Second Workshop on Metacognitive Prediction of AI Behavior

Perspectives on Hyperdimensional Computing for Metacognitive Artificial Intelligence

Zhenzhu A. Nelson, *United States Military Academy, West Point, NY, 10996, USA*

Mohsen Imani, *University of California - Irvine, Irvine, CA, 92697 USA*

Nathaniel D. Bastian, *United States Military Academy, West Point, NY, 10996, USA*

Abstract—Operating in the modern battlefield requires a level of adaptability that conventional artificial intelligence (AI) currently lacks; it remains a technology that requires large amounts of data, storage, and training while lacking the ability to recognize potential errors in decision-making and adapt accordingly. Metacognition is the human brain's awareness of its own cognitive processes and its framework for analyzing and adapting those processes to meet the cognitive needs required of a task. Metacognitive AI, which is AI technology capable of introspection, self-monitoring and self-adaptation, has the potential to revolutionize the field of AI by reducing the need for continual human oversight and retraining with each modification of its tasks or environments. In this paper, we proffer the use of hyperdimensional computing (HDC), which is a robust method of representing large-scale data under a single class while allowing for efficient comparisons less susceptible to noise, techniques to achieve metacognitive AI by creating robust, energy-efficient systems capable of self-regulation, self-analysis, and self-adaptation. Recent research in HDC and metacognitive AI is discussed, highlighting the potential efficacy of utilizing HDC's capabilities to realize metacognitive AI-enabled systems. The intent of this position paper is to outline the potential of combining these two approaches to increase the transparency and assurance of AI.

Artificial intelligence (AI) has revolutionized the way people perform work, becoming an invaluable tool with great potential in fields ranging from education to the defense industry.[1] However, effective AI usage in safety-critical settings is contingent on increasing transparency and assurance of AI.[2] Deep neural networks are the most commonly utilized deep learning based AI methods, but have key limitations in generalization, transparency, and resource efficiency. A survey of deep learning highlighted how neural networks are data hungry, lack transparency, and have limited capacity for transfer.[3] These weaknesses call for the implementation of new architectures for developing and deploying AI-enabled systems.

For military applications, AI failures have serious consequences; thus, accountability and transparency are paramount. The U.S. military has established responsibility, equity, traceability, reliability, and quality of being governable as the five principles of AI usage.[2] Bridging the ethical gap on the battlefield requires increased transparency. In addition, self-adaptability, self-regulation, and low-resource consumption are required to increase the usability of AI on the battlefield.

Metacognition, defined as the perception and regulation of cognitive processes, is integral in influencing human cognition and behavior.[4] In humans, cognitive maps are the mental representations of structures and relationships used to navigate problem solving in both spatial and nonspatial contexts.[5] In AI, cognitive maps are the representations of task structures and adaptive planning. Research in deep neural network-trained large language models revealed how traditional methods of AI lack the capacity to construct cognitive maps, rendering them unable to generalize and adapt decision making.[6] We propose that metacognitive AI

Digital Object Identifier 10.1109/XXX.0000.0000000

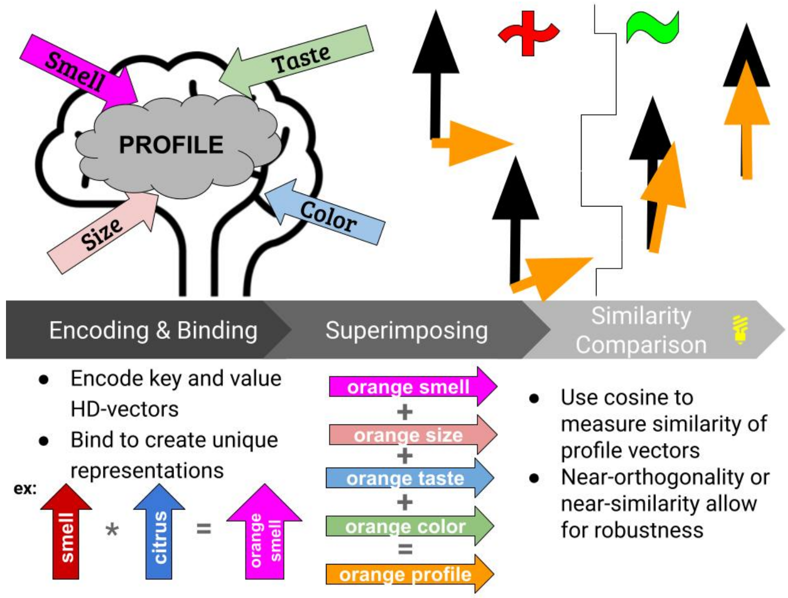

FIGURE 1. Conceptual overview of HDC and its three main operations: encoding, binding, superimposing

capable of representing cognitive maps will permit AI-enabled systems to store and recall previous learning, enabling transfer learning in new contexts and increasing transparency and trust in decision-making.

Key to the utilization of cognitive maps in metacognitive AI is the resource-efficient representation of models capable of training on limited examples. Hyperdimensional computing (HDC), in comparison to traditional architectures like deep neural networks, offers a less resource-intensive method of computing.[7] Furthermore, HDC's ability to outpace deep neural networks when data is limited makes it ideal for computing on the edge with resource and data constraints.[8]

For the remainder of this paper, we identify the unique strengths of HDC and how it can be combined with principles of metacognition to create a unified representation layer capable of storing and comparing cognitive maps to support a larger metacognitive AI architecture that brings self-regulation and self-adaptation to AI-enabled systems.

HYPERDIMENSIONAL COMPUTING

HDC is a method of computing that deviates from traditional von Neumann computing by utilizing vectors existing in hyperdimensional space to perform operations. Inspired by the human brain, HDC is designed to be noise-tolerant and capable of storing, retrieving, and calculating similarity in a less resource-intensive way than traditional computing.[9]

It is possible to achieve such resource-efficient computing because the ability to classify using near similarity removes the requirement for deterministic computing, increasing robustness towards noise. These properties make HDC ideal for situations that require lightweight, transparent and data-efficient computation and multimodal comparison.[9]

Hyperdimensional vectors can encode a multitude of data structures, including sets, binary trees, and graphs, and are capable of storing and computing probabilities without explicit counting, which reduces the memory and computing requirement for HDC.[9] In Figure 1, we use the example of classifying a fruit as an orange to demonstrate HDC's use of multiplication-bound key-value vectors and how profile vectors are formed and compared using superimposing, permutation and cosine similarity.[9]

AI Model Training on the Edge

Recent research has identified potential for using HDC in edge training.[8] To measure HDC performance, researchers compared its performance with traditional convolutional neural networks (CNN). Compared to CNNs, HDC achieved about 20% higher inference accuracy when training with only a single sample per class.[8] Further, HDC only requires that the encoding be performed once (compared to the many epochs required by CNNs). Performing model adjustments after encoding incur lower costs, further supporting the potential of using HDC in training on the edge.[8]

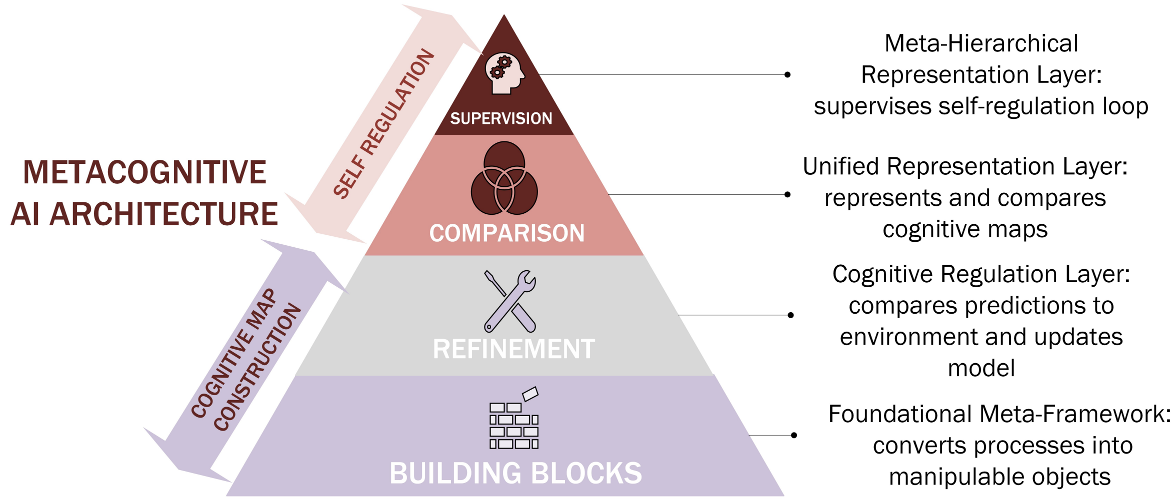

FIGURE 2. Conceptual overview of Metacognitive AI architecture

METACOGNITIVE AI

AI currently lacks metacognition, which is 'cognition about cognition' that provides self-monitoring processes to the brain and that can be used as inspiration for developing metacognitive AI.[10] Tankelevitch et al. suggests that as cognition responsibility is increasingly offloaded to AI, it is increasingly necessary for AI-enabled systems to have the ability to monitor and recognize when its own cognitive processes are inadequate and require resource-efficient adaptation.[11]

Tankelevitch et al. identified four AI failures that can be addressed by metacognitive AI: transparency, reasoning, adaptation, and perception. Metacognitive AI permits AI models to explain outputs and decisions, self-reflect on logic, adapt behavior and strategy, and perceive sensory information in the environment.[11] Thus, use of metacognitive AI-enabled systems increases transparency and enables the operation of fast-paced environments, transferring and adapting knowledge from different contexts in an efficient way.

Metacognitive AI Architecture

We propose a metacognitive AI architecture, depicted in Figure 2, that regulates and adapts cognitive processes, comprising a foundational meta-framework layer, cognitive regulation layer, unified representation layer, and a meta-hierarchical representation layer. We illustrate this architecture using a "hello world" example of a trained AI model for handwritten digit classification adapting to examples outside its training data.

The foundational meta-framework, rooted in category theory, provides a structured approach to cognitive representation by formalizing relational knowledge and hierarchical abstractions.[12] In this layer, the system abstracts processes and concepts as manipulable objects. In our example, the model would identify and represent each potential process. For example, one method of classifying digits from untrained domains might be to create a new, "unknown" category, while another might be to categorize examples of untrained digits into the most similar digit the model knows.

The cognitive regulation layer introduces predictive coding mechanisms, enabling AI to self-monitor decision-making by minimizing prediction error.[14] A key component is the Cognitive Map Learner (CML), which encodes an environment and associated actions into a unified representation space. The CML constructs cognitive maps by learning state-action relationships and refining its internal model over time.

The CML is supported by HDC, which provides a dual representation framework for both superposition-based state encoding and geometric transformations in planning. HDC allows the CML to encode high-dimensional state-action representations using binding and superposition operations, facilitating efficient storage, retrieval, and adaptation. The ability to maintain quasi-orthogonal hypervectors ensures robust pattern recognition and generalization, particularly in environments with partial observability, where one-to-one state mappings are not always available. HDC's associative memory further enables inference of missing state information, reducing dependence on explicit retraining.

In our example of handwritten digit classification, the cognitive regulation layer would identify and adjust the threshold between identifying examples as either belonging to one of the digits within the trained domain or an unknown digit to minimize prediction errors.

The unified representation layer leverages HDC to store and compare cognitive maps as they evolve. Unlike traditional deep learning models that rely on weight updates, HDC dynamically adjusts high-dimensional representations through fast similarity-based operations, allowing real-time adaptation. This self-updating capability ensures that representations remain flexible without requiring extensive retraining. In our example, the unified representation layer would leverage HDC to represent each of the models utilizing the different problem-solving approaches, enabling effective and efficient comparison of their performances.

Finally, the meta-hierarchical representation layer integrates self-regulation mechanisms across all layers, enabling AI to learn from past experiences while adapting to new challenges. In our example, the meta-hierarchical representation layer would identify priorities based on the task at hand and adjust the model accordingly. If the task requires maximizing the accuracy of predictions for each category, it would select models that do not categorize examples identified as being outside the training scope. Yet, if the task requires that all examples be categorized into labels included in the training data, then accuracy may be de-prioritized in favor of labeling each example. This oversight ensures decision-making consistency while maintaining flexibility. By embedding HDC-driven cognitive map learning within this framework, we establish a robust AI-enabled system capable of continuous learning, transparent decision-making, and real-time adaptation.

Metacognitive AI Deployment Considerations

The real-world deployment of metacognitive AI with HDC necessitates further exploration of hardware constraints and system integration. Implementing metacognitive AI at-scale requires advancements in low-power neuromorphic computing, real-time adaptation mechanisms, and more efficient HDC encoding strategies. Future research should investigate how multi-agent metacognitive architectures can be applied to distributed AI decision-making, particularly in environments where multiple autonomous systems must dynamically coordinate under uncertain conditions.

Beyond resource-efficiency, the integration of metacognitive AI with HDC has significant implications for AI explainability and trustworthiness. Traditional deep learning models often operate as black-box systems, limiting their interpretability. However, by structuring AI cognition through HDC-encoded cognitive maps, metacognitive AI can provide more interpretable decision pathways. This structured encoding fosters greater human-AI collaboration, allowing operators to query, audit, and understand AI decisions in real time. Future research should explore how HDC-encoded knowledge representations can be used to generate human-readable AI explanations, further improving trust and transparency in AI-assisted decision-making.

HDC for Self-Regulation and Adaptation

A crucial aspect of implementing metacognitive AI at-scale is ensuring that the underlying architecture remains computationally efficient while dynamically adapting to new scenarios. Traditional deep learning models require extensive retraining when encountering novel conditions, making them impractical for real-time applications in dynamic environments such as cyber defense, aided target recognition, multi-domain operations, and autonomous decision-making. In contrast, HDC enables a lightweight and energy-efficient mechanism for encoding, storing, and comparing cognitive maps, allowing AI-enabled systems to refine their strategies without full-scale retraining. This capability is particularly advantageous for edge computing applications, where real-time self-regulation and low-power adaptability are critical.

Moreover, metacognitive AI can be further enhanced by incorporating predictive coding mechanisms that proactively minimize uncertainty in decision-making processes. Inspired by neuroscience, predictive coding allows AI-enabled systems to continuously compare their expected outcomes against environmental feedback, iteratively reducing errors. When integrated with HDC's ability to efficiently store and manipulate structured cognitive maps, this approach permits AI to conduct real-time self-assessment and autonomous course correction. The result is an AI-enabled system that not only generalizes across diverse tasks but also dynamically fine-tunes its cognitive models, reinforcing robustness, trust, and interpretability in AI-driven decision-making.

By leveraging HDC as the foundation for a meta-hierarchical architecture, we propose a scalable framework that operates efficiently without the computational overhead of traditional neural networks. The ability to encode relationships, actions, and contextual information in a unified high-dimensional space facilitates on-the-fly adaptation, which is essential for military, cybersecurity, and autonomous system applications. Future work should investigate optimizations in hardware accelerators for HDC, ensuring that metacognitive AI can function effectively under real-world constraints while maintaining robustness and explainability.

The Unified Representation Layer

The unified representation layer leverages HDC for efficient cognitive map representation and comparison. Unlike traditional AI models requiring extensive retraining, HDC encodes structured knowledge in high-dimensional spaces, enabling real-time adaptation. Key mathematical frameworks include Fourier Holographic Reduced Representation (FHRR) and Generalized Holographic Reduced Representation (GHRR).

FHRR employs complex-valued hypervectors and Fourier transformations for efficient binding and noise resilience, enhancing dynamic learning,[15] whereas GHRR extends holographic computing with non-binary encoding, improving adaptability and compositional memory.[17] GraphHD further expands HDC's capabilities by encoding graph-based structures, optimizing relational knowledge representation.[19] By integrating these methods, the unified representation layer supports adaptive learning, robust memory encoding, and efficient cognitive map comparisons.[16] These approaches enhance AI self-regulation and edge computing applications, ensuring scalable and data-efficient AI.

CONCLUSION

AI must be trustworthy and assured to be useful on the modern battlefield and in industry at-large. Today, the resource-intensive nature of current deep neural network based AI-enabled systems and their brittleness in the face of ever-changing conditions limits the implementation of AI on the battlefield - particularly in scenarios of multi-domain operations (where navigating changing contexts is key) and autonomy (which must function without constant human supervision).

As such, we proffer the use of HDC to create a unified representation layer to enable self-regulation and self-adaptation necessary for metacognitive AI. HDC provides a resource-efficient method of representation uniquely suited for systems on the edge requiring data-efficient storage and comparison of cognitive maps. Furthermore, it enables the demystification of AI decision-making and increases the operator's trust in the AI-enabled system's ability to self-regulate and self-adaptation in new operating domains.

Future work must further explore frameworks for representing cognitive processes as hyperdimensional vectors and how to increase data-efficiency by adjusting the hardware used in edge devices. Through further investigation of increasing the data-efficiency of HDC and methods of effectively leveraging HDC for metacognitive AI, the development of a meta-architecture capable of self-regulation and self-adaptation will provide increased transparency, trust, and accountability in deployed AI-enabled systems.

DISCLAIMER

The views expressed in this paper are those of the authors and do not reflect the official policy or position of the United States Military Academy, Department of the Army, Department of Defense, or U.S. Government.

References

1. U.S. Department of Defense, "Dod data, analytics, and ai adoption strategy," U.S. Department of Defense, Tech. Rep., 2023, Accessed: Feb. 7, 2025. [Online]. Available: https://media.defense.gov/2023/Nov/02/2003333300/-1/-1/1/DOD_DATA_ANALYTICS_AI_ADOPTION_STRATEGY.PDF.
2. R. Whitty, "2022 international joint conference on neural networks (ijcnn)," *Expeditions with MCUP*, vol. 2022, no. 1, pp. 1–17, 2022, ISSN: 2688-5395. [Online]. Available: https://muse.jhu.edu/pub/419/article/857599.
3. G. Marcus, "Deep learning: A critical appraisal," *CoRR*, vol. abs/1801.00631, 2018. arXiv: 1801.00631. [Online]. Available: http://arxiv.org/abs/1801.00631.
4. J. H. Flavell, "Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry," *American Psychologist*, vol. 34, no. 10, pp. 906–911, 1979. DOI: 10.1037/0003-066X.34.10.906.
5. R. A. Epstein, E. Z. Patai, J. B. Julian, and H. J. Spiers, "The cognitive map in humans: Spatial navigation and beyond," *Nature Neuroscience*, vol. 20, no. 11, pp. 1504–1513, 2017. DOI: 10.1038/nn.4656.
6. I. Momennejad, H. Hasanbeig, F. Vieira, *et al.*, *Evaluating cognitive maps and planning in large language models with cogeval*, 2023. arXiv: 2309.15129 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2309.15129.
7. Z. Yan, S. Wang, K. Tang, and W.-F. Wong, "Efficient hyperdimensional computing," in *Machine Learning and Knowledge Discovery in Databases:*

Research Track. Springer Nature Switzerland, 2023, 141–155, ISBN: 9783031434150. DOI: 10.1007/978-3-031-43415-0_9. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-43415-0_9.
8. P. R. Genssler, S. A. Wasif, M. Wael, R. Novkin, and H. Amrouch, "Frontiers in edge ai with risc-v: Hyperdimensional computing vs. quantized neural networks," in *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2024, pp. 1–6. DOI: 10.23919/DATE58400.2024.10546832.
9. P. Kanerva, "Hyperdimensional computing: An algebra for computing with vectors," in *Advances in Semiconductor Technologies: Selected Topics Beyond Conventional CMOS*. 2023, pp. 25–42. DOI: 10.1002/9781119869610.ch2.
10. H. Wei, P. Shakarian, C. Lebiere, B. Draper, N. Krishnaswamy, and S. Nirenburg, *Metacognitive ai: Framework and the case for a neurosymbolic approach*, 2024. arXiv: 2406.12147 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2406.12147.
11. L. Tankelevitch, V. Kewenig, A. Simkute, *et al.*, "The metacognitive demands and opportunities of generative ai," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ACM, May 2024, 1–24. DOI: 10.1145/3613904.3642902. [Online]. Available: http://dx.doi.org/10.1145/3613904.3642902.
12. T. Leinster, *Basic category theory*, 2016. arXiv: 1612.09375 [math.CT]. [Online]. Available: https://arxiv.org/abs/1612.09375.
13. S. Phillips, "What is category theory to cognitive science? compositional representation and comparison," *Frontiers in Psychology*, vol. 13, p. 1 048 975, 2022. DOI: 10.3389/fpsyg.2022.1048975.
14. C. Stöckl, Y. Yang, and W. Maass, "Local prediction-learning in high-dimensional spaces enables neural networks to plan," *Nature Communications*, vol. 15, p. 2344, 2024. DOI: 10.1038/s41467-024-46586-0.
15. T. A. Plate, "Holographic reduced representations," *IEEE Transactions on Neural Networks*, vol. 6, no. 3, pp. 623–641, 1995. DOI: 10.1109/72.377968.
16. P. S. Frady and F. T. Sommer, "Hyperdimensional computing with holographic and adaptive encoder," *Frontiers in Neuroscience*, vol. 15, 2021. DOI: 10.3389/fnins.2021.683300.
17. M. Imani, I. Tanj, A. Qazi, and T. Rosing, "Generalized holographic reduced representation for efficient learning," *Proceedings of the 57th Design Automation Conference (DAC)*, pp. 1–6, 2020. DOI: 10.1145/3394885.3407831.
18. M. Imani, A. Zakeri, H. Alimohamadi, and T. Rosing, "Learning from hypervectors: A survey on hypervector encoding," *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–38, 2023. DOI: 10.1145/3516825.
19. M. Imani, H. Alimohamadi, A. Zakeri, and R. Cammarota, "Graphd: Graph-based hyperdimensional memorization for brain-like cognitive learning," *Frontiers in Neuroscience*, vol. 16, p. 893, 2022. DOI: 10.3389/fnins.2022.812734.
20. P. Sutor, D. Yuan, D. Summers-Stay, C. Fermuller, and Y. Aloimonos, "Gluing neural networks symbolically through hyperdimensional computing," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–10. DOI: 10.1109/IJCNN55064.2022.9892622.
21. E. Osipov, D. Kleyko, and A. Legalov, "Associative synthesis of finite state automata model of a controlled object with hyperdimensional computing," Nov. 2017. DOI: 10.1109/IECON.2017.8216554.

Zhenzhu A. Nelson is currently an undergraduate student in the Department of Mathematical Sciences at the United States Military Academy at West Point, where she is pursuing a B.S. degree in Applied Statistics and Data Science. Her primary research interest is artificial intelligence for defense applications. Contact her at zhenzhu.nelson@westpoint.edu.

Mohsen Imani is currently an Assistant Professor in the Department of Computer Science at the University of California, Irvine and Director of the Bio-Inspired Architecture and Systems Lab. He received his Ph.D. from the University of California, San Diego. His primary research interests are hyperdimensional computing and cognitive learning for artificial intelligence with defense applications. Contact him at mohseni@uci.edu.

Nathaniel D. Bastian is currently an Assistant Professor in the Department of Mathematical Sciences at the United States Military Academy at West Point, as well as Chief Scientist and Director, Office of Science & Engineering at the Army Cyber Institute. He received his Ph.D. from the Pennsylvania State University. His primary research interests are artificial intelligence assurance, neuro-symbolic artificial intelligence, and uncertainty quantification with defense applications. Contact him at nathaniel.bastian@westpoint.edu.