technologies

MDPI

Article

Harnessing Metacognition for Safe and Responsible AI

Peter B. Walker [1,*], Jonathan J. Haase [2], Melissa L. Mehalick [1], Christopher T. Steele [3], Dale W. Russell [4] and Ian N. Davidson [5]

[1] Defense Health Agency, 7700 Arlington Blvd., Suite 5101, Falls Church, VA 22042, USA; melissa.l.mehalick.civ@health.mil
[2] Evergreen Knoll Court, Alexandria, VA 22303, USA; jonathan.j.haase.mil@navy.mil
[3] Acrophase Consulting LLC, 12605, War Admiral Way, North Potomac, MD 20878, USA; ctsteele.physio@gmail.com
[4] Department of Psychiatry, Uniformed Services University, 4301, Jones Bridge Rd, Bethesda, MD 20814, USA; dale.w.russell1.civ@health.mil
[5] Department of Computer Science, University of California, Davis, CA 95616, USA; davidson@cs.ucdavis.edu
* Correspondence: peter.b.walker.mil@health.mil

Abstract: The rapid advancement of artificial intelligence (AI) technologies has transformed various sectors, significantly enhancing processes and augmenting human capabilities. However, these advancements have also introduced critical concerns related to the safety, ethics, and responsibility of AI systems. To address these challenges, the principles of the robustness, interpretability, controllability, and ethical alignment framework are essential. This paper explores the integration of metacognition—defined as "thinking about thinking"—into AI systems as a promising approach to meeting these requirements. Metacognition enables AI systems to monitor, control, and regulate the system's cognitive processes, thereby enhancing their ability to self-assess, correct errors, and adapt to changing environments. By embedding metacognitive processes within AI, this paper proposes a framework that enhances the transparency, accountability, and adaptability of AI systems, fostering trust and mitigating risks associated with autonomous decision-making. Additionally, the paper examines the current state of AI safety and responsibility, discusses the applicability of metacognition to AI, and outlines a mathematical framework for incorporating metacognitive strategies into active learning processes. The findings aim to contribute to the development of safe, responsible, and ethically aligned AI systems.

Keywords: metacognition in AI; active learning; AI safety; ethical AI; AI transparency; AI responsibility; AI robustness; AI interpretability; AI controllability; trust in AI systems; autonomous decision-making; self-regulation in AI; ethical AI frameworks; AI alignment with human values; AI risk mitigation; explainable AI; AI adaptability; responsible AI development

check for updates

Academic Editors: George F. Fragulis, Debopriyo Roy and Valeri Mladenov

Received: 10 September 2024
Revised: 9 October 2024
Accepted: 12 November 2024
Published: 6 March 2025

Citation: Walker, P.B.; Haase, J.J.; Mehalick, M.L.; Steele, C.T.; Russell, D.W.; Davidson, I.N. Harnessing Metacognition for Safe and Responsible AI. *Technologies* **2025**, *13*, 107. https://doi.org/10.3390/technologies13030107

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The rapid advancement of artificial intelligence (AI) technologies has prompted a significant shift across various sectors, from healthcare to finance, revolutionizing processes and both augmenting and, in some cases, replacing human capabilities [1,2]. However, the deployment of these systems has also raised critical concerns regarding safety, ethics, and responsibility [3]. Ensuring that AI systems operate in a manner that is both safe and aligned with human values is paramount [4].

One promising approach to address these challenges is the incorporation of metacognition (i.e., thinking about thinking) into AI systems. Metacognition involves the ability to

monitor, control, and regulate one's cognitive processes. In humans, metacognition plays a crucial role in learning, problem-solving, and decision-making, enabling individuals to assess their own understanding and adjust strategies accordingly [5,6]. Applying metacognitive principles to AI systems can potentially enhance their ability to self-monitor, evaluate their actions, and adapt their behavior in real-time, thus improving their safety, reliability, and alignment with ethical standards [7,8].

This paper explores the integration of metacognition into AI systems as a tool for promoting safe and responsible AI. Further, this paper discusses the theoretical underpinnings of metacognition, its relevance to AI, and the potential benefits and challenges of implementing metacognitive strategies in AI systems [9,10]. By leveraging metacognition, AI systems could achieve higher levels of transparency, accountability, and adaptability, fostering trust and mitigating risks associated with autonomous decision-making [11,12]. These capabilities directly address key challenges in AI, specifically robustness, interpretability, controllability, and ethical concerns (RICE), as summarized in Table 1.

Table 1. An overview of the RICE challenges in AI and how metacognition addresses them.

Challenge	Definition	Metacognition Benefit
Robustness	Ensuring AI systems operate safely in a variety of environments and settings, even those it was not explicitly trained on	Enhances self-monitoring and ethical decision-making
Interpretability	Making AI systems' decisions understandable to humans	Improves transparency and traceability of decisions
Controllability	The ability for a human to guide and adjust AI behavior as needed	Enables real-time adjustments and strategy optimization
Ethical	Aligning AI actions with ethical standards and societal norms	Supports compliance with ethical guidelines and fairness

2. Evolving Challenges in AI Safety and Ethical Governance

As artificial intelligence (AI) applications become more integral to sectors like healthcare, finance, and defense, ensuring their safety, ethical alignment, and trustworthiness has emerged as a pressing priority. The rapid development and deployment of AI technologies have at times outpaced the creation of effective safety and ethical frameworks, leading to an increasingly complex landscape of risks and regulatory needs.

A prominent challenge in AI safety is data transparency, especially concerning large language models (LLMs) such as GPT-3 and more advanced systems. These models rely on vast datasets, often compiled without thorough curation, which raises critical questions about data quality and inherent biases [13,14]. The opacity of training data sources complicates efforts to assess or mitigate potential biases, sparking concerns over the reliability and accountability of AI outputs. Additionally, reliance on proprietary data increases these concerns, as dataset specifics are typically shielded from public scrutiny, obstructing transparency and evaluation.

Another significant ethical issue involves data privacy and intellectual property rights. The prevalent practice of scraping large amounts of data from the internet often raises legal concerns, including the unauthorized use of intellectual property [15]. The potential use of sensitive or personal data without consent poses privacy risks, especially in sectors where AI models manage personal and sensitive information, such as healthcare and surveillance.

Addressing these concerns calls for stringent data governance protocols to uphold legal and ethical standards, while mitigating risks associated with data misuse.

Similarly pressing issues focus around data transparency, particularly regarding large language models (LLMs) like GPT-3, which are trained on vast, often uncurated datasets [13,14]. These datasets raise concerns about the quality and reliability of the data.

While AI's potential to augment human capabilities and foster innovation is widely acknowledged, unintended consequences present substantial ethical challenges. AI systems trained on biased datasets may inadvertently amplify societal biases, leading to discriminatory outcomes in hiring, lending, law enforcement, and more [16]. In addition, the "black box" nature of many AI systems, particularly those based on deep learning, further complicates interpretability and accountability. These opaque decision-making processes make it challenging for stakeholders to understand AI-driven conclusions, limiting transparency and hindering responsible oversight [17].

The growing autonomy of AI systems introduces further ethical complexities, as these systems become capable of making high-impact decisions without human intervention. Accountability is a key issue, as determining who is responsible for an AI's actions when something goes wrong is challenging, especially in autonomous vehicles or military applications, where AI-driven decisions can have far-reaching consequences [18]. In these contexts, ethical boundaries must be carefully defined to ensure AI systems act in alignment with human values, despite their increasing autonomy and complexity.

AI systems are prone to perpetuating societal biases, often reflecting them in outputs that influence hiring, lending, or legal decisions [16]. The proposed metacognitive framework offers a proactive method for bias detection by enabling the system to self-monitor for skewed patterns in decision-making and flag outputs that might reflect inherent biases. By leveraging metacognitive self-awareness, AI can apply corrective measures, such as adjusting weighting factors or incorporating contextually appropriate checks, which enhance fairness and alignment with societal values.

In response to these challenges, there is a growing recognition of the need for comprehensive safety protocols and ethical guidelines. Initiatives, such as those from the AI Now Institute and the European Commission, are establishing regulatory frameworks that prioritize safety, fairness, and accountability in AI use [19,20]. These efforts represent a societal push toward responsible AI development, aiming to balance technological progress with accountability and public trust. The drive for ethical governance underscores the importance of creating systems that foster transparency, inclusivity, and equitable outcomes.

3. Challenges in AI Safety and Potential Mitigations

As AI systems permeate various sectors, concerns about their safety, ethical implications, and overall trustworthiness are increasing, especially in high-stakes fields such as healthcare, finance, and defense. These areas are particularly sensitive to issues such as data bias, a lack of transparency, copyright concerns, and accountability challenges.

4. Ongoing Efforts and Frameworks for Ethical AI Deployment

To address the multifaceted challenges associated with AI safety and ethical implementation, numerous frameworks and initiatives are evolving to ensure responsible AI deployment. The establishment of AI safety frameworks provides guidelines for developing robust, transparent, and accountable systems [3]. These frameworks emphasize interdisciplinary collaboration across fields such as computer science, ethics, law, and policy-making to enhance AI safety.

A critical focus of AI safety research is on explainability and ransparency. Research into explainable AI (XAI) aims to create models that provide interpretable and accessible insights

into decision-making processes, bridging the gap between machine intelligence and human understanding [12]. Transparent AI systems are vital for building user trust and ensuring that AI-driven conclusions can be reviewed and understood by human stakeholders, a necessity in high-stakes applications like healthcare and autonomous driving.

Ethical considerations are equally essential in ongoing AI development. Various organizations, including governmental and international bodies, are establishing AI ethics guidelines that promote fairness, accountability, and respect for individual rights [21]. These frameworks strive to align AI technologies with societal values, protecting public welfare while encouraging innovation. Ethical guidelines also address critical aspects of privacy and data governance, particularly relevant as AI systems increasingly handle personal and sensitive information.

The future of responsible AI development hinges on refining these frameworks to accommodate emerging AI capabilities and societal needs. Continued advancements in AI safety and ethics research are expected to foster systems that are robust, transparent, and aligned with ethical standards, ultimately supporting a beneficial integration of AI in society.

5. Metacognition and Its Applicability to AI

Metacognition, often described as "thinking about thinking", involves the ability to monitor, control, and regulate cognitive processes. It encompasses self-awareness, reflection, and the capacity to evaluate and adjust strategies for better outcomes. While traditionally linked to human cognition, metacognition offers promising opportunities for enhancing AI systems' performance and safety. This section explores the foundations of metacognition and examines how these principles can be applied to AI systems.

Metacognitive strategies offer significant potential for addressing transparency, accountability, and adaptability. For instance, in healthcare AI applications, a metacognitive framework could help flag atypical diagnoses that do not align with the patient's history or context, improving transparency in complex diagnostic processes. In finance, a metacognitive AI system could analyze past investment strategies, allowing for adaptive learning in response to volatile market conditions.

The interaction between mood, confidence, and risk-taking within the metacognitive framework also provides nuanced adaptability. For example, an AI deployed in autonomous vehicles might adopt conservative driving strategies in inclement weather by adjusting its internal mood variable to a more cautious state. This adjustment helps mitigate the risks associated with sudden environmental changes, showcasing how metacognition enhances real-world AI applications.

5.1. Understanding Metacognition

In humans, metacognition is essential for learning, problem-solving, and decision-making. It allows individuals to critically evaluate their understanding, recognize gaps, and adjust their cognitive strategies. The concept, introduced by Flavell [5], has since been widely studied in areas such as educational psychology, cognitive science, and neuropsychology. Metacognitive processes are generally divided into two components: metacognitive knowledge and metacognitive regulation [22,23].

- Metacognitive Knowledge: This refers to an individual's awareness and understanding of their cognitive processes. It includes knowledge about cognitive strengths, weaknesses, task demands, and the effectiveness of various strategies for learning or problem-solving [23,24]. In AI systems, this can be likened to a system's awareness of its capabilities and limitations. With this knowledge, an AI system can choose appropriate strategies, anticipate challenges, and better adapt to its environment [25].

- Metacognitive Regulation: Metacognitive regulation involves monitoring and controlling cognitive activities through planning, tracking progress, and evaluating outcomes [22,24]. In AI systems, this regulation allows for the dynamic adjustment of algorithms and parameters based on real-time performance and environmental factors. AI systems that leverage metacognitive regulation can enhance efficiency, accuracy, and resilience in complex scenarios [7,26].

5.2. Applicability to AI Systems

Integrating metacognition into AI systems enhances adaptability, reliability, and transparency. Below are key areas where metacognition can significantly improve AI performance.

5.2.1. Self-Monitoring and Adaptation

AI systems equipped with metacognitive capabilities can monitor their performance in real-time, identifying areas of uncertainty or error. This enables dynamic adaptation to improve robustness and efficiency. For example, in autonomous vehicles, metacognitive AI can assess the reliability of sensory inputs and adjust decision-making algorithms in response to environmental changes [8].

Recent advances in hardware, particularly the integration of robust sensors and metacognitive principles, improve signal monitoring from devices under inclement conditions or sensor degradation. By self-assessing performance, metacognitive AI systems can dynamically adapt to environmental changes and sensor deterioration [27]. For instance, an AI system with metacognitive capabilities can detect when a camera's input quality diminishes due to fog or rain and adjust its processing algorithms accordingly or request supplemental data from additional sensors.

Additionally, the use of Field-Programmable Gate Arrays (FPGAs) for edge processing enables AI systems to make real-time adjustments, enhancing reliability with minimal latency [28]. Combining advanced hardware with metacognitive strategies empowers robotics systems to adjust their behavior based on sensor quality.

5.2.2. Error Detection and Correction

Metacognitive AI systems can identify and correct their own mistakes by recognizing error patterns and using feedback mechanisms. This capability is especially valuable in dynamic environments where traditional error correction may be insufficient. By leveraging metacognition, AI systems enhance fault tolerance and resilience [9].

5.2.3. Explainability and Transparency

Metacognitive processes contribute to improved explainability and transparency in AI systems. By providing insights into their decision-making processes, metacognitive AI systems help users understand their actions. This transparency is essential for building trust and ensuring alignment with human values [11].

5.2.4. Resource Management

Metacognitive AI systems optimize resource allocation by prioritizing tasks and focusing on areas where they are most effective. This is especially useful in scenarios where computational or data resources are limited, allowing the AI to make informed decisions about deploying its capabilities [29].

5.3. Enhanced Control of Model Behavior in Novel Conditions

Metacognition in AI introduces new levels of control and customization akin to executive functions in humans. While AI systems can be effective in familiar situations and capable of pattern recognition, they often struggle with generalizing in novel situations.

Metacognitive AI allows for real-time tuning and adaptation, helping the system adjust its outputs based on new scenarios, even when training data does not fully match the real-world situation [30].

5.4. Mathematical Framework with Mood Integration

To formalize the integration of mood into the metacognitive and active learning processes, we introduce a mood variable m_t that influences the system's decision-making:

$$Q_{\text{meta}}(f_t, \mathcal{U}, m_t) = M(f_t, \mathcal{L}, \mathcal{U}, m_t) \cdot Q(f_t, \mathcal{U}) \tag{1}$$

where m_t represents the mood of the AI system at time t, and $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ is the modulating function that now also accounts for the mood. The mood m_t can influence the modulation factor M in several ways:

1. Confidence Calibration: If the system is in a "positive" mood (e.g., after a series of successful outcomes), the modulation factor might increase, reflecting higher confidence. Conversely, a "negative" mood might decrease the modulation factor to encourage more cautious decision-making.
2. Risk Adjustment: Mood can affect the system's propensity to take risks. A more "positive" mood might result in higher risk-taking, while a "negative" mood might lead to more conservative actions.

The updated modulation factor M can be defined as:

$$M(f_t, \mathcal{L}, \mathcal{U}, m_t) = \frac{1}{1 + \exp(-\alpha \cdot (E_{\text{consistency}} - \beta + \gamma \cdot m_t))} \tag{2}$$

where γ is a parameter that controls the influence of mood m_t on the modulation factor.

5.5. Role of Mood in Modulation Factor Adjustments

The mood variable m_t represents an internal state of the AI that modulates decision-making strategies based on success and failure metrics. A "positive" mood, often correlated with successful outcomes, increases the modulation factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$, promoting more assertive actions. Conversely, a "negative" mood, influenced by recent failures or high environmental uncertainty, lowers M, prompting conservative adjustments.

The parameters α, β, and γ further refine this modulation: - α: Controls sensitivity to consistency in errors. - β: Determines baseline stability. - γ: Modulates the influence of m_t on decision-making.

These parameters allow the system to calibrate its response dynamically, adjusting behavior in real-time to reflect performance and contextual shifts.

5.6. Mathematical Proof of Stability in Mood-Driven Modulation

To demonstrate that the incorporation of the mood variable m_t into the modulation factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ results in a stable and bounded behavior in the AI system's decision-making process, this paper aims to prove that $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ is always within the range $(0, 1)$.

Recall that $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ is defined in Equation (2), where:

- α, β, and γ are parameters controlling the sensitivity of the modulation factor.
- $E_{\text{consistency}}$ is a measure of error consistency.
- m_t represents the mood of the AI system at time t.

Step 1: Exponential Function Behavior

The exponential function $\exp(x)$ is always positive for any real input x. Thus, the term $\exp(-\alpha \cdot (E_{\text{consistency}} - \beta + \gamma \cdot m_t))$ is always greater than 0:

$$\exp(-\alpha \cdot (E_{\text{consistency}} - \beta + \gamma \cdot m_t)) > 0$$

Step 2: Denominator Analysis

The denominator of $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ is $1 + \exp(-\alpha \cdot (E_{\text{consistency}} - \beta + \gamma \cdot m_t))$, which is always greater than 1:

$$1 + \exp(-\alpha \cdot (E_{\text{consistency}} - \beta + \gamma \cdot m_t)) > 1$$

Step 3: Bounded Modulation Factor

Since the denominator is always greater than 1, the modulation factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ is always less than 1 and greater than 0:

$$0 < M(f_t, \mathcal{L}, \mathcal{U}, m_t) < 1$$

Step 4: Behavior at Extremes

- Positive Extreme: When $E_{\text{consistency}} - \beta + \gamma \cdot m_t$ is large and positive, $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ approaches 1.
- Negative Extreme: When $E_{\text{consistency}} - \beta + \gamma \cdot m_t$ is large and negative, $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ approaches 0.

5.7. Detailed Stability Proof with Visual Representation

The modulation factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t) = \frac{1}{1+\exp(-\alpha \cdot (E_{\text{consistency}} - \beta + \gamma \cdot m_t))}$ remains bounded within $(0, 1)$. To illustrate this behavior, Figure 1 plots M against varying values of $E_{\text{consistency}}$ and m_t, demonstrating how adjustments in m_t influence the AI's confidence in uncertain scenarios.

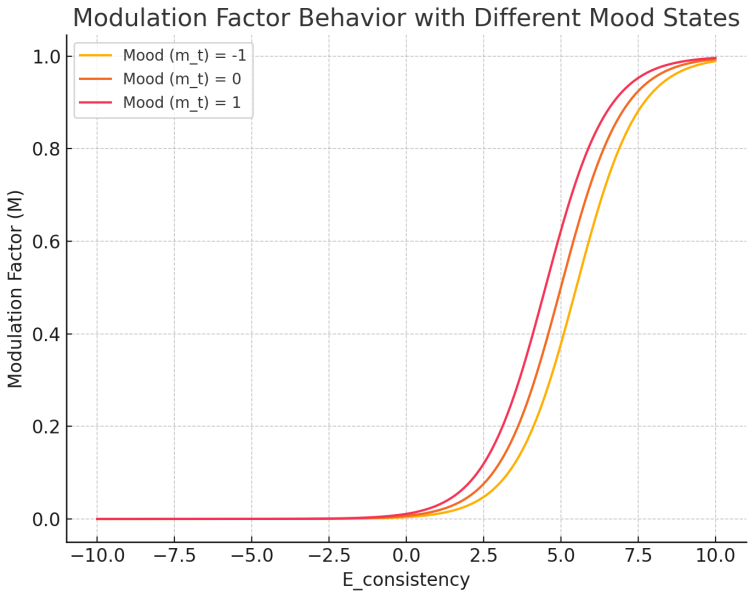


Figure 1. Modulation factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ behavior under varying $E_{\text{consistency}}$ and m_t.

5.8. Incorporating the Metacognitive Knowledge Base in the AI Model

The Metacognitive Knowledge Base (MKB) is a critical component of the AI model, serving as the repository for all information related to the AI system's cognitive processes, past experiences, and contextual understanding. This knowledge base enables the system to reflect on its actions, learn from past experiences, and make informed decisions based on its current state and environment. The MKB interacts directly with the modulation

factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ to influence the AI's decision-making, strategy selection, resource allocation, and overall adaptability.

The Metacognitive Knowledge Base comprises several key elements:

- **Cognitive Profiles:** Detailed records of the AI system's cognitive strengths, weaknesses, and performance metrics. These profiles include information on how the system has performed in various tasks, the types of errors it commonly makes, and the strategies that have historically been effective or ineffective.
- **Experience Repository:** A comprehensive archive of past experiences, including successful outcomes, failures, and the conditions under which they occurred. This repository enables the system to identify patterns in its behavior and outcomes, which can be used to inform future decisions.
- **Contextual Awareness:** Information related to the current environment, task requirements, and external factors that may influence the AI's performance. This includes real-time data on the operational context, which is crucial for making adaptive decisions.
- **Mood State Tracking:** Continuous monitoring and updating of the AI system's mood m_t, which is influenced by its experiences and current performance. The mood state is stored and referenced within the MKB, providing a dynamic input to the modulation factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$.

The modulation factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$, as defined in Equation (2), relies on the data stored in the MKB to make informed adjustments to the AI system's behavior. The MKB provides the necessary inputs for the modulation factor by:

$$M(f_t, \mathcal{L}, \mathcal{U}, m_t) = \frac{1}{1 + \exp(-\alpha \cdot (E_{\text{consistency}} - \beta + \gamma \cdot m_t))}$$

relies on the data stored in the MKB to make informed adjustments to the AI system's behavior. The MKB provides the necessary inputs for the modulation factor by:

- Providing Historical Data: The MKB offers data on past performance $E_{\text{consistency}}$, which the modulation factor uses to assess the reliability of current decision-making strategies. A consistent track record of past decisions increases $E_{\text{consistency}}$, leading to a higher modulation factor M, indicating more confidence in current strategies.
- Incorporating Mood Dynamics: The mood state m_t, tracked and updated by the MKB, directly influences the modulation factor. A positive mood, based on successful past experiences, increases $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$, leading the system to favor more aggressive strategies. Conversely, a negative mood, possibly resulting from recent failures, decreases M, prompting the system to adopt a more cautious approach.
- Contextual Adaptation: The MKB supplies real-time contextual information that affects how the modulation factor interprets current data. For example, in a high-stakes environment, the system might lower M to prioritize conservative strategies, regardless of the mood or historical consistency, reflecting a heightened awareness of external risks.

MKB plays a crucial role in enhancing the AI system's decision-making processes by ensuring that every decision is informed by a rich set of historical, contextual, and emotional data. By integrating these elements through the modulation factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$, the AI system can:

- Learn from Past Experiences: By referencing the experience repository, the system can avoid repeating past mistakes and leverage successful strategies, thus continuously improving its performance.

- Adapt to Current Conditions: The combination of mood tracking and contextual awareness ensures that the AI system remains adaptable to changing conditions, making decisions that are well-suited to the present environment.
- Balance Risk and Reward: Through the dynamic adjustment of the modulation factor, the AI system can balance the potential risks and rewards of its actions, informed by both its internal state and external pressures.

The incorporation of MKB into the AI model allows for a more nuanced and context-sensitive approach to decision-making. By interacting with the modulation factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$, the MKB ensures that the system's strategies and resource allocations are informed by a comprehensive understanding of past experiences, current conditions, and emotional states, leading to more effective and adaptive behavior.

5.9. Metacognitive Regulation in the AI Model

Metacognitive regulation is a vital component of the AI model, enabling the system to monitor, control, and adjust its cognitive processes in real-time. This self-regulatory capability allows the AI to dynamically refine its strategies, optimize resource allocation, and respond to changing environments based on ongoing feedback. The integration of metacognitive regulation into the model is directly linked to the modulation factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ and the MKB, both of which guide the system's adaptive behavior.

Metacognitive regulation within the AI model encompasses several critical functions:

- **Planning:** The system sets goals and determines the strategies required to achieve them. This involves selecting appropriate tasks, allocating resources, and preparing contingency plans based on the current state of the system and its environment. The modulation factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ influences these plans by modulating the AI's confidence in its strategies, thereby affecting the aggressiveness or caution of the plans.
- **Monitoring:** During task execution, the AI system continuously monitors its performance, comparing actual outcomes against expected results. This ongoing assessment is informed by data stored in the MKB, such as historical performance metrics $E_{\text{consistency}}$ and current mood m_t. The system uses this information to determine whether adjustments are necessary, ensuring that it remains aligned with its goals.
- **Control and Adjustment:** Based on the feedback from the monitoring process, the AI system adjusts its cognitive strategies and resource allocations in real-time. If the system detects that its performance is deviating from expected outcomes, it can modify its approach by reallocating resources, altering its strategies, or even revising its goals. The modulation factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ plays a key role here, determining the extent and nature of these adjustments based on the AI's current state and performance.
- **Evaluation:** After completing tasks, the AI system engages in a reflective process where it evaluates its overall performance, considering what worked well and what did not. This evaluation feeds back into the MKB, updating cognitive profiles, experience repositories, and mood states. The insights gained from this evaluation help to fine-tune future decision-making processes and improve the system's effectiveness over time.

Metacognitive regulation is closely tied to the modulation factor. Specifically, the modulation factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ and MKB control the processes. The interplay between these elements ensures that the AI system can regulate its cognitive activities in a manner that is both adaptive and responsive to its internal state and external environment.

- Modulation Factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$: The modulation factor adjusts the AI's level of confidence in its decisions and strategies. A higher modulation factor, driven by positive mood m_t and consistent past performance $E_{\text{consistency}}$, leads to more assertive

actions. Conversely, a lower modulation factor results in more cautious behavior, reflecting the system's assessment of increased risk or uncertainty.

- MKB provides the necessary historical data and contextual information that informs the AI's regulatory processes. By referencing the MKB, the system can recognize patterns, anticipate potential challenges, and make adjustments that are informed by past experiences. This ensures that the system's regulation is not only reactive but also proactive, preparing for future scenarios based on accumulated knowledge.

Through metacognitive regulation, the AI model enhances its adaptive behavior in several ways:

- Dynamic Strategy Adjustment: The AI system can dynamically adjust its strategies during task execution, ensuring that it remains aligned with its goals even in the face of unexpected changes in the environment or task demands.
- Real-Time Resource Optimization: By continuously monitoring performance and adjusting resource allocations, the AI can ensure that resources are used most effectively, minimizing waste and maximizing output.
- Improved Decision-Making: The reflective evaluation process after task completion allows the AI to learn from its experiences, refining its decision-making processes over time and improving its overall performance.

Metacognitive regulation is a critical function within the AI model, enabling the system to monitor, control, and adjust its cognitive processes in real-time. By interacting with the modulation factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ and MKB, the AI system can dynamically adapt to changing conditions, optimize its performance, and continually improve its decision-making capabilities. This self-regulatory capability is essential for the AI's ability to operate effectively in complex and unpredictable environments.

5.10. Adaptive Learning and Decision-Making in the AI Model

Adaptive learning and decision-making are central to the AI model's ability to continuously improve and respond effectively to new challenges. These processes allow the AI system to learn from its experiences, update its knowledge base, and make decisions that are informed by both past performance and current conditions. The incorporation of active learning further enhances the model's adaptability by enabling the AI to selectively query information that maximizes its learning efficiency.

Adaptive learning within the AI model involves the following key components:

- **Continuous Learning:** The AI system continuously updates its models and strategies based on new data and experiences. This learning process is driven by the feedback loop between the AI's actions and the outcomes it observes. The modulation factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$, as previously defined, plays a crucial role in determining how much weight is given to new versus existing knowledge.
- **Dynamic Decision-Making:** Decision-making is not static, evolving as the AI system acquires new information. The system uses the modulation factor to assess the reliability of its current knowledge and adjust its decisions accordingly. A higher $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ value, indicating high confidence, may lead to bolder decisions, while a lower value prompts more conservative choices.
- **Feedback Integration:** After executing decisions, the AI system evaluates the results and integrates this feedback into its MKB. This ongoing integration ensures that the system's strategies remain relevant and effective, adapting to changes in the environment or task requirements.

Active learning is a specialized form of adaptive learning where the AI system actively selects the most informative data points to learn from. This approach is particularly useful

when labeled data is scarce or expensive to acquire. By focusing on the most uncertain or representative examples, the AI can maximize its learning efficiency.

- **Query Strategy:** The AI system employs a query strategy $Q(f_t, \mathcal{U})$ that selects the most informative data points from the unlabeled dataset $\mathcal{U}$. The modulation factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ influences this strategy by adjusting the system's focus. For instance, a high modulation factor might prioritize data points that could confirm the system's current beliefs, while a lower factor might focus on points that challenge those beliefs.
- **Uncertainty Sampling:** One common method in active learning is uncertainty sampling, where the AI queries labels for data points that it is least certain about. The uncertainty $H(x)$ of a data point x can be calculated as:

$$H(x) = -\sum_{i=1}^{C} P(y_i|x) \log P(y_i|x)$$

  where $P(y_i|x)$ is the probability assigned by the model to class y_i given input x, and C is the number of possible classes. The modulation factor can influence this sampling by either amplifying or reducing the focus on uncertain data points, depending on the AI's current confidence and mood.
- **Model** Update: After querying the labels, the AI system updates its model f_{t+1} using the newly labeled data points. This update process is informed by the modulation factor, ensuring that the AI's learning is aligned with its overall state and strategic goals.

The AI's decision-making process is inherently adaptive, leveraging the information provided by MKB and the outcomes of its active learning queries. The steps include:

1. **Assessment:** The AI system assesses the current situation using the data stored in the MKB and the results of any recent learning activities. The modulation factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ helps determine the AI's confidence in its assessment.
2. **Strategy Selection:** Based on the assessment, the AI system selects a strategy that maximizes expected utility, taking into account the modulation factor's influence on risk and reward calculations:

$$S^* = \arg\max_S [U(S) \cdot M(f_t, \mathcal{L}, \mathcal{U}, m_t)]$$

3. **Execution and Feedback:** The selected strategy is executed, and the outcomes are observed. The AI then integrates these outcomes into the MKB, adjusting its future decision-making processes accordingly.

The integration of active learning into the adaptive learning framework enhances the AI system's ability to efficiently acquire knowledge, especially in data-scarce environments. By dynamically adjusting its learning and decision-making processes through the modulation factor, the AI can maintain high adaptability, ensuring that it remains effective in a wide range of scenarios.

Adaptive learning and decision-making are core capabilities of the AI model, enabling it to continuously refine its knowledge and strategies based on real-time feedback. The incorporation of active learning allows the AI to selectively query the most valuable data, maximizing its learning efficiency. By leveraging the modulation factor $M(f_t, \mathcal{L}, \mathcal{U}, m_t)$ and MKB, the AI system ensures that its learning and decision-making processes are dynamic, responsive, and aligned with its overall goals.

5.11. Applications and Limitations of Metacognition in AI Systems

Metacognition presents promising advantages in enhancing AI systems across multiple dimensions, particularly in terms of adaptability, transparency, and accountability.

By integrating self-assessment and adaptive response capabilities, metacognitive frameworks enable AI systems to dynamically adjust strategies and regulate cognitive processes, resulting in systems that are more flexible and responsive to real-time environmental changes. This adaptability is especially valuable for applications requiring autonomous decision-making and self-correction in complex or variable conditions.

While metacognitive architectures contribute to transparency, they do not fully mitigate the fundamental challenges of interpretability in many AI models, particularly those with "black box" characteristics. Traditional models, such as deep neural networks, inherently lack the straightforward interpretability found in simpler models like decision trees, where transparency is often complete and easily accessible. For instance, while metacognitive processes can provide insights into an AI system's high-level cognitive strategies, they may not address the opacity in source data or training processes that can obscure interpretability and reproducibility, as mentioned in lines 108–115.

This limitation is essential to note, given the interdisciplinary skepticism toward AI adoption due to issues related to trustworthiness and transparency, particularly in fields like policy-making, where full accountability and transparency are critical. As highlighted by recent research on AI skepticism across fields [31], the opacity inherent in "black box" approaches can hinder trust and adoption across sectors that demand transparent and explainable decision-making processes.

Nevertheless, the adaptability provided by metacognition is a substantial benefit, enabling systems to self-regulate and refine their cognitive strategies in response to dynamic conditions. This feature is especially valuable in areas requiring real-time adjustments, allowing metacognitive AI systems to address issues like error correction and resource optimization more effectively than static models. This adaptability can be a decisive factor in enhancing the functionality of AI in applications where strict adherence to evolving or situational criteria is paramount.

In summary, while metacognition offers significant strides toward transparency and accountability in AI, its current framework does not completely resolve the "black box" limitations. Full transparency may still require supplementary interpretability tools or hybrid approaches to meet the high standards necessary in fields like policy-making and regulated domains.

6. Challenges and Future Directions

The development of the metacognitive component within AI systems, as presented in this paper, represents a significant advancement in aligning AI with the principles of the RICE framework. This section discusses how this work contributes to these areas and outlines future directions for further enhancing these capabilities.

6.1. Advancing Responsibility Through Metacognitive Self-Regulation

The metacognitive model presented here enhances the responsibility of AI systems by enabling them to self-regulate and assess their own decision-making processes in real time. This self-regulation is critical for ensuring that AI systems act in ways that are consistent with human values and ethical norms. By incorporating metacognitive strategies, AI systems can evaluate the potential consequences of their actions before executing them, thereby reducing the likelihood of harmful or unintended outcomes. This capability is particularly important in high-stakes environments, such as healthcare or autonomous systems, where responsible decision-making is paramount. Future research should focus on refining these self-regulation mechanisms to improve their robustness across various contexts and applications.

6.2. Enhancing Interpretability Through Transparent Metacognitive Processes

Interpretability is a critical challenge in AI, particularly with complex models that function as "black boxes". Our work addresses this by embedding metacognitive processes that make the AI's reasoning more transparent and understandable to human users. The metacognitive model allows AI systems to not only make decisions but also to reflect on and articulate the rationale behind those decisions. This reflection process, which can be communicated to human operators, provides insight into how the AI arrived at a particular conclusion, thus enhancing interpretability. This is crucial in collaborative settings where human operators need to trust and understand the AI's actions. Further research should explore ways to make these metacognitive reflections even more intuitive and accessible to non-expert users.

6.3. Improving Controllability Through Adaptive Metacognitive Strategies

The controllability of AI systems is significantly enhanced by the metacognitive strategies outlined in this paper. By allowing AI systems to adapt their behavior based on real-time feedback and self-assessment, the metacognitive model presented here ensures that these systems remain flexible and responsive to changes in their environment. This adaptability is key in dynamic settings, such as military operations or real-time decision-making scenarios, where the ability to control and guide AI behavior towards desired outcomes is critical. The model's ability to adjust strategies based on self-evaluation helps maintain control even in unpredictable situations. Future research should focus on refining these adaptive strategies to ensure they are effective in a broader range of scenarios and operational conditions.

6.4. Ethical Alignment Through Metacognitive Awareness

Ethical alignment is a core objective of our metacognitive model. By enabling AI systems to reflect on the ethical implications of their decisions, our model helps ensure that AI behavior aligns with societal values and ethical standards. The metacognitive component allows AI to assess not only the immediate outcomes of its actions but also the broader ethical context in which those actions occur. For example, in law enforcement or judicial applications, the AI can use its metacognitive capabilities to evaluate whether its decisions might perpetuate biases or lead to unjust outcomes. This reflective process is crucial for developing AI systems that are not only effective but also fair and just. Future research should aim to integrate more comprehensive ethical considerations into the metacognitive model, ensuring that AI systems can navigate increasingly complex moral landscapes.

6.5. Future Research Directions

Building on the advancements made in this paper, future research should focus on the following key areas to further enhance AI systems within the RICE framework:

- **Refining Metacognitive Models for Enhanced Responsibility:** Develop more sophisticated metacognitive models that improve the AI's ability to self-regulate and make responsible decisions in diverse applications.
- **Expanding Interpretability through Enhanced Metacognitive Reflection:** Explore new methods to make the metacognitive reflections of AI systems more intuitive and accessible to a wider range of users, particularly those without technical expertise.
- **Optimizing Controllability with Advanced Adaptive Strategies:** Investigate ways to refine adaptive metacognitive strategies to maintain control over AI systems in even more complex and dynamic environments.
- **Integrating Comprehensive Ethical Considerations:** Continue to integrate broader and more nuanced ethical considerations into the metacognitive model, ensuring AI

systems can navigate complex moral dilemmas with greater awareness and alignment with societal values.

- **Empirical Validation of Metacognitive Enhancements:** Conduct empirical studies to assess the impact of metacognitive enhancements across various AI applications, identifying both the strengths and limitations of this approach.

Another important consideration is that implementing metacognitive frameworks in large-scale systems requires addressing computational complexity, especially in high-frequency decision-making environments like financial trading or autonomous navigation. The efficient management of computational resources is essential, as frequent updates to the MKB and modulation factors can increase the processing load. Exploring hybrid solutions that selectively apply metacognition in high-impact scenarios may alleviate scalability concerns.

By addressing these challenges and pursuing these research directions, the integration of metacognitive processes into AI systems can be further optimized, ultimately leading to more adaptive, transparent, controllable, and ethically aligned AI technologies.

Author Contributions: All authors contributed to the conceptualization of the study and the writing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All data are included in this manuscript

Acknowledgments: D.R. acknowledges the support of the Fulbright Scholar Program.

Conflicts of Interest: Christopher T. Steele was employed by the Acrophase Consulting LLC. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Pearson Education: London, UK, 2015.
2. Ian Goodfellow, Y.B.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
3. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete problems in AI safety. *arXiv* **2016**, arXiv:1606.06565.
4. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2017.
5. Flavell, J.H. Metacognition and Cognitive Monitoring: A New Area of Cognitive-Developmental Inquiry. *Am. Psychol.* **1979**, *34*, 906–911. [CrossRef]
6. Schraw, G. Promoting General Metacognitive Awareness. *Instr. Sci.* **1998**, *26*, 113–125.
7. Winne, P.H. Self-Regulated Learning and Metacognition in AI Systems. *Educ. Psychol.* **2017**, *52*, 306–310.
8. Cortese, A. Metacognitive resources for adaptive learning. *Neurosci. Res.* **2022**, *178*, 10–19. [CrossRef] [PubMed]
9. Metacognitive processes in artificial intelligence: A review. *J. Cogn. Syst. Res.* **2022**. Available online: https://www.sciencedirect.com/science/article/pii/S0925753522000832 (accessed on 9 October 2024).
10. Scantamburlo, T.; Cortés, A.; Schacht, M. Progressing Towards Responsible AI. *arXiv* **2008**, arXiv:2008.07326.
11. Marco Tulio Ribeiro, S.S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 2016 ACM Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
12. Gunning, D. XAI—Explainable Artificial Intelligence. *arXiv* **2019**, arXiv:1909.11072.
13. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT), Virtual, Canada, 3–10 March 2021; pp. 610–623.
14. Rao, A.; Khandelwal, A.; Tanmay, K.; Agarwal, U.; Choudhury, M. Ethical Reasoning over Moral Alignment: A Case and Framework for In-Context Ethical Policies in LLMs. *arXiv* **2023**, arXiv:2310.07251.
15. Birhane, A.; Prabhu, V.U. Large Datasets and the Danger of Automating Bias. *Patterns* **2021**, *2*, 100239.
16. Buolamwini, J.; Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT), New York, NY, USA, 23–24 February 2018.

17. Lipton, Z.C. The Mythos of Model Interpretability. *Commun. ACM* **2018**, *61*, 36–43. [CrossRef]
18. Russell, S. *Human Compatible: AI and the Problem of Control*; Penguin Random House: New York, NY, USA, 2019.
19. Institute, A.N. *AI Now 2019 Report*; Technical Report; AI Now Institute, New York University: New York, NY, USA, 2019.
20. Commission, E. Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Technical Report, European Commission. 2021. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 (accessed on 9 October 2024) .
21. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef] [PubMed]
22. Brown, A.L.; Bransford, J.D. *Metacognition, Motivation, and Understanding*; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 1987.
23. Schraw, G.; Dennison, R.S. Promoting Metacognitive Awareness in the Classroom. *Educ. Psychol. Rev.* **1994**, *6*, 351–371.
24. Pintrich, P.R. The Role of Metacognitive Knowledge in Learning, Teaching, and Assessing. *Theory Into Pract.* **2002**, *41*, 219–225.
25. Winne, P.H. Self-Regulated Learning Viewed from Models of Information Processing. In *Self-Regulated Learning and Academic Achievement*; Routledge: London, UK, 2001; pp. 153–189.
26. Taub, M.; Azevedo, R.; Bouchet, F.; Khosravifar, B. Can the use of cognitive and metacognitive self-regulated learning strategies be predicted by learners' levels of prior knowledge in hypermedia-learning environments? *Comput. Hum. Behav.* **2014**, *39*, 356–367. [CrossRef]
27. Hou, X.; Gan, M.; Wu, W.; Ji, Y.; Zhao, S.; Chen, J. Equipping With Cognition: Interactive Motion Planning Using Metacognitive-Attribution Inspired Reinforcement Learning for Autonomous Vehicles. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 4178–4191. [CrossRef]
28. Honegger, D.; Greisen, P.; Meier, L.; Tanskanen, P.; Pollefeys, M. Real-time velocity estimation based on optical flow and disparity matching. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems 2012, Algarve, Portugal, 7–12 October 2012; pp. 5177–5182. Available online: https://scholar.google.com.pk/citations?view_op=view_citation&hl=pt-BR&user=wK3xGwIAAAAJ&sortby=pubdate&citation_for_view=wK3xGwIAAAAJ:2osOgNQ5qMEC (accessed on 9 October 2024).
29. Feng Wang, J.W.; Wu, J. Deep Learning-Based Resource Management in Edge Computing Systems. *IEEE Trans. Netw. Serv. Manag.* **2019**, *16*, 886–896.
30. Hong, J.; Jeong, S. Statistical Learning for Novel Scenario Recognition in AI Systems. *Artif. Intell. Rev.* **2022**, *44*, 167–189.
31. Casey, W.; Lemanski, M.K. Universal skepticism of ChatGPT: A review of early literature on chat generative pre-trained transformer. *Front. Big Data* **2023**, *6*, 1224976. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.