Proceedings of 3rd International Conference on Interdisciplinary Humanities and Communication Studies
DOI: 10.54254/2753-7064/42/20242524

The Neural Basis and Computational Models of Metacognition

Nashao Zhong[1,a,*]

[1]*Department of Psychology, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom*
a. nashaoz@outlook.com
**corresponding author*

Abstract: The ability to reflect on one's own thinking is what makes human cognition "meta." Metacognition, the capability to assess, reflect on, and control first-order cognitive processes, is essential for flexible and adaptive behaviors across various contexts. This review explores the neural mechanisms and computational models underpinning metacognition. The involvement of brain regions, including the insula, precuneus, medial prefrontal cortex, and dorsolateral prefrontal cortex in metacognitive judgments is examined. How distinct regions support both domain-general and domain-specific metacognitive processes is also explored. Furthermore, the neural correlates of metacognitive executive functions, such as error monitoring and cognitive control, are investigated, with a focus on the prefrontal and anterior cingulate cortex and their roles in regulating working memory and performance monitoring. This review also discusses the Bayesian models of human metacognitive processes proposed by Fleming and Daw. Studies on human metacognition have significant implications for the development of artificial intelligence, evidenced by the H-CogAff architecture, revealing how integrating metacognitive frameworks could enhance AI's transparency, reasoning, adaptability, and perception. The findings suggest that investigating the neural mechanisms and computational models of metacognition is crucial not only for understanding human cognitive processes but also for improving the resilience and flexibility of AI systems. Future studies in this field should expand the scope by integrating broader and more qualitative dimensions, such as affective self-assessment and social cognition, while maintaining the precision of current evaluation approaches.

Keywords: Neural basis, computational models, metacognition.

1. Introduction

Metacognition refers to the ability to reflect on, assess, and control first-order cognitive processes, including decision-making, perception, and memory [1]. Accurate metacognition, commonly evaluated by how well subjective confidence tracks objective performance, is essential for adaptive and flexible behavior across various contexts. Dysfunctional metacognition is often associated with adverse clinical, educational, and interpersonal consequences. A primary emphasis on metacognition studies has been placed on the judgment of confidence, or error detection, as a fundamental metacognitive process that monitors first-order performances.

DOI: 10.54254/2753-7064/42/20242524

There are two primary components of metacognition – metacognitive knowledge (meta-knowledge) and metacognitive control (meta-control) [2]. Metacognitive knowledge refers to individuals' knowledge of their cognitive processes and capacity to monitor and reflect upon them [3]. Metacognitive control encompasses an individual's self-regulatory mechanisms, including planning behaviors and adaptation depending on outcomes. Nelson [4] characterized metacognitive knowledge as the information flow and processing from the object level to the meta-level; inversely, Nelson characterized metacognitive control as the information flow from the meta-level to the object level (see Figure 1). The object level comprises cognitive functions including object identification and discrimination, semantic encoding, decision-making, and representation of space. At the meta-level, top-down regulation of object-level functions is applied, and information generated from the object level is processed.

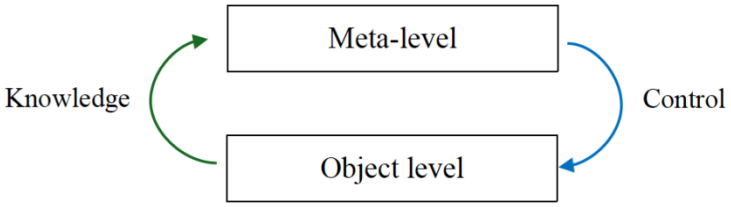

Figure 1: Metacognitive process model [3]

Executive functions (EFs) have been one primary focus of cognitive neuroscience studies for a long time, and they are tightly linked to metacognitive processes [5]. EFs and metacognition are both higher-order cognitive processes essential for goal-directed and adaptive behaviors. They share the role of self-regulation – EFs offer cognitive control to manage and adapt to tasks, and metacognition provides reflective monitoring essential for strategy and behavior adjustment. For example, academics have recently examined metacognition with regard to introspective judgments of task performance [3]. However, EFs are mainly concerned with executing top-down cognitive control, whereas metacognition adds a layer of metacognitive monitoring, which is more explicitly measured in studies. Metacognitive monitoring offers theoretical insights into the conditions and reasons for executive processes to be initiated, modified, or terminated. It also renders metacognition crucial for refining the operations of EFs.

The explosion of interest and progress in metacognitive research led to increasing attention to the development of objective instruments for metacognitive measures, such as metacognitive judgments. The advances in these measures have allowed researchers to quantify the individual-level evaluations of cognitive performance and self-assessment more precisely.

This review analyzes the neural basis and computational models underlying metacognition. The review is structured by first exploring the neural basis of metacognitive judgments and executive functions with evidence of the findings from neuroimaging studies, then examining the Bayesian computational models of metacognition, and finally considering how these findings may imply future directions for AI development, as suggested by the TRAP model, with examples of the H-CogAff architecture.

2. Neural Basis of Metacognitive Judgements

Metacognitive judgments are the decisions that learners make about how confident they were in learning a specific material [6]. The most widely employed paradigms are Judgements of Learning (JOL) [7] and Feelings of Knowing (FOK) [7, 8]. JOL consist of making predictions during learning about whether an item will be remembered or not in the future tests [8]. Conversely, FOK consists of

DOI: 10.54254/2753-7064/42/20242524

a person's confidence that they have an answer to a question at hand from the options available, even though they are unable to recall the answer directly [7]. Therefore, metacognitive judgments are rendered after the recall attempt.

Metacognitive judgments are typically measured by asking individuals to retrospectively evaluate their performance in a two-alternative forced choice task (2-AFC) [9]. The 2-AFC tasks require participants to select the alternative from the two with the greatest criterion value. The tasks include various domains, such as memory and perception. For instance, participants might be asked to identify new words from previously learned ones, detect Gabor patches with greater contrast, or visually differentiate between two boxes based on the number of dots in them [3]. The participants make metacognitive judgments by appraising their level of confidence regarding their decision in the given task. One's metacognitive bias, sensitivity, and efficiency can be evaluated depending on their responses. Metacognitive bias refers to the aggregate confidence level throughout a task; metacognitive sensitivity refers to the capacity to differentiate between accurate and inaccurate judgments; metacognitive efficiency refers to the assessment of metacognitive sensitivity while accounting for task performance [10].

A domain-general neural network linked to high versus low confidence judgments was unraveled with a recent meta-analysis of 47 neuroimaging studies on metacognition based on memory perception tasks, FOK and JOL, and decision-making tasks such as the 2-AFCs [11]. This network comprises the insula, precuneus, medial prefrontal cortex (mPFC), and lateral prefrontal cortex (lPFC). Specifically, the bilateral parahippocampal cortex was associated with the FOK and JOL, and the right anterior dorsolateral prefrontal cortex (dlPFC) was associated with the 2-AFCs. Moreover, the right insula, the posterior mPFC, and the left dlPFC were linked to the prospective judgments (JOL). Conversely, the left inferior frontal gyrus and bilateral parahippocampal cortex were linked to retrospective judgments (FOK).

Some studies also indicate that the anterior prefrontal cortex (aPFC) is associated with metacognition in 2-AFC tasks relating to perception, while the precuneus is specifically involved in 2-AFC tasks relating to memory [12,13]. This may indicate that metacognitive processes activate specific regions in a domain-specific way, whereas other regions function in a domain-general manner. A more recently published meta-analysis reveals that the domain-specific and domain-general responses may share the same circuitry. Yet, their neural signatures differ based on the kind of activity or task [14]. Additionally, Boldt and Gilbert [15] revealed that brain patterns linked to the inclination for cognitive offloading (meta-control) partially coincide with those related to meta-knowledge. This shows that meta-control is influenced by non-metacognitive and metacognitive processes or by a synthesis of domain-specific meta-knowledge processes.

3. Neural Basis of Metacognitive Executive Functions

Processes within executive function relating to meta-knowledge could include error monitoring/detection and effort monitoring [16]. Those relating to meta-control could include resource allocation, error correction, and inhibitory control. The neural correlates of these processes are examined by having participants engage in tasks with their brain activity monitored and recorded using electroencephalography (EEG) or functional magnetic resonance imaging (fMRI) [3]. The tasks commonly performed in laboratory settings include Stroop, Flanker, Demand Selection, and Motion Discrimination Tasks. Moreover, individuals with brain lesions could be assessed alongside healthy subjects to determine the functional significance of the impacted brain areas in metacognition.

In Shimamura's [17] review paper, the author considered the prefrontal cortex (PFC) to be a critical region for the top-down control and monitoring of the cognitive processes relevant to metacognition. The review suggests that a network comprising regions of the dorsomedial, dorsolateral, and ventrolateral PFC (dmPFC, dlPFC, and vlPFC) are involved in the top-down

DOI: 10.54254/2753-7064/42/20242524

regulation of cognition, enabling working memory, selective attention, and cognitive control. These regions are both interconnected and connected with cortical and subcortical regions outside the PFC [3]. Shimamura [17] proposed that the PFC controls the dynamic filtering of information, exerting executive control to amplify pertinent signals while attenuating irrelevant ones. This function is necessary for conflict resolution between different cognitive operations, which has been evidenced by research using the Stroop Tasks or working memory tasks such as the "n-back."

Neuroimaging studies have shown that different areas of the PFC are likely involved in separate control processes [17]. For example, ventrolateral PFC (vlPFC) is involved in selecting task-relevant stimuli, dorsomedial PFC (dmPFC) is involved in managing conflicting stimuli, and dorsolateral PFC (dlPFC) is involved in information manipulation in working memory. The studies also suggest that the PFC interacts with posterior sensory cortices that usually modulate cognitive functions such as learning, memory retrieval, and sensory processing. Injury to the PFC causes problems in interference control from extraneous information, resulting in impairments in the performance of tasks that necessitate cognitive control and flexibility. The review by Taylor et al. [18] explored the role of the anterior cingulate cortex (ACC) and its surrounding regions in error monitoring. It suggests the dorsal ACC (dACC) is a crucial region for error processing, which is part of a larger network involved in performance monitoring, motivation, and task adjustment. Research demonstrates that dACC is consistently engaged during error processing, but the rostral ACC (rACC) seems to handle the emotional dimensions of errors. Taylor et al. [18] indicate two theoretical models that explain error processing: Conflict Theory and Reinforcement Learning Theory. According to Conflict Theory, error arises in situations when cognitive conflict is present (e.g., due to conflicting response inclinations). The ACC monitors such as conflict and signals the need for enhanced cognitive control. According to Reinforcement Learning Theory, error processing is affected by feedback signals from negative outcomes or unmet expectations, and ACC is modulated by efferent impulses from dopamine neurons that signal such "worse-than-expected results."

4. Computational Models of Metacognition

4.1. Bayesian Models of Human Metacognitive Processes

Fleming and Daw [19] present a conceptual framework and utilize Bayesian models to explain the metacognitive processes of humans. They suggest that self-evaluation of decision-making can be modeled as a “second-order” interference, where the brain independently assesses its decision-making process. Within this framework, assessing an individual’s confidence in a decision is computationally similar to assessing such confidence in another person. The second-order computation differs from the simpler “first-order” models, in which decisions and confidence originate from the same internal state. One of the major advantages of the second-order model is its capacity to explain confidence judgments and error detection within one system. According to this model, confidence is treated as an estimate of the probability for a decision to be correct. On the other hand, error detection occurs when the confidence level drops below a certain threshold, which suggests that an error has likely arisen. However, in the first-order models, confidence and decision-making are closely coupled, making it difficult to elucidate the mechanism by which individuals identify their own errors without external feedback. The second-order model distinguishes decisions and confidence judgments, leading to a more flexible and precise explanation of self-evaluation. This model addresses the disassociation between task performance and metacognitive accuracy, accounting for the variance in people's abilities to correctly self-assess independently of their actual task performance. Therefore, the accuracy of the confidence variable and its link with decision-making may fluctuate, resulting in variations in metacognitive sensitivity. This framework also significantly contributes to explaining how actions influence confidence judgments. It posits that

DOI: 10.54254/2753-7064/42/20242524

when a decision is rendered, such an action itself yields additional information that might alter the confidence judgment. This means confidence judgments made post-action are likely to exhibit more sensitivity and accuracy than those made pre-decision. In contrast, the first-order models do not consider how actions may retroactively influence confidence. Overall, the Bayesian framework suggested by Fleming and Daw [19] offers a comprehensive and flexible explanation for various behavioral manifestations of metacognition, including confidence in decisions, error detection, and individual differences in self-evaluation.

4.2. Metacognitive AI

The studies that model human metacognitive processes have significant implications for AI development, especially in fostering more adaptive, transparent, and flexible AI systems. Kennedy [20] discussed the computational modeling of metacognition in emotion regulation in artificial agents built on the H-CogAff architecture. The model includes three layers: reactive, deliberative, and metacognitive (see Figure 2 and Figure 3). The reactive layer manages automatic and instantaneous input responses. This layer operates in a highly parallel manner, which means it can simultaneously process multiple reactions without demanding conscious deliberation. The deliberative layer involves more intricate sequential reasoning, including decision-making and planning. It plays a significant role when the agent encounters more strategic or nuanced choices. This layer takes over in situations when the automatic response of the reactive layer is insufficient.

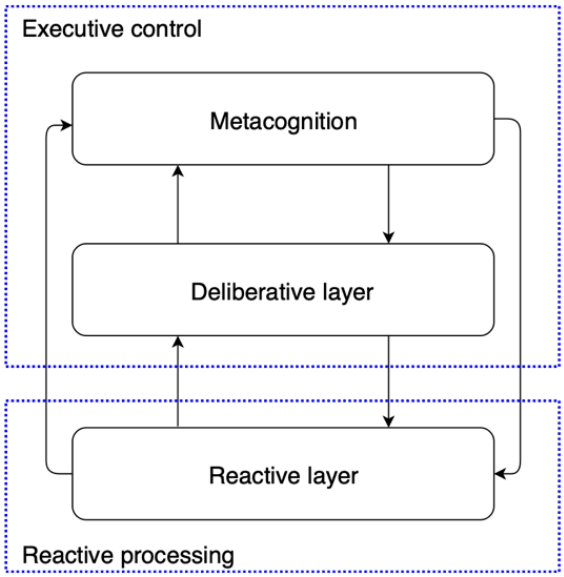

Figure 2: A simplified version of H-CogAff architecture, demonstrating the distinction between the components of executive control and the reactive layer [20]

DOI: 10.54254/2753-7064/42/20242524

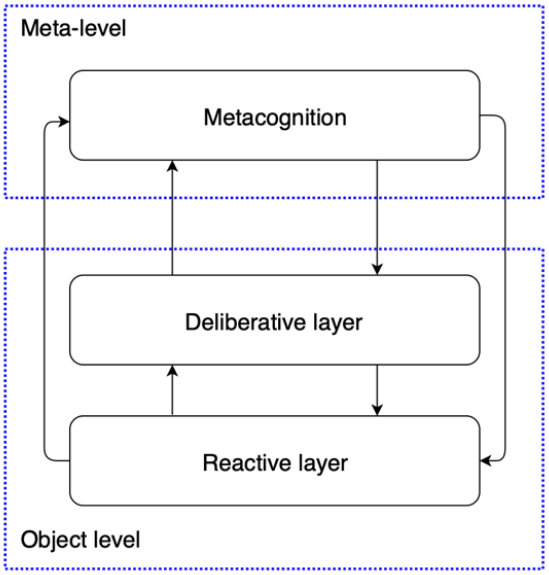

Figure 3: A simplified version of H-CogAff architecture, demonstrating the distinction between meta-level and object-level [20]

At the top of this framework sits the metacognitive layer, which oversees and steers processes in both the relative and deliberative layers. As emotions become a confounding factor for the execution of the goal of an agent, the metacognitive layer, in the case of emotion regulation, detects this and makes corrections. Strategies such as reappraisal focus on reframing emotions so that the situation or circumstances responsible for evoking these emotions have less emotional impact. Second, the model includes affective force that allows positive motivations to override negative emotions, thereby assuring goal-directed behavior. The development of this affective layer is crucial for effectively regulating emotional disturbances in AI systems in a manner similar to human emotion regulation.

Metacognitive processes, including self-assessment, error detection, and strategy adjustment, provide a framework for improving AI's capacity for real-time self-monitoring and adaptation [21]. It is consistent with the AI metacognitive framework of TRAP (transparency, reasoning, adaptability, and perception). Such a framework would empower AI systems to adapt dynamically to environmental changes. For instance, integrating metacognitive models into reinforcement learning can decrease training duration and improve adaptability, as evidenced by model-based reflection methods. These systems would be able to generalize strategies on different tasks, as shown by neural networks that are trained on multi-dimensional inputs, and were then able to use those learned rules in new environments. One example of these systems is neurosymbolic AI, which integrates neural networks with symbolic logic, offering a concrete approach for AI systems to identify and correct errors through abductive learning. Consider an AI-powered industrial robot tasked with categorizing various items according to their size and shape. By implementing a metacognitive layer through neurosymbolic AI, the system could detect discrepancies between its categorization and symbolic knowledge, such as the rules about object shapes. It might subsequently revise its categorizing rules or solicit human help to rectify the issue. Integrating such metacognitive abilities could make AI systems more resilient in fields such as autonomous driving, robotics, and healthcare diagnostics, where unforeseen circumstances and environmental variability frequently challenge current systems.

DOI: 10.54254/2753-7064/42/20242524

5. Limitations and Directions of Future Research

In the field of metacognitive research, the current issue of balancing measurement rigor with construct breadth exists [1]. Although breakthroughs in computational models have resulted in more precise approaches for evaluating confidence and monitoring performance, such a focus has restricted the scope of metacognitive research. Concerns exist that the more comprehensive, qualitative dimensions of metacognition, such as the establishment of solid metacognitive knowledge, affective self-assessment, and interactions in social cognition, are being neglected. Future studies may seek to restore the richness of metacognitive investigation by including these dimensions while maintaining the precision of existing measurement approaches. This will likely require the creation of novel tasks and models to investigate self-other evaluations, the connections between local and global metacognition, and metacognitive judgments in areas without a definitive ground truth.

6. Conclusion

This review has delved deeply into the neural mechanisms of human metacognition, computational models of metacognition of humans, and the implications of metacognitive studies for AI development. Metacognitive judgments, including assessments of confidence in knowledge or decision-making, are commonly quantified through experimental paradigms such as JOL and FOK and are associated with brain regions of the insula, precuneus, and PFC. These areas are specifically activated when processing memory and perceptual tasks. Furthermore, the PFC and ACC are closely connected with executive functions in metacognition, such as error detection and cognitive control. The PFC regulates tasks like working memory and attention, while the ACC is responsible for monitoring errors and processing the cognitive and emotional aspects of errors. The Bayesian framework proposed by scholars models the human metacognitive process as a second-order computational process, which not only explains how metacognitive judgments and error detection can be independent of task performance but also provides a more flexible model for self-assessment. In addition, metacognitive research has significant implications for the development of artificial intelligence, promoting the creation of metacognitive AI frameworks such as the H-CogAff architecture, which are better qualified in transparency, reasoning, adaptability, and perception.

Despite the progress made in metacognitive research, it still faces the challenge of balancing precise measurement with the inclusion of broader and more qualitative dimensions of metacognition, such as social cognition and self-assessment. Future research should expand the scope of research to include these dimensions while maintaining the rigor of evaluation methods, which may require the development of new models and tasks.

References

[1] *Katyal, S., & Fleming, S. M. (2024). The future of metacognition research: Balancing construct breadth with measurement rigor. Cortex, 171, 223–234.*

[2] *Flavell, J. H., & Wellman, H. M. (1975). Metamemory. Institute of Child Development, University of Minnesota. National Institute of Child Health and Human Development, National Science Foundation. ERIC*

[3] *Fleur, Damien S., Bredeweg, B., & van den Bos, W. (2021). Metacognition: Ideas and insights from neuro- and Educational Sciences. Npj Science of Learning, 6(1).*

[4] *Nelson, T.O. (1990). Metamemory: A Theoretical Framework and New Findings. Psychology of Learning and Motivation, 26, 125-173.*

[5] *Roebers, C. M. (2017). Executive Function and Metacognition: Towards a unifying framework of cognitive self-regulation. Developmental Review, 45, 31–51.*

[6] *Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. Philosophical Transactions of the Royal Society of London. Series B, Biological sciences, 367(1594), 1338–1349.*

[7] *Hart, J. T. (1965). Memory and the feeling-of-knowing experience. Journal of Educational Psychology, 56(4), 208–216.*

DOI: 10.54254/2753-7064/42/20242524

[8] *Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. Journal of Experimental Psychology, 81(1), 126–131.*
[9] *Fechner, G. T. (1948). Elements of psychophysics, 1860. In W. Dennis (Ed.), Readings in the history of psychology,206–213.*
[10] *Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. Science, 329(5998), 1541–1543.*
[11] *Vaccaro, A. G., & Fleming, S. M. (2018). Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. Brain and neuroscience advances, 2.*
[12] *Baird, B., Smallwood, J., Gorgolewski, K. J., & Margulies, D. S. (2013). Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. The Journal of neuroscience: the official journal of the Society for Neuroscience, 33(42), 16657–16665.*
[13] *McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., de Lange, F. P., & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. The Journal of neuroscience: the official journal of the Society for Neuroscience, 33(5), 1897–1906.*
[14] *Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. Biological psychiatry, 84(6), 443–451.*
[15] *Boldt, A., & Gilbert, S. J. (2019). Confidence guides spontaneous cognitive offloading. Cognitive Research: Principles and Implications, 4(1).*
[16] *Fernandez-Duque, D., Baird, J. A., & Posner, M. I. (2000). Executive attention and Metacognitive Regulation. Consciousness and Cognition, 9(2).*
[17] *Shimamura, A. P. (2008). A neurocognitive approach to metacognitive monitoring and control. In J. Dunlosky & R. A. Bjork (Eds.), Handbook of metamemory and memory, 373–390. Psychology Press.*
[18] *Taylor, S. F., Stern, E. R., & Gehring, W. J. (2007). Neural systems for error monitoring: recent findings and theoretical perspectives. The Neuroscientist: a review journal bringing neurobiology, neurology and psychiatry, 13(2), 160–172.*
[19] *Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. Psychological Review, 124(1), 91–114.*
[20] *Kennedy, C. M. (2018). Computational modelling of metacognition in emotion regulation.*
[21] *Wei, H., Shakarian, P., Lebiere, C., Draper, B., Krishnaswamy, N., & Nirenburg, S. (2024). Metacognitive AI: Framework and the Case for a Neurosymbolic Approach. In International Conference on Neural-Symbolic Learning and Reasoning, 60-67.*