In that paper, a reasoning problem (Soduku) is attempted by a neural model, producing a symbolic output. This output is then evaluated by another neural model to identify components that are erroneous and are then corrected by an application-specific symbolic solver. This represents one approach to a metacognitive cue - a model trained to identify errors. This neural model is similar to the intuitive metacognitive cues of "Feeling of Rightness" observed in cognitive psychology experiments. Notice that both the intuitions observed in psychology experiments and the machine learning evaluator of NASR are both black boxes. The idea with both a machine learning based cue and an intuitive cue is not to address the problem, but rather provide insight in that something is incorrect.

Just as the cognitive psychologists have identified different cues in human subjects, there have been different approaches to creating metacognitive cues in AI research. The creation of cues to identify errors is probably the area where the largest advances have been made. Here we outline a few clear categories of this research.

1. *Models that detect an error state.* The strategy here is given a perceptual result (e.g., the result of an image classification model), an auxiliary model, trained on the same training data, can predict errors. Early work in this include (Daftry et al. 2016) which is used to predict vision errors - perhaps the most straight-forward strategy for creating such a model. The aforementioned NASR (Cornelio et al. 2022) trained a model to predict reasoning errors, here erroneous data was generated to train the model. Abductive learning with new concepts (Huang et al. 2023) instead relies on an out-of-distribution detector to identify errors.
2. *Use of an alternative model for the same task.* In this approach, in addition to the initial model trained for the task, one or more alternative models is trained for the same task (on the same data), the results of those models is treated as a feature to learn an error detection module. For example, in (Lee et al. 2024) the authors trained a large number of auxiliary neural models of various architectures on the same data for the same task and then learned rules as to which of the auxiliary models providing a different result is indicative of an error in the primary model.
3. *Critique Models.* Critique models have become a prevalent source of cues in large language models (Shinn et al. 2023; Xiong et al. 2025; Zhang et al. 2025; Yang et al. 2025). Often these models are trained on errors of LLM output (to include reasoning paths), for example, for example, Critic-V (Zhang et al. 2025) is trained on degraded GPT-4o reasoning paths. A key aspect that differentiates this approach from approaches 1-2 above is that the critique model not only is identifying possible errors, but also provides natural language feedback - which can be employed as a gradient signal for further refinement.
4. *Consistency-based approaches.* For models providing a symbolic output, there is the potential for checking the consistency of that output (i.e. the use of a verifier), as that too could lead to errors. This has been exam-

ined based on human-created requirements (Yang, Neary, and Topcu 2024), but have also been examined from the standpoint of rules learned from the training data about configurations of label assignments relating to error (Kricheli et al. 2024).

While cues can lead to improved performance, this is not necessarily the case, even when a cue can detect errors with high accuracy. For example, in a multi-label classification problem, even a cue that can very accurately identify an error may not lead to a correction that can provide improved overall model accuracy (i.e., see Theorem 4.2 of (Shakarian, Simari, and Bastian 2025)). However, that theoretical result lies on a notion of reclassification - a relatively simple process. As mentioned earlier, from a psychological perspective (e.g., (Evans and Stanovich 2013; Ackerman and Thompson 2017)), a cue is viewed as a computationally inexpensive trigger that leads to more resources intensive mental processes. Viewed this way, the effectiveness of a cue is perhaps best judged in the context of a larger system. For example, the recent work on LLM metacognition (Yang et al. 2025) leverages cues that identify errors with fairly low precision of 0.488 yet leads to high accuracy for the overall problem results (0.877). As the community gain clarity on reference architecture(s) suitable for metacognition, it may then become more useful to research cues by themselves, as they can be evaluated in the context of an overall metacognitive approach. We discuss architectures in the next section.

4 Metacognitive Architectures

There have been several proposed architectures for artificial metacognition. Such work has generally focused on meeting practical desiderata such at the proposed TRAP (Transparency, Reasoning, Adaptability, and Perception) criteria (Wei et al. 2024). From a more cognitively inspired perspective, researchers in the cognitive modeling community have proposed a series of conjectures related to the creation of cognitively plausible metacognitive framework that would extend the ACT-R framework (Lebiere et al. 2025). Likewise, there are recent proposed extensions to the standard model of cognition (Laird, Lebiere, and Rosenbloom 2017) that add metacognitive capabilities (Laird et al. 2025).

Implemented architectures have varied greatly in approach and capabilities. One category includes architectures focused on correcting perceptual outputs. We can think of this as a "System 1" correction, e.g., the type of correction we do as people when we mistake one person for another. Some of this work has focused on model training how a metacognitive signal can be used to retrain a model - for example (Taparia et al. 2025) leverages VLM output as metacognitive cues in the training of a vision model while (Sagar, Taparia, and Senanayake 2024) uses reinforcement learning to characterize the failure landscape allowing for fine-tuning. Likewise, hyperdimensional computing (HDC) (Kanerva 2009) has also been studied from a metacognitive perspective by creating HDC-based frameworks to leverage axillary models on residual errors. A different approach to metacognitive training, similar to the earlier-mentioned NASR, is the use of abductive inference to correct perceptual errors (Dai et al. 2019). Recent work in this area leverages network activations to quickly identify perceptual errors and cue abductive inference (Hu et al. 2025).

At inference time, error detection rules provide a framework for employing a variety of different metacognitive cues focused on detection (Lee et al. 2024), learning of constraints (Kricheli et al. 2024), and correction (Xi et al. 2025). The use of rules as a data structure to manage metacognitive cues has a few advantages. In an application to LLM correction, "Error Detection and Correction for Interpretable Mathematics" (EDCIM) (Yang et al. 2025) leverages the explainability of metacognitive rules to prompt a more powerful LLM to make corrections. As with training, abductive inference can also be used at test-time, leveraging error detection rules to ensemble models in out-of-distribution environments (Leiva et al. 2025).

The architectures described in this section represent a step toward realizing a metacognitive AI system. However, most of the work is relegated to improving perceptual tasks or LLM output, as opposed to controlling cognitive resources - the cognitive resource regulation observed in humans (Evans and Stanovich 2013; Ackerman and Thompson 2017). We examine this as part of future inquiry in the next section.

5 Avenues for Future Inquiry

Cognitive psychologists have referred to humans as "cognitive misers" (Toplak, West, and and 2014) due to their reluctance to allocate cognitive resources. As mentioned earlier, "cognitive control" refers to the ability of humans to self-regulate resources, in particular working memory, for a given task. Further, in reasoning tasks, humans are observed to repeatedly make decisions on whether to continue with the tasks using special metacognitive cues such as "judgment of solvability" (JoS). This is referred to as "metareasoning" (Ackerman and Thompson 2017). While the architectures of the previous section have made advances toward metacognition, there is very little work on cognitive control and metareasoning.

Some of the work does tend toward this direction though. Critic-V (Zhang et al. 2025) performs rudimentary metareasoning by examining reasoning traces from LLM output, but does not leverage this information in metacognitive control. Meanwhile, EDCIM (Yang et al. 2025) does provide a form a metacognitive control, calling to a more sophisticated LLM when encountering error cues, but is doing so primarily in the transformation of text to math equations - not for reasoning. Federated learning (McMahan et al. 2017) also provides a potential framework for metacognitive control, as tradeoffs between accuracy, efficiency, and security (Zhang et al. 2023; Yan et al. 2023) and cues relating to anomalies, reputation, and trust are already being studied in this paradigm (Chuprov, Memon, and Reznik 2023; Chuprov, Bhatt, and Reznik 2023).

There are other research issues that would require further study to advance these aims. In particular, a need for datasets and benchmarks. Some recent work such as Multiple Distribution Shift – Aerial (MDS-A) dataset (Ngu et al. 2025) and the Natural Robustness Toolkit (NRTK) (Kitware 2025) are

steps in this direction. Evaluation is also a challenge, and has been a topic of discussion in recent workshops (Lanus and Freeman 2025).

Additionally, several other subfields of AI have the potential to contribute to our understanding of artificial metacognition. For example, belief revision (Darwiche and Pearl 1997) may inform correction of a knowledgebase based on what we would call a "consistency-based" metacognitive monitoring approach. In particular recent work combining ideas from belief revision with machine learning (Schwind et al. 2025; Aravanis 2025) can provide insights into a framework for the development of a deliberate metacognitive system (e.g., based on belief revision) based on automatic metacognitive cues (e.g., based on detection of inconsistency). Another area of AI is the work on uncertainty estimation (Cui, Mouchel, and Faltings 2025; Mena, Pujol, and Vitrià 2021; He et al. 2025). We can think of these methods as also providing a form of metacognitive cue - what we earlier referred to as "models that detect an error state."

As a closing thought, the idea of using metacognition to communicate internal cognitive state across agents, as discussed in (Shea et al. 2014) may provide a longer-term goal for metacognitive AI research. If such a line of research is successful, it would lead to agents with the ability to not only conduct metacognitive monitoring and use that information for cognitive control, but also provide a representation of that information to other agents to accomplish a goal in a multi-agent setting - which would likely mark significant progress in the field of AI.

Acknowledgments

This research was supported by Army Research Office (ARO) grant W911NF-24-1-0007.

References

Ackerman, R. 2013. The Diminishing Criterion Model for Metacognitive Regulation of Time Investment. *Journal of Experimental Psychology: General*.

Ackerman, R. 2019. Heuristic Cues for Meta-Reasoning Judgments: Review and Methodology. *Psychological Topics*, 28.

Ackerman, R.; and Thompson, V. A. 2017. Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. *Trends in Cognitive Sciences*, 21(8): 607–617.

Ackerman, R.; and Undorf, M. 2017. The puzzle of study time allocation for the most challenging items. *Psychonomic Bulletin & Review*, in press.

Anderson, J. R.; and Fincham, J. M. 2014. Extending problem-solving procedures through reflection. *Cognitive Psychology*, 74: 1–34.

Aravanis, T. 2025. Towards machine learning as AGM-style belief change. *International Journal of Approximate Reasoning*, 183: 109437.

Chuprov, S.; Bhatt, K. M.; and Reznik, L. 2023. Federated learning for robust computer vision in intelligent transportation systems. In *2023 IEEE Conference on Artificial Intelligence (CAI)*, 26–27. IEEE.

Chuprov, S.; Memon, M.; and Reznik, L. 2023. Federated learning with trust evaluation for industrial applications. In *2023 IEEE Conference on Artificial Intelligence (CAI)*, 347–348. IEEE.

Cornelio, C.; Stuehmer, J.; Hu, S. X.; and Hospedales, T. 2022. Learning where and when to reason in neuro-symbolic inference. In *The Eleventh International Conference on Learning Representations*.

Cox, M. T. 2005. Metacognition in computation: A selected history. In *AAAI Spring Symposium: Metacognition in Computation*, 1–17.

Cox, M. T.; and Raja, A. 2011. *Metareasoning: Thinking about thinking*. MIT Press.

Cui, S.; Mouchel, L.; and Faltings, B. 2025. Uncertainty in Causality: A New Frontier. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8022–8044. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Daftry, S.; Zeng, S.; Bagnell, J. A.; and Hebert, M. 2016. Introspective Perception: Learning to Predict Failures in Vision Systems. .

Dai, W.-Z.; Xu, Q.; Yu, Y.; and Zhou, Z.-H. 2019. Bridging machine learning and logical reasoning by abductive learning. *NeurIPS*, 32.

Darwiche, A.; and Pearl, J. 1997. On the logic of iterated belief revision. *Artificial Intelligence*, 89(1): 1–29.

Demetriou, A.; Efklides, A.; Platsidou, M.; and Campbell, R. L. 1993. The architecture and dynamics of developing mind: Experiential structuralism as a frame for unifying cognitive developmental theories. *Monographs of the society for research in child development*, i–202.

Evans, J. S. B.; and Stanovich, K. E. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3): 223–241.

Flavell, J. H. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10): 906.

He, W.; Jiang, Z.; Xiao, T.; Xu, Z.; and Li, Y. 2025. A Survey on Uncertainty Quantification Methods for Deep Learning. arXiv:2302.13425.

Henmon, V. A. C. 1911. The relation of the time of a judgment to its accuracy. *Psychological Review*, 18(3): 186–201.

Hu, W.-C.; Dai, W.-Z.; Jiang, Y.; and Zhou, Z.-H. 2025. Efficient Rectification of Neuro-Symbolic Reasoning Inconsistencies by Abductive Reflection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16): 17333–17341.

Huang, Y.-X.; Dai, W.-Z.; Jiang, Y.; and Zhou, Z.-H. 2023. Enabling Knowledge Refinement upon New Concepts in Abductive Learning.

Kahneman, D. 2012. *Thinking, Fast and Slow*. London: Penguin. ISBN 9780141033570, 0141033576.

Kanerva, P. 2009. Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with

High-Dimensional Random Vectors. *Cognitive Computation*, 1(2): 139–159.

Kitware. 2025. Natural Robustness Toolkit (NRTK).

Kricheli, J. S.; Vo, K.; Datta, A.; Ozgur, S.; and Shakarian, P. 2024. Error detection and constraint recovery in hierarchical multi-label classification without prior knowledge. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 3842–3846.

Laird, J.; Lebiere, C.; Rosenbloom, P.; and Stocco, A. 2025. A Proposal to Extend the Common Model of Cognition with Metacognition. arXiv:2506.07807.

Laird, J. E.; Lebiere, C.; and 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.

Lanus, E.; and Freeman, L. J. 2025. Combinatorial Testing Applications for Artificial Intelligence Metacognition. Presented at METCOG-25.

Lebiere, C.; Thomson, R.; Stocco, A.; Orr, M.; and Morrison, D. 2025. An Architectual Appraoch to Metacognition. In *Metacognitive Artificial Intelligence*.

Lee, N.; Ngu, N.; Sahdev, H. S.; Motaganahall, P.; Chowdhury, A. M. S.; Xi, B.; and Shakarian, P. 2024. Metal Price Spike Prediction via a Neurosymbolic Ensemble Approach. In *IEEE International Conference on Data Mining, ICDM 2024 - Workshops, Abu Dhabi, United Arab Emirates, December 9, 2024*, 106–110. IEEE.

Leiva, M.; Ngu, N.; Kricheli, J. S.; Taparia, A.; Senanayake, R.; Shakarian, P.; Bastian, N.; Corcoran, J.; and Simari, G. 2025. Consistency-based Abductive Reasoning over Perceptual Errors of Multiple Pre-trained Models in Novel Environments. arXiv:2505.19361.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Mena, J.; Pujol, O.; and Vitrià, J. 2021. A Survey on Uncertainty Estimation in Deep Learning Classification Systems from a Bayesian Perspective. *ACM Comput. Surv.*, 54(9).

Metcalfe, J.; and Shimamura, A. P., eds. 1994. *Metacognition: Knowing about Knowing*. Bradford Books. Cambridge, MA: MIT Press. ISBN 9780262631693.

Nelson, T. O. 1990. Metamemory: A Theoretical Framework and New Findings. In Bower, G. H., ed., *Psychology of Learning and Motivation*, volume 26 of *Psychology of Learning and Motivation*, 125–173. Academic Press.

Nelson, T. O. 1996. Consciousness and metacognition. *American psychologist*, 51(2): 102.

Nelson, T. O.; and Narens, L. 1994. Why investigate metacognition. *Metacognition: Knowing about knowing*, 13: 1–25.

Ngu, N.; Taparia, A.; Simari, G. I.; Leiva, M.; Corcoran, J.; Senanayake, R.; Shakarian, P.; and Bastian, N. D. 2025. Multiple Distribution Shift – Aerial (MDS-A): A Dataset for Test-Time Error Detection and Model Adaptation. In *AAAI Spring Symposium*.

Reder, L. M.; and Ritter, F. E. 1992. What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(3): 435–451.

Sagar, S.; Taparia, A.; and Senanayake, R. 2024. Failures are fated, but can be faded: characterizing and mitigating unwanted behaviors in large-scale vision and language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Schwind, N.; Inoue, K.; Konieczny, S.; and Marquis, P. 2025. Iterated belief change as learning. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI 2025)*.

Shakarian, P.; Bastian, N. D.; Simari, G. I.; and Leiva, M. 2025. METACOG-25. https://metacognition.syracuse.edu/metacog-25/. Accessed: 6 September 2025.

Shakarian, P.; Simari, G. I.; and Bastian, N. D. 2025. Probabilistic Foundations for Metacognition via Hybrid-AI. In *AAAI Spring Symposium*.

Shakarian, P.; and Wei, H., eds. 2025. *Metacognitive Artificial Intelligence*. Cambridge ; New York, NY: Cambridge University Press. ISBN 978-1-009-52247-2.

Shea, N.; Boldt, A.; Bang, D.; Yeung, N.; Heyes, C.; and Frith, C. D. 2014. Supra-personal cognitive control and metacognition. *Trends in cognitive sciences*, 18(4): 186–193.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 8634–8652.

Taddeo, M.; and Floridi, L. 2007. A praxical solution of the symbol grounding problem. *Minds and Machines*, 17(4): 369–389.

Taparia, A.; Ngu, N.; Leiva, M.; Kricheli, J. S.; Corcoran, J.; Bastian, N. D.; Simari, G.; Shakarian, P.; and Senanayake, R. 2025. VLC Fusion: Vision-Language Conditioned Sensor Fusion for Robust Object Detection. arXiv:2505.12715.

Thompson, V. A. 2009. Dual-process theories: A metacognitive perspective. In Evans, J.; and Frankish, K., eds., *In Two Minds: Dual Processes and Beyond*, 171–195. Oxford University Press.

Thompson, V. A.; Prowse Turner, J. A.; and Pennycook, G. 2011. Intuition, reason, and metacognition. *Cognitive Psychology*, 63(3): 107–140.

Toplak, M. E.; West, R. F.; and and, K. E. S. 2014. Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, 20(2): 147–168.

Undorf, M.; Livneh, I.; and Ackerman, R. 2021. Metacognitive Control Processes in Question Answering: Help Seeking and Withholding Answers. *Metacognition and Learning*, 16.

Wason, P.; and Evans, J. 1974. Dual processes in reasoning? *Cognition*, 3(2): 141–154.

Wei, H.; Shakarian, P.; Lebiere, C.; Draper, B. A.; Krishnaswamy, N.; and Nirenburg, S. 2024. Metacognitive AI: Framework and the Case for a Neurosymbolic Approach. In Besold, T. R.; d'Avila Garcez, A.; Jiménez-Ruiz, E.; Confalonieri, R.; Madhyastha, P.; and Wagner, B., eds., *Neural-Symbolic Learning and Reasoning - 18th International Conference, NeSy 2024, Barcelona, Spain, September 9-12, 2024, Proceedings, Part II*, volume 14980 of *Lecture Notes in Computer Science*, 60–67. Springer.

Xi, B.; Scaria, K.; Bavikadi, D.; and Shakarian, P. 2025. Rule-Based Error Detection and Correction to Operationalize Movement Trajectory Classification. In *IJCAI Workshop on Spatio-Temporal Reasoning and Learning*.

Xiong, T.; Wang, X.; Guo, D.; Ye, Q.; Fan, H.; Gu, Q.; Huang, H.; and Li, C. 2025. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13618–13628.

Yan, Z.; Li, D.; Zhang, Z.; and He, J. 2023. Accuracy–Security Tradeoff With Balanced Aggregation and Artificial Noise for Wireless Federated Learning. *IEEE Internet of Things Journal*, 10(20): 18154–18167.

Yang, Y.; Cornelio, C.; Leiva, M.; and Shakarian, P. 2025. Error Detection and Correction for Interpretable Mathematics in Large Language Models. In *AAAI Fall Symposium*.

Yang, Y.; Neary, C.; and Topcu, U. 2024. Multimodal Pretrained Models for Verifiable Sequential Decision-Making: Planning, Grounding, and Perception. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, 2011–2019. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400704864.

Zhang, D.; Lei, J.; Li, J.; Wang, X.; Liu, Y.; Yang, Z.; Li, J.; Wang, W.; Yang, S.; Wu, J.; et al. 2025. Critic-v: Vlm critics help catch vlm errors in multimodal reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9050–9061.

Zhang, X.; Kang, Y.; Chen, K.; Fan, L.; and Yang, Q. 2023. Trading off privacy, utility, and efficiency in federated learning. *ACM Transactions on Intelligent Systems and Technology*, 14(6): 1–32.