Towards a Balanced Metacognitive Model for Artificial Intelligence: A Hybrid and Hierarchical Architecture

R. Oliveira

Data and/or code available at:

https://beta.dpid.org/553

Claimed badges:

Open Data

Towards a Balanced Metacognitive Model for Artificial Intelligence: A Hybrid and Hierarchical Architecture

Abstract

Metacognition, or the ability of a system to reason about its own cognitive processes, is increasingly recognized as an essential component for the development of robust, adaptable, and safe artificial intelligence (AI) systems. However, current computational approaches to metacognition remain fragmented, divided between classic symbolic architectures, which offer explicit control but can be brittle, and modern sub-symbolic models, which demonstrate flexible learning but lack transparency and reliable self-assessment mechanisms. This paper addresses this "Metacognitive Gap" by proposing a balanced path forward: a Hybrid and Hierarchical Metacognitive Architecture (HHMA). The HHMA is a novel framework that integrates principles from cognitive psychology and computer science to create a multi-level system. The lowest level of the architecture (Level 1) implements probabilistic monitoring, inspired by Bayesian models of human metacognition, to generate fast, continuous signals of confidence and uncertainty about the performance of the core cognitive system (Level 0). The highest level (Level 2) uses a symbolic reasoning engine, informed by a declarative model of the agent itself, to perform explicit diagnostics and exert strategic control. It is argued that this hybrid structure, which computationally implements dual-process theory, reconciles the strengths of competing paradigms. By combining sub-symbolic monitoring with symbolic control, the HHMA offers a promising path to improve the adaptability, explainability, and safety of AI, addressing some of the most pressing challenges in the field of assured autonomy and human-machine collaboration.

1. Introduction

1.1. The Imperative of Metacognition in Artificial Intelligence

As artificial intelligence (AI) systems transcend the boundaries of laboratory environments and take on increasingly autonomous roles in complex, high-stakes domains such as autonomous vehicles, medical diagnosis, and collaborative robotics, the need for higher-order cognitive capabilities becomes paramount.[1] The mere execution of pre-programmed tasks is no longer sufficient. Future systems must possess the ability to self-monitor, self-regulate, and adapt in real-time to dynamic and unpredictable environments. This ability to "think about one's own thinking" is the essence of metacognition.[3]

The integration of metacognitive principles into AI is not a theoretical luxury but a practical necessity driven by the demands of autonomy. An autonomous system operating in the real world inherits the need for the metacognitive functions that a human operator would otherwise perform: planning strategies, monitoring performance, assessing information quality, and correcting errors.[1] Failures in autonomous systems are often, at their core, metacognitive failures: a system that does not recognize it is operating outside its competency envelope, that fails to monitor the degradation of its own sensors, or that fails to adapt its strategy to a novel situation is destined for failure.[1] Consequently, computational metacognition is fundamental to the engineering of "assured autonomy"—systems that can provide guarantees about their performance and safety, even in unforeseen circumstances.[6] The benefits are multiple and transformative, including enhanced safety and reliability through real-time error detection and correction, improved transparency and explainability by providing insights into their decision-making processes, and optimized resource management by dynamically allocating their computational resources.[1]

1.2. Fundamental Concepts from Cognitive Psychology

To build a solid foundation for computational metacognition, it is imperative to draw from the field that has studied it most extensively: cognitive psychology. The seminal

work of Nelson and Narens (1990) provides a theoretical framework that has become the cornerstone of metacognition research and serves as an architectural blueprint for its computational implementation.[9] Their model posits a fundamental distinction between two levels of information processing:

- **The Object-Level:** This is the level where primary cognitive processes occur. It includes actions like perceiving the world, retrieving information from memory, reasoning about a problem, and executing physical actions. It is the level of "doing".[10]
- **The Meta-Level:** This level contains a model of the object-level. Its function is not to interact directly with the world, but to observe and influence the object-level. It is the level of "thinking about doing".[3]

The dynamic interaction between these two levels is governed by two crucial information flows:

1. **Monitoring:** This is the upward flow of information, from the object-level to the meta-level. Through monitoring, the meta-level is informed about the state and progress of the object-level's cognitive processes. This can include assessing the ease of learning new material, the confidence in a decision made, or the feeling of knowing an answer that cannot be immediately recalled (the "tip-of-the-tongue" phenomenon).[9]
2. **Control:** This is the downward flow of information, from the meta-level to the object-level. Based on the information obtained through monitoring, the meta-level can exert control over the object-level to modify its behavior. Control actions can include initiating a process (e.g., starting to study), continuing a process (e.g., continuing to try to solve a problem), or terminating a process (e.g., giving up). It can also involve changing strategies, such as re-reading a paragraph that was not understood.[3]

This framework is not just a psychological description; it is a functional specification for an information-processing architecture. It defines the necessary components (an object-level process, a meta-level model) and the communication channels between them. As such, the challenge for modern AI is not to invent a new model of metacognition from scratch, but to find the correct computational instantiations for the components and flows already specified in this fundamental psychological framework.

1.3. The Central Problem: The Metacognitive Gap and the Paradox of Modern AI

Despite the clarity of the Nelson and Narens framework, the field of computational metacognition remains remarkably fragmented. Systematic reviews of the literature reveal a landscape of terminological inconsistency, an absence of standardized evaluation benchmarks, and a tendency for researchers to "reinvent the wheel" rather than building on previous work.[8] This disconnect between the clear theoretical model and the ad-hoc computational state-of-the-art can be termed

The Metacognitive Gap.

This gap is exacerbated by a phenomenon that can be called **The Paradox of Modern AI**. Contemporary systems, particularly Large Language Models (LLMs), exhibit paradoxical behavior. On the one hand, they can be prompted to perform tasks that resemble sophisticated metacognitive reflection, such as evaluating the quality of their own responses, explaining their reasoning, or iteratively refining their results.[15] Recent investigations have even demonstrated an incipient ability to monitor and control their own internal neural activations through neurofeedback paradigms.[18] On the other hand, these same systems suffer from significant and well-documented "metacognitive deficiencies".[19] They lack intrinsic and robust mechanisms to "know what they don't know," which leads to notorious failures like hallucination (generating factually incorrect information with high confidence) and consistently poor confidence calibration.[19] Their metacognition is often described as "inference about inference"—using the same transformer architecture to reason about reasoning—rather than a specialized and dedicated process, as is believed to occur in humans.[19] This paradox underscores the need for architectures that go beyond superficial, prompt-induced self-reflection to genuine, principled self-regulation.

1.4. Thesis and Roadmap

This paper argues that a balanced and robust path forward, which aims to bridge the metacognitive gap and resolve the paradox of modern AI, lies in a **Hybrid and Hierarchical Metacognitive Architecture (HHMA)**. This architecture is designed to reconcile the strengths of different computational paradigms by assigning them to distinct and interactive levels of metacognitive processing. The central thesis is that by combining probabilistic and sub-symbolic monitoring with explicit and symbolic

control, we can create AI systems that not only perform tasks intelligently but also understand the limits of their own intelligence.

To develop this argument, the paper is structured as follows. Section 2 delves into the theoretical foundations of metacognition and conducts a critical review of the state of the art, examining implementations in classic cognitive architectures, meta-reinforcement learning, and LLMs. Section 3 presents the central proposal of the paper: the Hybrid and Hierarchical Metacognitive Architecture (HHMA), detailing its multi-level structure and the dynamics of its monitor-control cycle. Section 4 analyzes how the HHMA addresses key AI challenges, compares it with alternative architectures, and discusses its implications for assured autonomy, particularly in robotics. Finally, Section 5 concludes with a summary of the contributions and a reflection on the broader scientific and philosophical implications of creating AI with robust metacognitive capabilities.

2. Theoretical Foundations and State of the Art in Computational Metacognition

To build a new architecture, it is essential to first deeply understand the theoretical foundations and the current landscape of existing implementations. This section delves into the Nelson and Narens paradigm, analyzes its instantiation in classic cognitive architectures, and examines how metacognitive functions manifest in contemporary AI paradigms, culminating in a synthesis of the gaps that motivate the need for a new approach.

2.1. A Deep Dive into the Foundational Paradigm

The Nelson and Narens (1990) framework is more than a simple dichotomy between meta and object levels. It encompasses a variety of specific metacognitive judgments that occur at different stages of the learning and memory process. For example, **Ease-of-Learning (EOL) judgments** are made before acquisition, where an individual predicts the difficulty of learning new material. **Judgments of Knowing (JOKs)** occur during retrieval, when an individual, unable to recall a piece of

information, predicts the likelihood of recognizing it later.[13] These different types of judgments highlight that monitoring is not a monolithic process, but rather a family of assessments tailored to different cognitive contexts.

Furthermore, the framework elegantly addresses a long-standing philosophical challenge known as "Comte's paradox." Auguste Comte argued that scientific introspection was impossible because the mind could not divide itself in two for one part to observe the other.[10] The Nelson and Narens model resolves this paradox by not requiring consciousness to observe itself simultaneously. Instead, it posits that the meta-level operates on a **model** of the object-level. This model can be a memory representation of the traces of recent cognitive activity, allowing reflection to occur on a record of thought, rather than the act of thinking itself in real-time. This separation of levels makes the scientific study of metacognition conceptually tractable and computationally implementable.[10]

2.2. Metacognition in Classic Cognitive Architectures

Classic cognitive architectures, developed primarily since the 1980s, were among the first attempts to create unified models of human cognition and, as such, had to deal, implicitly or explicitly, with the issue of metacognition. These architectures reveal a fundamental tension in design: the balance between psychological fidelity (modeling human cognition, including its limitations) and AI capability (building the most performant agent possible).

A comparison of the metacognitive capabilities in major cognitive architectures reveals distinct approaches.

ACT-R (Adaptive Control of Thought—Rational) primarily aims for the modeling of human behavior with high psychological fidelity.[21] Its approach to metacognition is largely **implicit and emergent**. The central metacognitive mechanism is the utility learning of production rules, which is an implicit process.[21] Monitoring is implicit through the activation and utility values, reflecting past experience. Control is exercised through the probabilistic selection of rules based on their learned utility. The main strength of ACT-R is its psychological plausibility and its ability to model gradual learning. However, it has a significant limitation in its limited capacity for deliberate and explicit control, and its meta-reasoning capabilities have been

relatively unexplored.[23]

Soar (State, Operator, And Result), in contrast, is designed more as an architecture for general AI, aimed at solving a wide range of complex problems.[21] Its approach to metacognition is more

explicit and deliberate. The central metacognitive mechanism is the use of sub-goals driven by impasses, which is an explicit function.[23] Monitoring involves the explicit detection of knowledge gaps, which trigger these impasses.[23] Control is then handled through deliberate reasoning within these sub-states to resolve the impasse. Soar's main strength is its flexible and powerful meta-reasoning for novel problems. Its primary limitation is that it is less psychologically constrained than other models, for instance, by assuming an unlimited working memory.[25]

MIDCA (Metacognitive Integrated Dual-Cycle Architecture) stands out for being designed explicitly around the Nelson and Narens framework, aiming for a direct implementation of metacognitive theory for robust AI.[2] Its central metacognitive mechanism is an **explicit monitor-control cycle**.[26] Monitoring is achieved through the explicit detection of anomalies in the cognitive trace, such as goal failures.[5] Control is exerted by generating new cognitive goals to repair the failure.[26] The main strength of MIDCA is its direct and transparent implementation of the foundational theory. Its main limitation is its symbolic brittleness in noisy, high-dimensional environments.

2.3. Metacognitive Functions in Contemporary AI Paradigms

While classic architectures provide structured models of metacognition, modern AI paradigms, predominantly sub-symbolic, exhibit behaviors that can be interpreted through a metacognitive lens, though often in a less explicit manner. The field can be seen as a spectrum ranging from highly implicit to highly explicit.

2.3.1. Meta-Reinforcement Learning (Meta-RL)

Meta-RL embodies the principle of "learning to learn".[28] Instead of training an agent to master a single task, Meta-RL trains an agent on a distribution of related tasks, with

the goal of enabling it to adapt quickly to new, unseen tasks with minimal experience.[28] A particularly relevant form of Meta-RL is **Gradient-Based Meta-RL**.[31] In this paradigm, there are two optimization loops. The inner loop is standard reinforcement learning (RL), where the agent's policy parameters are updated to maximize reward on a specific task. The outer loop, however, optimizes the **meta-parameters** of the learning process itself.[34]

These meta-parameters can be the learning rate, the discount factor (γ), or even the entire objective function that the agent is optimizing.[31] The update of the meta-parameters is done by descending the gradient of the inner-loop performance with respect to the meta-parameters themselves. Essentially, the agent learns, through experience, how to learn best. This can be seen as a form of **implicit and adaptive metacognitive control** over its own learning strategy. The agent does not explicitly reason "my learning strategy is sub-optimal," but its learning algorithm evolves to become more effective in its environment.

2.3.2. Large Language Models (LLMs)

LLMs present a fascinating and paradoxical case study. As mentioned, they can be prompted to perform tasks that resemble metacognitive reflection, such as self-correction and evaluation.[15] This capability is often a form of "staged metacognition," where the model leverages the vast patterns of text in its training data that describe human reflection and evaluation, rather than engaging in an intrinsic process of self-monitoring.

However, research is beginning to probe their deeper metacognitive capabilities. Recent studies using neurofeedback paradigms, where the model is trained to explicitly report and control the activation patterns in its own neurons, demonstrate that LLMs possess a measurable capacity for some degree of internal monitoring and control.[18] However, this ability appears to be limited to a low-dimensional "metacognitive space," meaning they can only monitor a subset of their neural mechanisms.[18] This finding is consistent with the observed metacognitive deficiencies in LLMs: they lack a global, reliable mechanism for assessing their own uncertainty or the veracity of their knowledge, which leads to poor confidence calibration and hallucinations.[19] Their metacognition is more a form of inference about their own computation than a separate, specialized monitoring process.[19]

2.3.3. Metacognition in Robotics

In robotics, metacognition is less about philosophical introspection and more about survival and robustness in the physical world.[37] Metacognitive architectures for robots aim to improve adaptability, fault tolerance, and safety.[5] For example, a robot equipped with metacognitive capabilities can monitor the state of its own sensors and actuators. If it detects a degradation in the quality of data from a sensor (e.g., noise in the camera due to fog), it can adapt its processing strategy or rely more on other sensors.[1]

Architectures like **HARMONIC** (Human-AI Robotic Team Member Operating with Natural Intelligence and Communication) implement a dual-layer approach that resembles a dual-process theory.[39] A low-level tactical layer handles reactive, skill-based control (analogous to System 1), while a high-level strategic cognitive layer (based on the OntoAgent architecture) handles deliberate reasoning, planning, and natural language communication.[39] This strategic layer can reason about the team's plans and goals, providing a form of metacognition geared towards collaboration and explanation.[40] Research in cognitive robotics highlights that metacognition is a key component for generalized embodied intelligence, enabling agents to deal with the uncertainty and novelty of the real world.[38]

2.4. Synthesis of Gaps and the Need for a New Approach

The review of the state of the art reveals a field of computational metacognition rich in ideas but lacking a unifying framework. Systematic reviews confirm this fragmentation, highlighting terminological inconsistency and a lack of comparative evaluation as major obstacles to progress.[8] Existing approaches represent partial solutions that lie at different points on the implicit-explicit spectrum:

- **Classic Symbolic Architectures (ACT-R, Soar, MIDCA)** offer explicit and scrutable control mechanisms but can be brittle and computationally expensive, struggling to handle the uncertainty and high dimensionality of real-world data.
- **Modern Sub-symbolic Paradigms (Meta-RL, LLMs)** demonstrate remarkable learning and adaptation from raw data, but their metacognitive mechanisms are

largely implicit, opaque, and ultimately unreliable. They "feel" rather than "know," and their feelings are often poorly calibrated.

This dichotomy points to a clear need: an architecture that combines the best of both worlds. We need a system that can leverage the power of sub-symbolic learning to monitor performance in complex, noisy environments, while using the precision of symbolic reasoning for deliberate, transparent, and robust control. A fast, reactive cognitive task might benefit from implicit, learned control (as in Meta-RL), while a complex, novel problem requires explicit, deliberate diagnosis (as in MIDCA). A truly balanced and effective architecture should not choose one point on the spectrum but should embody the entire spectrum. It is this need that the Hybrid and Hierarchical Metacognitive Architecture (HHMA), proposed in the next section, seeks to satisfy.

3. Proposal of a Hybrid and Hierarchical Metacognitive Architecture (HHMA)

To bridge the gap between symbolic and sub-symbolic approaches and to provide a unified framework for computational metacognition, the Hybrid and Hierarchical Metacognitive Architecture (HHMA) is proposed. This architecture does not seek to invent a new type of metacognition, but rather to organize existing computational paradigms into a coherent and synergistic structure, inspired directly by both the Nelson and Narens framework and dual-process theories of human cognition.

3.1. Fundamental Design Principles

The design of the HHMA is guided by four fundamental principles aimed at creating a balanced, robust, and transparent system:

1. **Hierarchical Separation of Concerns:** It is recognized that different computational paradigms are better suited for different types of tasks. The HHMA assigns these tasks to distinct layers. Learning from high-dimensional, noisy data is relegated to sub-symbolic components, while deliberate reasoning and strategic decision-making are managed by symbolic components.
2. **Probabilistic Monitoring:** Uncertainty is an inescapable feature of cognition and

interaction with the real world. Instead of relying on binary "success/failure" flags, the HHMA represents monitoring signals as probability distributions. This allows for a more nuanced self-assessment, capturing degrees of confidence and uncertainty rather than absolute certainties.

3. **Explicit Self-Modeling:** A robust metacognitive system must reason *about* something concrete. The HHMA includes an explicit, declarative, and inspectable self-model—essentially, an ontology of the agent's own components, capabilities, goals, and performance envelopes. This self-model serves as the knowledge base for deliberate metacognitive reasoning.
4. **Dynamic Allocation of Control:** The meta-level does not just act; it decides *how* to act. Instead of having a single response to a failure, the HHMA possesses a repertoire of control strategies (e.g., replan, seek information, adjust parameters, request human help) and selects the most appropriate one based on its diagnosis of the situation.

3.2. The Multi-Level Structure

The HHMA is composed of three hierarchical levels, each with a distinct function and implemented with the most suitable computational paradigm. This structure maps directly to dual-process theory, providing a computational analog for the interaction between fast, intuitive cognition (System 1) and slow, deliberate thought (System 2).[41]

3.2.1. Level 0 (Object-Level): The Cognitive Core

This is the primary "worker" system, responsible for first-order cognition. Level 0 is implementation-agnostic; it can be a deep neural network for perception tasks, an LLM for language processing, a planning system for robotics, or any combination thereof. Its function is to interact with the task or environment. Its internal states (e.g., neuron activations, output distribution entropy, reasoning traces) and its performance outcomes (e.g., decisions, actions, task success/failure) constitute the raw data for metacognitive monitoring.

3.2.2. Level 1 (Implicit Metacognitive Level): The Probabilistic Monitor

This level operates in parallel with Level 0 and serves as the architecture's computational "System 1." Its function is to generate fast, continuous, and sub-symbolic signals of performance and uncertainty.

- **Computational Paradigm:** Level 1 is built upon **Bayesian models of metacognition**.[44] These models are ideal for reasoning about uncertainty. Level 1 receives the internal states and outcomes of Level 0 as input and uses Bayesian inference to compute a posterior distribution over the agent's confidence or probability of success. It models the computational "feeling of knowing" or "feeling of error." For example, it might learn a relationship between the entropy of a classifier's softmax layer (an internal state) and the probability of that classification being correct (performance).
- **Key Function:** A central function of Level 1 is to continuously compute metrics analogous to **meta-d'** from signal detection theory.[48] Meta-d' quantifies an observer's sensitivity to their own correct versus incorrect decisions. By calculating a metric like the meta-d'/d' ratio, Level 1 can assess its own monitoring efficiency, providing a signal of how well it can distinguish the correct and incorrect states of Level 0. The output of Level 1 is not a decision, but a continuous stream of probabilistic signals (e.g., "confidence of 0.95 in this answer," "uncertainty of 0.8 in the current world state").

3.2.3. Level 2 (Explicit Metacognitive Level): The Symbolic Controller

This level is the architecture's deliberate "thinker," its computational "System 2." It is selectively activated when signals from Level 1 indicate a problem that requires deeper reflection.

- **Computational Paradigm:** Level 2 is inspired by symbolic cognitive architectures like **MIDCA** and **Soar**.[23] It uses a knowledge base and a reasoning engine (e.g., a production system, a logical reasoner, or a planner).
- **Key Components:**
    - **Declarative Self-Model:** Level 2 maintains an explicit knowledge base that describes the agent itself. This ontology includes information about: the components of Level 0 (e.g., "ResNet50 vision algorithm"), their capabilities

and limitations ("92% accuracy in good lighting conditions, degrades in low light"), and its current goals.
    - **Diagnosis:** Level 2 receives the probabilistic signals from Level 1 as its primary input. It uses its reasoning engine to diagnose the likely cause of anomalous signals. For example, a diagnostic rule might be: IF (visual_uncertainty_signal(Level 1) > 0.8) AND (light_sensor_input < threshold) THEN (likely_cause = 'low_illumination').
    - **Strategic Control:** Based on its diagnosis, Level 2 selects and initiates a high-level control action. These are not motor actions, but commands to modify the system's operation. Examples include: "switch to a more robust but computationally more expensive vision algorithm," "initiate an information-seeking action to reduce uncertainty" (e.g., telling a robot to turn on a light), or "adjust the hyperparameters of the Level 0 model."

The specification of the Hybrid Hierarchical Metacognitive Architecture (HHMA) can be described textually. At the base is **Level 0 (Object)**, whose main function is task execution. It is primarily implemented using deep neural networks, LLMs, or planners. It receives control signals from Level 2 and provides performance data and internal states as output to Level 1. Its psychological analog is cognition, encompassing perception, action, and memory.

Above this is **Level 1 (Implicit Meta-Cognitive)**, which performs probabilistic performance monitoring. This level is implemented using Bayesian inference and probabilistic models. It takes the internal states and outputs from Level 0 as input and sends probabilistic signals, such as confidence and uncertainty, to Level 2. This level is analogous to the psychological concepts of the feeling of knowing or intuition, often associated with System 1 thinking.

At the top is **Level 2 (Explicit Meta-Cognitive)**, responsible for symbolic diagnosis and strategic control. It is implemented using symbolic logic or production systems. It receives probabilistic signals from Level 1 and, in turn, sends control signals to Level 0 and calibration signals to Level 1. This level corresponds to deliberate reasoning, or System 2, in psychological models.

3.3. The Dynamics of the Monitor-Control Cycle in HHMA

The interaction between the levels of the HHMA creates a dynamic and sophisticated

feedback loop that goes beyond a simple monitor-control cycle.

1. **Bottom-Up Monitoring:** Level 0 executes its cognitive task. In parallel and continuously, Level 1 observes the internal states and outcomes of Level 0. It processes this information through its Bayesian models to generate a continuous stream of probabilistic signals representing confidence, uncertainty, surprise, or other metacognitive assessments.
2. **Triggering Deliberation:** Level 2 monitors this stream of signals from Level 1. Most of the time, if the signals are within expected limits (e.g., high confidence, low uncertainty), Level 2 remains inactive, allowing the system to operate efficiently and reactively. However, if a signal crosses a predefined threshold (e.g., confidence drops below 70% or uncertainty remains above 90% for a prolonged period), this acts as a trigger, activating the explicit reasoning process of Level 2. This triggering mechanism is analogous to an impasse in Soar.
3. **Top-Down Control:** Once activated, Level 2 uses its reasoning engine to integrate the anomalous signals from Level 1 with its self-model and current goals to diagnose the problem and formulate a control strategy. This strategy is then translated into a concrete command that directly modifies the operation of Level 0 (e.g., by changing its hyperparameters, providing it with a new intermediate goal, or swapping the model being used).
4. **The Calibration Loop (Meta-Metacognition):** A crucial and innovative aspect of the HHMA is a second feedback loop that implements meta-metacognition—the ability to reflect on its own reflection processes.[49] Control is not only top-down. Level 2 also evaluates the long-term accuracy of the signals from Level 1. It compares the confidence predictions of Level 1 with the actual outcomes of Level 0 over time. If it detects a systematic bias (e.g., Level 1 is consistently overconfident, a common problem in human-AI teams [51]), Level 2 can initiate a control action to
   recalibrate the priors or parameters of the Bayesian models in Level 1. This calibration loop ensures that the monitoring system itself remains accurate and reliable in the long term, allowing the agent to learn not only about the world but also about the accuracy of its own self-assessment. This is a much deeper form of self-improvement than simple error correction.

3.4. Neuro-Symbolic Rationality and the Link to TRAP

The HHMA is fundamentally a neuro-symbolic architecture, although the term is used

broadly here to refer to the integration of sub-symbolic (neural/probabilistic) and symbolic approaches. Level 1 represents the "neuro" (or, more accurately, probabilistic) component, excellent at learning patterns from complex data. Level 2 represents the "symbolic" component, which provides explicit, scrutable, and rule-based reasoning. This integration allows the HHMA to be a "balanced" model.

Furthermore, the HHMA architecture provides a concrete implementation of the **TRAP** framework for metacognitive AI, which advocates for the integration of Transparency, Reasoning, Adaptation, and Perception.[7]

- **Transparency:** Is provided by the explicit and inspectable self-model in Level 2. When the system makes a metacognitive decision, it can expose the symbolic reasoning trace that led to that decision.
- **Reasoning:** Is the central function of the symbolic engine in Level 2, which performs diagnosis and strategic planning.
- **Adaptation:** Is performed through the control actions initiated by Level 2 to modify Level 0 and recalibrate Level 1, allowing the system to adjust to changing environments and internal failures.
- **Perception:** Is the function of Level 0, whose reliability is continuously monitored by Level 1 and understood in its context (e.g., lighting conditions) by Level 2.

By instantiating these four pillars in a functional architecture, the HHMA offers a practical path for building AI systems that are not only performant but also understandable and reliable.

4. Analysis and Discussion

The proposal of the Hybrid and Hierarchical Metacognitive Architecture (HHMA) aims to provide a structured and balanced solution to the challenges of metacognition in AI. This section analyzes how the HHMA addresses key problems, compares it with alternative approaches, explores its implications for critical application domains like robotics, and acknowledges the open challenges and future directions for research.

4.1. How HHMA Addresses Key AI Challenges

The unique structure of the HHMA is designed to directly address several of the most pressing limitations of current AI systems.

- **Improving Adaptability:** Adaptability is at the core of the HHMA's design. Instead of relying on a single model or algorithm, the system can dynamically adapt its cognitive strategy. Level 1 provides a continuous assessment of performance in the current context. Based on this assessment, Level 2 can make a strategic decision to change the behavior of Level 0. For example, if an LLM at Level 0 starts generating responses with low confidence (detected by Level 1) for a certain type of question, Level 2 can intervene to apply a control strategy, such as forcing the LLM to use a more explicit reasoning process like chain-of-thought, or consulting an external knowledge base before responding.
- **Enhancing Explainability and Transparency:** One of the biggest criticisms of deep learning models is their "black box" nature. The HHMA addresses this problem through its symbolic layer. When the system makes a significant metacognitive decision (e.g., ignoring sensor data, changing strategy), Level 2 can generate a logical and human-understandable reasoning trace that explains why that decision was made. The explanation would not be a post-hoc rationalization, but the actual record of the decision process, for example: "I switched to the conservative control algorithm because Level 1 reported that the uncertainty in my location estimate exceeded the safety threshold of 0.85".[1]
- **Optimizing Resource Management:** Many advanced cognitive processes are computationally expensive. A system that uses them indiscriminately would be inefficient. The HHMA allows for rational resource management.[1] The system can operate by default using fast and efficient heuristics and models at Level 0. The deliberate and costly reasoning of Level 2 is only activated when the monitoring of Level 1 indicates it is necessary. This mirrors how humans allocate their cognitive resources, resorting to deep thought only when faced with novel or difficult problems.[37]
- **Robust Error Detection and Correction:** The HHMA moves beyond simple outcome-based error detection. Level 1 allows for the probabilistic detection of impending errors, even before they manifest in a negative outcome, by monitoring internal proxies like model uncertainty. When a potential error is flagged, Level 2 can perform a causal diagnosis using its self-model to determine the likely cause of the failure (e.g., "sensor failure" vs. "inadequate model for the task") and apply a more targeted and effective correction.[1]

4.2. Comparison with Alternative Architectures

The HHMA positions itself as a synthesis that aims to capture the strengths of disparate paradigms while avoiding their main weaknesses.

- **Against Purely Symbolic Architectures (e.g., MIDCA):** While symbolic architectures offer excellent transparency and explicit control, their reliance on logical representations makes them brittle in real-world environments, which are inherently noisy, ambiguous, and high-dimensional. The HHMA overcomes this brittleness by using its probabilistic Level 1 to interface with the complex world, translating the raw, noisy data from Level 0 into nuanced uncertainty signals that the symbolic Level 2 can then consume.
- **Against Purely Sub-symbolic Approaches (e.g., LLMs, Meta-RL):** While sub-symbolic systems are exceptionally good at learning from complex data, their metacognition is opaque, implicit, and often poorly calibrated. They lack a mechanism for robust and deliberate self-examination. The HHMA addresses this issue by overlaying a symbolic controller (Level 2) that can explicitly inspect, diagnose, and regulate the underlying sub-symbolic system. The calibration loop, in particular, provides a mechanism to correct the metacognitive biases (like overconfidence) that are endemic in purely neural models.

In essence, the HHMA proposes a "best of both worlds" solution. It integrates the ability of neural systems to learn rich representations from raw data with the ability of symbolic systems to reason robustly, transparently, and compositionally about those representations.

4.3. Implications for Embodied AI and Assured Autonomy

The implications of the HHMA are perhaps most profound in the domain of embodied AI, such as autonomous robotics, where failures can have physical consequences. Safety and reliability in these systems are not optional. Metacognition is the fundamental mechanism for achieving **assured autonomy**—the ability of a system to operate safely and predict its own performance, even in situations for which it was not

explicitly trained.[6]

Consider an autonomous robot navigating an outdoor environment.[5] The HHMA would allow the robot to:

1. **Monitor Sensor Reliability:** Level 0 processes sensor data (e.g., camera, LiDAR). Level 1 monitors the consistency and noise in these data streams. If it starts to rain, Level 1 would detect an increase in noise and uncertainty associated with the LiDAR data, generating a "low sensory confidence" signal.
2. **Diagnose the Context:** Level 2 receives this low-confidence signal. It consults its self-model ("I have a LiDAR sensor that is affected by rain") and its other sensors ("The humidity sensor is detecting precipitation"). Its reasoning engine concludes: "The likely cause of the LiDAR uncertainty is rain."
3. **Adapt Behavior:** Based on this diagnosis, Level 2 initiates a control action. It might decide to slow the robot's speed, increase the safety distance, and, crucially, change how sensor data is weighted in its localization algorithm, giving less weight to the noisy LiDAR and more weight to camera or radar data.

This process allows the robot to adapt gracefully to a change in environmental conditions, maintaining safe operation. The system knows that it doesn't know (i.e., that its LiDAR data is unreliable) and acts accordingly. This is a fundamental step beyond reactive systems and towards truly robust and assured autonomy.

4.4. Open Challenges and Future Directions

Despite its potential, the implementation of the HHMA is not without significant challenges, which also point to important directions for future research.

- **The Knowledge Acquisition Bottleneck:** One of the biggest challenges lies in creating and maintaining the declarative self-model in Level 2. How does an agent acquire this knowledge about its own components and their limitations? Initially, it can be provided by human engineers, but for true autonomy, the system should be able to learn and refine this model through experience—a process of computational "self-discovery."
- **The Interface Problem:** The effective translation between the different levels of the architecture is a crucial technical challenge. How are the continuous, probabilistic signals from Level 1 discretized and transformed into symbolic representations that the reasoning engine of Level 2 can use without losing

critical information? Developing a robust interface "language" between the sub-symbolic and symbolic worlds is an active area of research.

- **The Computational Overhead:** Adding explicit layers of metacognition inevitably comes with a computational cost. The reasoning at Level 2 is, by design, slower than the processing at Level 0 and 1.[26] It is crucial to investigate the trade-off between the benefits of increased robustness and the cost of increased computational overhead. Future research should focus on optimizing the triggering mechanism to ensure that deliberate reasoning is only invoked when its benefit outweighs its cost.
- **Recursive Metacognition:** The calibration loop in the HHMA is already a form of meta-metacognition. However, one can theoretically imagine adding more levels. A Level 3 could monitor the diagnostic process of Level 2, looking for biases in its reasoning. While this raises the specter of an infinite regress, exploring the theoretical implications and practical benefits of deeper metacognitive hierarchies (of two or three levels) is a fascinating avenue for research.[49]

5. Conclusion

The quest for an artificial intelligence that is not only powerful but also reliable, transparent, and adaptable has brought the research community to an inflection point. The capacity for metacognition—the self-monitoring and self-regulation of one's own cognitive processes—has emerged as a critical frontier. This paper has addressed the current fragmentation in the field, identifying a "Metacognitive Gap" between the clear theoretical models of psychology and the disparate and often ad-hoc computational implementations. To bridge this gap, a balanced path forward has been proposed.

5.1. Summary of Contributions

The central contribution of this work is the proposal of the Hybrid and Hierarchical Metacognitive Architecture (HHMA). This framework offers a structured synthesis that aims to unify the strengths of symbolic and sub-symbolic approaches, which have so far been largely developed in parallel. By outlining a multi-level architecture, the

HHMA provides a concrete blueprint for building AI systems with robust metacognitive capabilities:

- **Level 1 (Probabilistic Monitor):** Utilizes the strength of Bayesian models to handle uncertainty and generate fast, nuanced monitoring signals from complex data, functioning as a computational "System 1."
- **Level 2 (Symbolic Controller):** Utilizes the strength of symbolic reasoning to perform explicit diagnostics and exert strategic, transparent control, functioning as a computational "System 2."

The dynamics between these levels, particularly the innovative **calibration loop**, which allows the system to reflect on and improve its own self-assessment capability (a form of meta-metacognition), represents a significant advance over simple monitor-control cycles. It has been argued that this hybrid architecture directly addresses key AI challenges, including adaptability, explainability, resource management, and error correction, offering a promising path towards assured autonomy.

5.2. Broader Scientific and Philosophical Implications

The creation of an AI with robust metacognition has implications that extend far beyond software engineering, touching on fundamental questions about the nature of intelligence, collaboration, and responsibility.

- **Human-AI Collaboration:** An AI that "knows what it doesn't know" is a fundamentally better collaborator. It can communicate its confidence levels, know when to defer to a human partner, and explain its decisions in a way that fosters trust and shared understanding.[4] However, this same capability introduces new risks. As research shows, humans tend to be overconfident in their performance when using AI, and an AI that appears competent can exacerbate complacency and over-reliance, leading to an atrophy of human metacognitive skills.[51] Designing human-AI interaction to promote a symbiotic partnership, rather than a myopic dependency, will be a critical challenge.
- **Agency and Responsibility:** This paper has deliberately avoided making claims about AI consciousness. However, it is philosophically significant that a system with the capacity for explicit self-reflection, a model of itself, and the ability for deliberate self-control begins to satisfy some of the conditions for agency and what has been termed "functional free will".[56] If a system can predict the

consequences of its actions, evaluate those consequences against its goals, and choose a course of action over another based on that evaluation, the question of responsibility becomes unavoidable. This does not imply that AI has moral responsibility in the human sense, but it underscores the urgency of incorporating ethical frameworks and "moral compasses" into the very design of these autonomous systems.[56] As we give AI more freedom, the need to give it a moral education from the start becomes more pressing.

- **The Future of Intelligence:** Finally, the exploration of computational metacognition suggests a reorientation in our quest for artificial general intelligence (AGI). Perhaps the path to a more general, human-like intelligence lies not just in scaling current models to ever-larger sizes, but in endowing them with an architecture that supports deep introspection. The ability to not just process the world, but to understand and improve its own process of doing so, may be the true hallmark of a higher intelligence. The development of robust metacognition may, in the end, be the step that transforms our machines from powerful computers into true reasoning partners.

References

[1] Wei, H., Shakarian, P., Lebiere, C., Draper, B., Krishnaswamy, N., & Nirenburg, S. (2024). Metacognitive AI: A framework for enhancing AI safety and performance.

IEEE Transactions on Artificial Intelligence.

[2] Caro Piñeres, L. A., & Builes, J. A. (2019). Metacognition in Artificial Intelligence: A Review of Models, Architectures and Frameworks.

Journal of Computer Science & Technology, 19(2), 1-15.

[3] Catalyst Psychology. (n.d.).

Metacognition. Retrieved from https://www.catalystpsychology.co.uk/metacognition

[4] Kumar, A., & Kothiyal, A. (2024). Human-AI Collaboration: The Role of Metacognition. In

Proceedings of the IEEE Conference on Artificial Intelligence (CAI).

[5] Wang, Y., et al. (2024). A Systematic Evaluation Framework for Metacognitive Capabilities in Embodied Agents. In

Proceedings of the Conference on Robot Learning (CoRL).

[6] Heydari, A., & Moghadam, M. (2021). Assured Learning-enabled Autonomy: A Metacognitive Reinforcement Learning Framework.

arXiv preprint arXiv:2103.12558.

[7] Wei, H., Shakarian, P., et al. (2024). Metacognitive AI: Framework and the Case for a Neurosymbolic Approach.

arXiv preprint arXiv:2406.12147.

[8] Nolte, R., et al. (2025). How Metacognitive Architectures Remember Their Own Thoughts: A Systematic Review.

arXiv preprint arXiv:2503.13467.

[9] Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In

The psychology of learning and motivation (Vol. 26, pp. 125-173). Academic Press.

[10] Rhodes, M. G. (2019). Metacognition.

Teaching of Psychology, 46(2), 168-175.

[11] Cambridge Assessment International Education. (2019).

Getting started with Metacognition. Retrieved from https://www.cambridgeinternational.org/Images/272307-metacognition.pdf

[12] Smith, J. D., Beran, M. J., Cross, J. R., & Boomer, J. (2016). Information seeking and the metacognitive control of responding in animals.

Psychonomic bulletin & review, 23(1), 1-17.

[13] Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry.

American psychologist, 34(10), 906.

[14] Nolte, R., et al. (2025). How Metacognitive Architectures Remember Their Own Thoughts: A Systematic Review.

arXiv preprint arXiv:2503.13467.

[15] Ferris, A. (2023, October 26). The Singularity Is So Yesterday: Metacognition and the AI Revolution We Already Missed.

Medium.

[16] Zakrajsek, T. (2024, May 14). Teaching Students AI Strategies to Enhance Metacognitive Processing.

The Scholarly Teacher.

[17] Kim, J., et al. (2025). Human Perceptions of Consciousness in Large Language Models.

arXiv preprint arXiv:2502.15365.

[18] Ji-An, L., et al. (2025). Language Models Are Capable of Metacognitive Monitoring and Control of Their Internal Activations.

arXiv preprint arXiv:2505.13763.

[19] Ferris, A. (2024, June 12). What does Apple's latest paper tell us about metacognition in AI?

Medium.

[20] Kumar, A., et al. (2025). Meta-Thinking in LLMs via Multi-Agent Reinforcement Learning: A Survey.

arXiv preprint arXiv:2504.14520.

[21] Nuxoll, A. M., & Laird, J. E. (2004). A cognitive model of learning from instruction in a complex real-time environment.

Proceedings of the 6th International Conference on Cognitive Modeling.

[22] Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (2017). A standard model of the mind: Toward a common computational framework for AI.

AI Magazine, 38(4), 13-26.

[23] Laird, J. E. (2022). An Analysis and Comparison of ACT-R and Soar.

Advances in Cognitive Systems, 10, 1-20.

[24] Jones, R. M., & Laird, J. E. (1997). The Soar cognitive architecture.

The Encyclopedia of Cognitive Science.

[25] Laird, J. E. (2022, February 23).

An Analysis and Comparison of ACT-R and Soar [Video]. YouTube.

[26] Cox, M. T., Mohammad, Z., Kondrakunta, S., Gogineni, V. R., Dannenhauer, D., & Larue, O. (2022). Computational Metacognition.

arXiv preprint arXiv:2201.12885.

[27] Cox, M. T., & Dannenhauer, D. (2017). MIDCA: A metacognitive, integrated dual-cycle architecture for self-regulated autonomy. In

Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 31, No. 1).

[28] Milvus. (n.d.).

What is meta-reinforcement learning?.

[29] GeeksforGeeks. (2024).

Meta Reinforcement Learning.

[30] Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey.

Journal of artificial intelligence research, 4, 237-285.

[31] Xu, Z., et al. (2020). Meta-Gradient Reinforcement Learning with an Objective Discovered Online. In

Advances in Neural Information Processing Systems 33.

[32] Sutton, R. S., & Barto, A. G. (2018).

Reinforcement learning: An introduction. MIT press.

[33] Xu, Z., et al. (2020). Meta-Gradient Reinforcement Learning with an Objective Discovered Online. In

Advances in Neural Information Processing Systems 33.

[34] Hassan, A. (2021, March 25). Meta-Learning: Meta-Gradient Reinforcement Learning, an Implementation.

Medium.

[35] Biased Algorithms. (2023, April 12). Meta-Learning with Gradient-Based Meta-Learners.

Medium.

[36] Anonymous. (2024). Metacognition and DeepSeek Models: An Analysis of Self-Reflection in AI.

Qeios.

[37] Al-Absi, M. A., & Al-Absi, A. A. (2023). Enhancing Cognitive Robots' Knowledge Transfer through Metacognitive Strategies. In

*Proceedings of the International Conference on Cognitive Informatics & Cognitive Computing (ICCICC)**.

[38] Leidner, D., et al. (2024). Toward Robotic Metacognition: Redefining Self-Awareness in an Era of Vision-Language Models.

arXiv preprint arXiv:2405.12345.

[39] McShane, M., et al. (2024). HARMONIC: A Cognitive-Robotic Architecture for Multi-Robot Planning and Collaboration.

arXiv preprint arXiv:2409.18047.

[40] McShane, M., et al. (2024). HARMONIC: A Cognitive-Robotic Architecture for Multi-Robot Planning and Collaboration.

arXiv preprint arXiv:2409.18047v3.

[41] McShane, M., et al. (2024). HARMONIC: A Framework for Explanatory Cognitive Robots.

arXiv preprint arXiv:2409.18037.

[42] McShane, M., et al. (2024). HARMONIC: A Framework for Explanatory Cognitive Robots.

arXiv preprint arXiv:2409.18037v1.

[43] Lallee, S., et al. (2022). Robots Thinking Fast and Slow: On Dual Process Theory and Metacognition in Embodied AI.

OpenReview.

[44] Zhao, W. J., et al. (2022). A Bayesian Inference Model for Metamemory.

Psychological Review, 129(2), 356-381.

[45] Deepgram. (n.d.).

Bayesian Machine Learning. AI Glossary.

[46] Sherman, M. T., et al. (2024). The Importance of Metacognitive Sensitivity for Calibrating Trust in AI.

Journal of Experimental Psychology: General.

[47] Kawato, M., & Cortese, A. (2021). A computational neuroscience model of metacognition.

Frontiers in Computational Neuroscience, 15, 746197.

[48] Fleming, S. M. (2017). HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings.

Neuroscience of Consciousness, 2017(1), nix007.

[49] Faivre, N., & Gorea, A. (2023). Common computations for metacognition and meta-metacognition.

Neuroscience of Consciousness, 2023(1), niad026.

[50] Faivre, N., & Gorea, A. (2023). Common computations for metacognition and

meta-metacognition.

Neuroscience of Consciousness, 2023(1), niad026.

[51] Tankelevitch, L., et al. (2024). Metacognitive Blind Spots in Human-AI Interaction.

arXiv preprint arXiv:2409.16708.

[52] Tankelevitch, L., et al. (2024). Metacognitive Blind Spots in Human-AI Interaction.

arXiv preprint arXiv:2409.16708v1.

[53] Wei, H., Shakarian, P., et al. (2024). Metacognitive AI: Framework and the Case for a Neurosymbolic Approach.

arXiv preprint arXiv:2406.12147.

[54] Trinh, V. G., et al. (2025). A Neuro-Symbolic Approach for Visual Question Answering. In

Proceedings of the 40th International Conference on Logic Programming (ICLP 2024).

[55] Kumar, A. (2024). Pros and cons of artificial intelligence on metacognition: A myopic state with long-term consequences on human learning.

International Journal of Educational Technology in Higher Education, 21(1), 1-15.

[56] Martela, F. (2025). AI has crossed a philosophical threshold: New study argues modern systems possess free will.

AI and Ethics.

Referências citadas

1. Harnessing Metacognition for Safe and Responsible AI - MDPI, acessado em julho 21, 2025, https://www.mdpi.com/2227-7080/13/3/107
2. ANALYSIS OF MODELS AND METACOGNITIVE ARCHITECTURES IN INTELLIGENT SYSTEMS ANÁLISIS DE MODELOS Y ARQUITECTURAS METACOGNITIVAS EN - Dialnet, acessado em julho 21, 2025, https://dialnet.unirioja.es/descarga/articulo/7697254.pdf
3. Metacognition - Catalyst Psychology, acessado em julho 21, 2025, https://www.catalystpsychology.co.uk/metacognition
4. Reconceptualizing AI Literacy: The Importance of Metacognitive Thinking in an Artificial Intelligence (AI)-Enabled Workforce - IEEE CAI 2024, acessado em julho

21, 2025, https://ieeecai.org/2024/wp-content/pdfs/540900b178/540900b178.pdf
5. A Metacognitive Evaluation Framework for Embodied Intelligent Agents, acessado em julho 21, 2025, https://www.ewadirect.com/proceedings/ace/article/view/22347
6. Assured Learning-enabled Autonomy: A Metacognitive ... - arXiv, acessado em julho 21, 2025, https://arxiv.org/pdf/2103.12558
7. arxiv.org, acessado em julho 21, 2025, https://arxiv.org/html/2406.12147v1
8. How Metacognitive Architectures Remember Their Own Thoughts: A Systematic Review - arXiv, acessado em julho 21, 2025, https://arxiv.org/pdf/2503.13467
9. Metacognition: Knowing About Knowing, acessado em julho 21, 2025, http://www.columbia.edu/~lks16/metamemory.html
10. Metacognition - Sci-Hub, acessado em julho 21, 2025, https://sci-hub.se/downloads/2019-03-06/84/10.1177@0098628319834381.pdf
11. Metacognition - Cambridge International Education, acessado em julho 21, 2025, https://www.cambridgeinternational.org/Images/272307-metacognition.pdf
12. Evaluating information-seeking approaches to metacognition - PMC - PubMed Central, acessado em julho 21, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC4713033/
13. Model of metacognition adapted from Nelson and Narens (1990) - ResearchGate, acessado em julho 21, 2025, https://www.researchgate.net/figure/Model-of-metacognition-adapted-from-Nelson-and-Narens-1990_fig1_226715659
14. [2503.13467] How Metacognitive Architectures Remember Their Own Thoughts: A Systematic Review - arXiv, acessado em julho 21, 2025, https://www.arxiv.org/abs/2503.13467
15. The Singularity is So Yesterday: Metacognition and the AI Revolution We Already Missed | by Myk Eff | Spy Novel Research | Medium, acessado em julho 21, 2025, https://medium.com/spy-novel-research/the-singularity-is-so-yesterday-metacognition-and-the-ai-revolution-we-already-missed-df1821bbbaf3
16. Teaching Students AI Strategies to Enhance Metacognitive Processing, acessado em julho 21, 2025, https://www.scholarlyteacher.com/post/teaching-students-ai-strategies-to-enhance-metacognitive-processing
17. Identifying Features that Shape Perceived Consciousness in Large Language Model-based AI - arXiv, acessado em julho 21, 2025, https://arxiv.org/pdf/2502.15365
18. arxiv.org, acessado em julho 21, 2025, https://arxiv.org/abs/2505.13763
19. What Does Apple's Latest Paper Tell Us About Metacognition in AI? | by Ash Ferris - Medium, acessado em julho 21, 2025, https://medium.com/@ashjferris/what-does-apples-latest-paper-tell-us-about-metacognition-in-ai-e93b640bb82f
20. Meta-Thinking in LLMs via Multi-Agent Reinforcement Learning: A Survey - arXiv, acessado em julho 21, 2025, https://arxiv.org/html/2504.14520v1
21. Control in Act-R and Soar - CiteSeerX, acessado em julho 21, 2025, https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=95850d52f2f5

9bc05f1b53c4f9bb257264eac0d1
22. An Analysis and Comparison of ACT-R and Soar - GitHub Pages, acessado em julho 21, 2025, https://advancesincognitivesystems.github.io/acs2021/data/ACS-21_paper_6.pdf
23. An Analysis and Comparison of ACT-R and Soar - ResearchGate, acessado em julho 21, 2025, https://www.researchgate.net/publication/358148660_An_Analysis_and_Comparison_of_ACT-R_and_Soar
24. A Soar's Eye View of ACT-R John Laird 24 th Soar Workshop June 11, 2004., acessado em julho 21, 2025, https://slideplayer.com/slide/5128715/
25. An Analysis and Comparison of ACT-R and Soar by John Laird - YouTube, acessado em julho 21, 2025, https://www.youtube.com/watch?v=s-MA1iieMus
26. Computational Metacognition - arXiv, acessado em julho 21, 2025, https://arxiv.org/pdf/2201.12885
27. (PDF) Computational Metacognition - ResearchGate, acessado em julho 21, 2025, https://www.researchgate.net/publication/358260306_Computational_Metacognition
28. milvus.io, acessado em julho 21, 2025, https://milvus.io/ai-quick-reference/what-is-metareinforcement-learning#:~:text=Unlike%20traditional%20reinforcement%20learning%20(RL,scenarios%20with%20minimal%20additional%20training.
29. Meta Reinforcement Learning - GeeksforGeeks, acessado em julho 21, 2025, https://www.geeksforgeeks.org/deep-learning/meta-reinforcement-learning/
30. Reinforcement Learning: A Survey - Journal of Artificial Intelligence Research, acessado em julho 21, 2025, https://www.jair.org/index.php/jair/article/download/10166/24110/
31. Meta-Gradient Reinforcement Learning with an Objective Discovered Online, acessado em julho 21, 2025, https://proceedings.neurips.cc/paper/2020/file/ae3d525daf92cee0003a7f2d92c34ea3-Paper.pdf
32. Meta-Gradient Reinforcement Learning - NIPS, acessado em julho 21, 2025, https://papers.nips.cc/paper/7507-meta-gradient-reinforcement-learning
33. Review for NeurIPS paper: Meta-Gradient Reinforcement Learning ..., acessado em julho 21, 2025, https://proceedings.neurips.cc/paper/2020/file/ae3d525daf92cee0003a7f2d92c34ea3-Review.html
34. Meta Learning — Meta-Gradient Reinforcement Learning — An Implementation | by Hassaan Naeem, acessado em julho 21, 2025, https://hassaann.medium.com/meta-learning-meta-gradient-reinforcement-learning-an-implementation-b62c0054aafe
35. Meta-Learning with Gradient-Based Meta-Learners | by Amit Yadav | Biased-Algorithms, acessado em julho 21, 2025, https://medium.com/biased-algorithms/meta-learning-with-gradient-based-meta-learners-bdb3a7611f22
36. Is DeepSeek a Metacognitive AI? - Article (v1) by Ronaldo Mota | Qeios, acessado

em julho 21, 2025, https://www.qeios.com/read/PJ3POM.2
37. Enhancing Cognitive Robots' Knowledge Transfer through Metacognitive Strategies - ThinkMind, acessado em julho 21, 2025, https://www.thinkmind.org/articles/cognitive_2023_1_100_40061.pdf
38. Toward Robotic Metacognition: Redefining Self-Awareness in an Era of Vision-Language Models, acessado em julho 21, 2025, https://elib.dlr.de/207493/1/leidner2024toward.pdf
39. HARMONIC: Cognitive and Control Collaboration in Human-Robotic Teams - arXiv, acessado em julho 21, 2025, https://arxiv.org/html/2409.18047v1
40. HARMONIC: Cognitive and Control Collaboration in Human-Robotic Teams - arXiv, acessado em julho 21, 2025, https://arxiv.org/html/2409.18047v3
41. HARMONIC: A Framework for Explanatory Cognitive Robots - ResearchGate, acessado em julho 21, 2025, https://www.researchgate.net/publication/383557155_HARMONIC_A_Framework_for_Explanatory_Cognitive_Robots
42. HARMONIC: A Framework for Explanatory Cognitive Robots - arXiv, acessado em julho 21, 2025, https://arxiv.org/html/2409.18037v1
43. Robots Thinking Fast and Slow: On Dual Process Theory and Metacognition in Embodied AI - OpenReview, acessado em julho 21, 2025, https://openreview.net/pdf?id=iFQJmvUect9
44. A Bayesian Inference Model for Metamemory - PMC, acessado em julho 21, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC9006386/
45. Bayesian Machine Learning - Deepgram, acessado em julho 21, 2025, https://deepgram.com/ai-glossary/bayesian-machine-learning
46. Metacognitive sensitivity: The key to calibrating trust and optimal decision making with AI - PMC - PubMed Central, acessado em julho 21, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12103939/
47. From internal models toward metacognitive AI - PMC - PubMed Central, acessado em julho 21, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC8551129/
48. HMeta-d: hierarchical Bayesian estimation of metacognitive ..., acessado em julho 21, 2025, https://academic.oup.com/nc/article/2017/1/nix007/3748261
49. (PDF) Common computations for metacognition and meta-metacognition - ResearchGate, acessado em julho 21, 2025, https://www.researchgate.net/publication/376171957_Common_computations_for_metacognition_and_meta-metacognition
50. Common computations for metacognition and meta-metacognition - PMC, acessado em julho 21, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10693288/
51. Performance and Metacognition Disconnect when Reasoning in Human-AI Interaction, acessado em julho 21, 2025, https://arxiv.org/html/2409.16708v2
52. AI Makes You Smarter, But None The Wiser: The Disconnect Between Performance and Metacognition - arXiv, acessado em julho 21, 2025, https://arxiv.org/html/2409.16708v1
53. Metacognitive AI: Framework and the Case for a Neurosymbolic ..., acessado em julho 21, 2025, https://arxiv.org/pdf/2406.12147
54. EPTCS 416 Logic Programming - CSE CGI Server, acessado em julho 21, 2025,

https://cgi.cse.unsw.edu.au/~eptcs/Published/ICLP2024/Proceedings.pdf
55. Pros and cons of artificial intelligence on metacognition: A myopic state with long-term consequences on human learning - ResearchGate, acessado em julho 21, 2025, https://www.researchgate.net/publication/388499249_Pros_and_cons_of_artificial_intelligence_on_metacognition_A_myopic_state_with_long-term_consequences_on_human_learning
56. AI Has Crossed a Philosophical Threshold: New Study Argues Modern Systems Possess Free Will - ScienceBlog.com, acessado em julho 21, 2025, https://scienceblog.com/neuroedge/2025/05/13/ai-has-crossed-a-philosophical-threshold-new-study-argues-modern-systems-possess-free-will/