# Metacognitive AI: Framework and the Case for a Neurosymbolic Approach

Hua Wei[1], Paulo Shakarian[1(✉)], Christian Lebiere[2], Bruce Draper[3], Nikhil Krishnaswamy[3], and Sergei Nirenburg[4]

[1] Arizona State University, Tempe, USA
{hua.wei,pshak02}@asu.edu
[2] Carnegie Mellon University, Pittsburgh, USA
cl@cmu.edu
[3] Colorado State University, Fort Collins, USA
{bruce.draper,nkrishna}@colostate.edu
[4] Rensselaer Polytechnic Institute, Troy, USA

**Abstract.** Metacognition is the concept of reasoning about an agent's own internal processes and was originally introduced in the field of developmental psychology. In this position paper, we examine the concept of applying metacognition to artificial intelligence. We introduce a framework for understanding metacognitive artificial intelligence (AI) that we call TRAP: transparency, reasoning, adaptation, and perception. We discuss each of these aspects in-turn and explore how neurosymbolic AI (NSAI) can be leveraged to address challenges of metacognition.

**Keywords:** Metacognition · Neurosymbolic AI

## 1 Introduction

Metacognition is the concept of reasoning about an agent's own internal processes and was originally introduced in the field of developmental psychology [17] as a description of higher-order cognition. This "cognition about cognition" is regarded by some as a self-monitoring process that is integral to the functioning of the human mind [13]. It has been studied extensively in the fields of manufacturing [28], aerospace [23], transportation [2,6,20], and military applications [32]. We argue for the study of metacognitive artificial intelligence which deals with the reasoning about an artificial agent's own processes. This idea actually has been studied on and off in the history of AI [8,9], but recent developments indicate that this area deserves a renewed focus. Specifically, despite large scale industry investments in AI, major failures still occur - which indicates that pure engineering solutions are unlikely to solve these fundamental failures. Consider the following examples:

– Large language model falsely accuses a professor of sexual harassment [30].


T. R. Besold et al. (Eds.): NeSy 2024, LNAI 14980, pp. 60–67, 2024.
https://doi.org/10.1007/978-3-031-71170-1_7

- Autonomous robot taxi in San Francisco accidentally drags a woman for 20 ft causing major injury [16].
- Reinforcement learning model has to be retrained to play with slight changes in the environment [24,35].
- Robot mistakes a man for a box in South Korea and crushes him to death [4].

Each of these case studies exhibits a different modality of AI failure. The first item illustrates a failure of **Transparency** - the system generated information that was false and could not provide a way to check itself on the facts. The second illustrates a failure in **Reasoning** - how the system synthesizes information and ultimately produces a decision. The third illustrates a failure of **Adaptation** - the system could not accommodate itself in a new environment. The fourth illustrates a failure in **Perception** - how the system recognizes entities in its environment. In this introduction, which stemmed from the *2023 ARO-sponsored Workshop on Metacognitive Prediction of AI Behavior*, we argue that the study of metacognitive AI should encompass these four areas (**TRAP**), as is shown in Fig. 1. In this paper, we build on our recent ARO-sponsored workshop event [36] and examine each of these aspects and then discuss how neurosymbolic AI can be used as an approach to address challenges in metacognition.

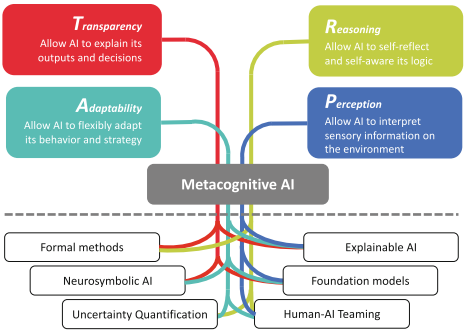


**Fig. 1.** Four aspects of metacognitive AI (TRAP) and approaches to achieve metacognition

## 2 The TRAP Framework for Metacognition

A traditional artificial intelligence (AI) system could be simplified as $y = f_\theta(x)$, where $x$ is the input for the AI system; $y$, depending on applications, could be a description, prediction, or actions to take; and $f_\theta$ is the operational function in most AI systems with parameters $\theta$. A metacognitive AI system could be an additional function $g$. With different metacognitive areas, $g$ is in different locations concerning $f$. With this framing in mind, we examine each aspect of the TRAP framework below.

*Transparency.* While traditional AI can sometimes be perceived as a 'black box', metacognitive AI enhances trust and transparency by making the decision-making process in black-box AI more understandable to users. This is achieved through the function $g(f(x), \theta)$ or the function $g|f$. The function $g(f(x), \theta)$ represents the process of generating explanations based on both the input $x$ and the parameters $\theta$ of the model $f$, while the function $g|f$ represents the function $f$ with a series of $g$. This function allows metacognitive AI to explain its decisions in terms of both the input data and its internal parameters, catering to different user expectations and motivations for seeking explanations.

On one hand, the nature of the explanation can vary significantly depending on whether it's intended for an expert with technical knowledge or a layperson. If they are looking to understand why certain outputs come from a global perspective, then the focus is to have $g(\theta)$ to make the $\theta$ transparent; if users are looking to understand certain cases, then the focus is to have $g(f(x))$ to understand a certain prediction $f(x)$ on $x$. On the other hand, the purpose behind an explanation necessitates a different approach to how explanations are formulated and presented: is the explanation for enhancing the performance of the system, reducing bias, increasing fairness, or simply deriving a clearer understanding of the AI's decision-making process? Enhancing the performance of the system through transparency could involve using the explanations to correct predictions or induce actions $g|f$ where the understanding outputted from $g$ is then processed by the AI system to better learn $f$. [29] argued building cognitive models of both the AI and the human user that could be introspected upon to adapt explanations according to the discrepancy between the two models, e.g., when the AI decision does not conform to the human model's expectations. [34] applied explainable AI framework to a real-world clinical machine learning (ML) use case, i.e., an explainable diagnostic tool for intensive care phenotyping. Co-designing with 14 clinicians, they provided five explanation strategies to mitigate decision biases and moderate trust. They implemented an early decision aid system to diagnose patients in an Intensive Care Unit (ICU) and found that users employed a diverse range of explainable AI facilities to reason.

*Reasoning.* Traditional AI systems $f$ often rely on predefined algorithms and data sets for reasoning or decision-making, which can limit their effectiveness in dynamic or unfamiliar scenarios. In contrast, metacognitive AI incorporates self-reflection and self-awareness into its logic, represented by $f(x; g(\theta))$. This indicates how the AI's self-reflection (through $g$) informs its decision-making process (through $f$), enhancing its reasoning capabilities. For instance, a metacognitive AI in healthcare could use this approach to evaluate and refine its diagnostic criteria over time, learning to differentiate between complex cases and refining its diagnostic criteria based on outcomes. This leads to decisions that are not only based on data but also enriched by the AI's growing experiential knowledge, resulting in more accurate and reliable outcomes.

In [33], the authors showed that model-based reflection may guide reinforcement learning with two benefits: The first is a reduction in learning time as compared to an agent that learns the task via pure reinforcement learning. Sec-

ondly, the reflection-guided RL agent shows benefits over the pure model-based reflection agent, matching the performance of that agent in the metrics measured in addition to converging to a solution in fewer trials. In addition, the augmented agent eliminates the need for an explicit adaptation library such as is used in the pure-model-based agent and thus reduces the knowledge engineering burden on the designer significantly. In [3], a novel technique called Hindsight Experience Replay was introduced, whose intuition is to re-examine the trajectories with a different goal - while a trajectory may not help learn how to achieve the desired goal, it tells us something about how to achieve the state in the actual trajectory. They demonstrated this approach on the task of manipulating objects with a robotic arm on three different tasks: pushing, sliding, and pick-and-place, while the vanilla RL algorithm fails to solve these tasks.

*Adaptability.* Adaptation in metacognitive AI encapsulates the system's ability to detect and correct errors of internal conditions and to flexibly adapt its behavior and strategies. This is represented by $f'(x; g(f(x), \theta))$, where $f'$ is the adapted model based on the metacognitive assessment $g$ of the original model's output $f(x)$ and parameters $\theta$. This notation reflects how metacognitive AI adapts by reassessing its outputs and parameters, allowing for more effective decision-making in the face of uncertainty and changing environments [10,39]. Additional adaptations could also be implemented with $g(x)?f(x) : h(x)$, where the metacognitive process $g$ decides whether to use the main function $f$ or an alternative function $h$ based on its analysis of the input $x$. This could model AI systems that choose different processing paths based on metacognitive assessment without modifying $f$. Such systems can adapt to new environments and tasks by understanding their learning process and limitations. [27] showed that uncertainty-informed decision referral can improve diagnostic performance. More recently, another metacognitive approach allowing for adaptability known as *error detection and correction rules* (EDCR) was introduced [37]. In this framework, function $g$ results in a set of learned rules that characterize the failure modes of $f$ and how to correct on those failure modes while $f'$ is an inference process conducted using these rules to erase or change the results of the underlying model $f$. In [37] the authors applied this technique to the classification of geospatial movement trajectories and examined performance improvement on $f$, where the current state-of-the-art is neural architecture known as LRCN [26]. EDCR was able to both improve over the state-of-the-art as well as exhibit the ability to improve performance when exposed to out-of-distribution data.

*Perception.* Perception refers to the ability to interpret sensory information to understand the environment. Perception in metacognitive AI involves the system's ability to interpret and understand sensory information, such as visual and auditory data, in a context-aware manner, represented by $f(g(x), x)$, where context is a metacognitive assessment of the AI's own capacity rather than an external context. Here, $f$ represents the primary perceptual processing function, interpreting sensory data like visuals or audio, while the metacognitive function $g$ specifically evaluates the accuracy and limitations of the AI's own sensory processing. The primary perceptual function $f$ then uses both the original sensory

input $x$ and the metacognitive assessment $g(x)$ to refine its interpretation. This dual-input model allows the AI to recognize and compensate for any inherent biases or weaknesses in its perception. This could include AI in autonomous vehicles that must interpret complex visual environments, in medical imaging distinguishing subtle diagnostic details, or in environmental monitoring systems that detect and analyze changes through sensory data.

## 3 Neurosymbolic AI for Metacognition

In the previous section, we introduced the TRAP framework of metacognitive AI , which provides a structured approach to understanding how metacognitive elements augment traditional AI systems. Building upon this foundation, in this section, we explore the emerging field of neurosymbolic AI (NSAI) and its profound implications for metacognition. NSAI refers to the integration of connections (e.g., neural) with symbolic (e.g., logical) systems. This term was coined in the early 2000s and has gained wider prominence in recent years [18, 21,25,31]. The key relationships between NSAI and metacognition relate to the ability to use symbolic knowledge and perceptual models to detect and correct errors in each other (adaptability) and the use of symbolic languages to express information about error modes of a perceptual model (transparency).

With the introduction of Logic Tensor Networks [5] the canonical paradigm for NSAI has consisted of guiding gradient descent with the addition of soft logical constraints in the loss function - and this was followed by related work [19,38]. In general, these loss-based approaches would not fit the metacognitive paradigm, as in these incarnations of NSAI, the symbolic logic is used as an additional optimization criteria - in much the same way as one would a regularization term. However, more recent views on NSAI do lend themselves to metacognition - in particular with respect to adaptability and transparency.

The key intuition in the use of NSAI for metacognitive adaptability is to leverage symbolic domain knowledge to explicitly identify errors in a neural model, allowing for some corrective action to be performed. One well-known approach for NSAI metacognitive adaptability is abductive learning (ABL) [11]. Using the paradigm of adaptability introduced in this paper, function $f'$ returns a result based on the combination of a perceptual model, a-priori domain knowledge (i.e., a logic program), and abduced error information (function $g$ in our framework). Here, function $g$ is abduced based on inconsistencies between the perceptual model and domain knowledge and can take the form of additional symbolic structures added to the logic program and/or updates to the perceptual model ($f$ in our framework). ABL has been shown to provide SOTA performance on combined perception-reasoning tasks as well as application to the identification of new concepts as shown in [22]. More recent applications of NSAI to metacognitive adaptability have sought to disentangle perceptual updates from the base perceptual model. Specifically, [7] introduces a framework where an additional transformer model ($g$ in our framework) is used to predict errors in the underlying neural ($f$) using symbolic knowledge and reinforcement learning to detect

and correct perceptual errors. The work of [37] also addresses perceptual errors but using a rule-learning approach - here rules are learned about the results of the neural model that allow for error detection and correction while providing the byproduct of an explanation of the errors.

Complementary to the NSAI work relevant to metacognitive adaptability is NSAI work relating to metacognitive transparency. Here, NSAI is used to reason directly about the inner workings of a perceptual model, often for a downstream task involving an explanation of the perceptual results. One example of such work is [1] where a binarized neural network is used to produce a symbolic theory of perception used in a downstream task of appreciation, providing an explanation of the perceptual results. Another application of NSAI to transparency deals with the use of concept induction [12] to map activations in a neural network to an explanation using description logic - thereby providing transparency.

## 4 Challenges

The idea of metacognitive AI leads to many open questions. In the application of NSAI to metacognition, there are several such as the challenge of creating symbolic structures to reason about (e.g., using inductive logic programming to obtain a knowledgebase [15], leveraging common sense knowledge [14] - both still have major challenges). More broadly, there are challenges that apply to metacognitive AI in general, which include the following:

- *Generalization to Diverse Dynamic Environments.* Metacognitive AI must be capable of adapting to rapidly changing and unpredictable environments, or at least know when it is incapable.
- *Designing for Continuous Self-Improvement.* Enabling AI systems to not only identify their weaknesses or errors but also to autonomously modify their behavior and learning strategies for continuous improvement.
- *Ensuring Ethical and Responsible Metacognition.* As metacognitive AI systems will have a higher level of autonomy in decision-making, ensuring that these decisions are ethically and morally responsible.
- *Interpreting Metacognitive Processes.* While explainability in AI is already a challenge, making the metacognitive processes of AI interpretable and understandable to humans adds an extra layer of complexity.
- *Benchmarking Datasets and Baselines.* Such benchmarks would afford validation of an AI's self-assessment and adaptive learning capabilities are functioning as intended, which requires human assessment.

While the ideas of metacognitive AI are still very new, the ideas of error correction have previously been used to establish the foundations for technologies such as computer networking and digital signal processing. We believe a formal study of the topic with respect to AI may yield similar advances in the future.

**Acknowledgement.** This work was funded by the Army Research Office (ARO) under grants W911NF2310345 and W911NF2410007.

**Disclosure of Interests.** The authors have no competing interests.

## References


1. Making sense of raw input: Artif. Intell. **299**, 103521 (2021)
2. Abduljabbar, R., Dia, H., Liyanage, S., Bagloee, S.A.: Applications of artificial intelligence in transport: an overview. Sustainability **11**(1), 189 (2019)
3. Andrychowicz, M., et al.: Hindsight experience replay. Adv. Neural Inform. Process. Syst. **30** (2017)
4. Atkinson, E.: Man crushed to death by robot in south korea (Nov 2023). https://www.bbc.com/news/world-asia-67354709
5. Badreddine, S., d'Avila Garcez, A., Serafini, L., Spranger, M.: Logic tensor networks. Artif. Intell. **303**, 103649 (2022)
6. Caesar, H., et al.: nuscenes: a multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11621–11631 (2020)
7. Cornelio, C., Stuehmer, J., Hu, S.X., Hospedales, T.: Learning where and when to reason in neuro-symbolic inference. In: The Eleventh International Conference on Learning Representations (2022)
8. Cox, M.T.: Metacognition in computation: a selected history. In: AAAI Spring Symposium: Metacognition in Computation, pp. 1–17 (2005)
9. Cox, M.T., Raja, A.: Metareasoning: Thinking about thinking. MIT Press (2011)
10. Da, L., Mei, H., Sharma, R., Wei, H.: Uncertainty-aware grounded action transformation towards sim-to-real transfer for traffic signal control. In: 2023 62nd IEEE Conference on Decision and Control (CDC), pp. 1124–1129. IEEE (2023)
11. Dai, W.Z., Xu, Q., Yu, Y., Zhou, Z.H.: Bridging machine learning and logical reasoning by abductive learning. Adv. Neural Inform. Process. Syst. **32** (2019)
12. Dalal, A., Sarker, M.K., Barua, A., Hitzler, P.: Explaining deep learning hidden neuron activations using concept induction (2023)
13. Demetriou, A., Efklides, A., Platsidou, M., Campbell, R.L.: 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, pp. i–202 (1993)
14. Du, L., Ding, X., Liu, T., Qin, B.: Learning event graph knowledge for abductive reasoning. In: Proc. of the 59th ACK. ACL (Aug 2021)
15. Evans, R., Grefenstette, E.: Learning explanatory rules from noisy data. J. Artifi. Intell. Res. **61**, 1–64 (2018)
16. Farivar, C.: Cruise robotaxi dragged woman 20 feet in recent accident, local politician says (Oct 2023). https://www.forbes.com/sites/cyrusfarivar/2023/10/06/cruise-robotaxi-dragged-woman-20-feet-in-recent-accident-local-politician-says/?sh=2d68e761466b
17. Flavell, J.H.: Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry. Am. Psychol. **34**(10), 906 (1979)
18. d'Avila Garcez, A., Lamb, L.C.: Neurosymbolic AI: the 3rd wave. CoRR (2020)
19. Giunchiglia, E., Lukasiewicz, T.: Coherent hierarchical multi-label classification networks. In: Proc. of the 34th International Conference of NIUPS, NIPS 2020, Red Hook, NY, USA (2020)
20. Grigorescu, S., Trasnea, B., Cocias, T., Macesanu, G.: A survey of deep learning techniques for autonomous driving. J. Field Robot. **37**(3), 362–386 (2020)
21. Hitzler, P., Sarker, M.K., Eberhart, A. (eds.): Compendium of Neurosymbolic Artificial Intelligence, Frontiers in Artificial Intelligence and Applications, vol. 369. IOS Press (2023)

22. Huang, Y.X., Dai, W.Z., Jiang, Y., Zhou, Z.H.: Enabling knowledge refinement upon new concepts in abductive learning (2023)
23. Izzo, D., Märtens, M., Pan, B.: A survey on artificial intelligence trends in spacecraft guidance dynamics and control. Astrodynamics **3**, 287–299 (2019)
24. Jayawardana, V., Tang, C., Li, S., Suo, D., Wu, C.: The impact of task underspecification in evaluating deep reinforcement learning. Adv. Neural. Inf. Process. Syst. **35**, 23881–23893 (2022)
25. Kautz, H.A.: The third ai summer: Aaai robert s. engelmore memorial lecture. AI Mag. **43**(1), 105–125 (2022)
26. Kim, J., Kim, J.H., Lee, G.: Gps data-based mobility mode inference model using long-term recurrent convolutional networks. Trans. Res. Part C: Emerging Technol. **135**, 103523 (2022)
27. Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S.: Leveraging uncertainty information from deep neural networks for disease detection. Sci. Rep. **7**(1), 17816 (2017)
28. Li, B.H., Hou, B.C., Yu, W.T., Lu, X.B., Yang, C.W.: Applications of artificial intelligence in intelligent manufacturing: a review. Front. Inform. Technol. Electronic Eng. **18**, 86–96 (2017)
29. Mitsopoulos, K., Somers, S., Schooler, J., Lebiere, C., Pirolli, P., Thomson, R.: Toward a psychology of deep reinforcement learning agents using a cognitive architecture. Top. Cogn. Sci. **14**(4), 756–779 (2022)
30. Mok, A.: Chatgpt reportedly made up sexual harassment allegations against a prominent lawyer (Apil 2023). https://www.businessinsider.com/chatgpt-ai-made-up-sexual-harassment-allegations-jonathen-turley-report-2023-4#:~:text=OpenAI's%20buzzy%20ChatGPT%20falsely%20accused,source%2C%20The%20Washington%20Post%20reported
31. Shakarian, P., Baral, C., Simari, G.I., Xi, B., Pokala, L.: Neuro Symbolic Reasoning and Learning. Springer Briefs in Computer Science, Springer (2023)
32. Svenmarck, P., Luotsinen, L., Nilsson, M., Schubert, J.: Possibilities and challenges for artificial intelligence in military applications. In: Proceedings of the NATO Big Data and Artificial Intelligence for Military Decision Making Specialists'Meeting, pp. 1–16 (2018)
33. Ulam, P., Goel, A., Jones, J., Murdock, W.: Using model-based reflection to guide reinforcement learning. Reasoning, Represent. Learn. Comput. Games **107** (2005)
34. Wang, D., Yang, Q., Abdul, A., Lim, B.Y.: Designing theory-driven user-centric explainable ai. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–15 (2019)
35. Wei, H., et al.: Honor of kings arena: an environment for generalization in competitive reinforcement learning. Adv. Neural. Inf. Process. Syst. **35**, 11881–11892 (2022)
36. Wei, H., Shakarian, P.: Metacognitive Artificial Intelligence. Cambridge University Press (2024)
37. Xi, B., Scaria, K., Shakarian, P.: Rule-based error detection and correction to operationalize movement trajectory classification (2023)
38. Xu, J., Zhang, Z., Friedman, T., Liang, Y., den Broeck, G.V.: A semantic loss function for deep learning with symbolic knowledge (2018)
39. Ye, K., Chen, T., Wei, H., Zhan, L.: Uncertainty regularized evidential regression. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38(15), pp. 16460–16468 (2024). https://doi.org/10.1609/aaai.v38i15.29583