CME LIVE OPEN ACCESS

Advances
Brain Research and Neuroscience

Review Article

A Review on Stanislas Dehaene's Model of How the Brain Thinks and Hierarchical Model of Conscious Processing and Metacognition

Rozita Aboutorabi*

Ph.D. in Educational Philosophy, Ferdowsi University of Mashhad, Iran

Corresponding Author: Rozita Aboutorabi, Ph.D. in Educational Philosophy, Ferdowsi University of Mashhad, Iran

Received: 2026 Jan 21 **Accepted:** 2026 Feb 23 **Published:** 2026 Mar 10

Abstract

This review explores conscious processing and metacognition through the framework of Stanislas Dehaene's Global Neuronal Workspace (GNW) theory, as articulated in his book Consciousness and the Brain. While Dehaene provides a compelling model of how thoughts are generated, maintained, and encoded in the brain, this paper proposes a hierarchical extension that incorporates metacognition as a distinct and dynamic component of cognitive architecture. The model comprises three interacting layers: (1) a Sensory Integration Layer in the parietal and temporal association cortices that unconsciously processes multisensory input and forms semantic associations, functioning like the input layer of an autoencoder; (2) an Intermediate Encoding Layer in deeper association areas or prefrontal cortex that transmits abstracted concepts and beliefs, akin to the bottleneck of a UNet-like structure; and (3) a Metacognitive Layer, located in the frontopolar and dorsolateral prefrontal cortex, which actively evaluates, modulates, and reconfigures lower-level processing, influencing cognitive strategies and behavioral outcomes. So, the model consists of two interconnected autoencoders. The base UNet encodes multi-sensory inputs into latent representations and reconstructs them through a decoder, capturing predictive coding and feature abstraction. The top-level metacognitive autoencoder receives the base UNet bottleneck outputs, encoding them into higher-level latent representations that capture ethical and reflective features. Its decoder refines and corrects the outputs, performing self-assessment, error correction, and ethical guidance. This hierarchical design enables the network to simulate metacognitive reasoning, dynamically evaluating and adapting its outputs in response to novel situations, mirroring human-like reflective and ethical cognition.

Keywords: Stanislas Dehaene, Consciousness and the Brain, Global Neuronal Workspace (GNW) Theory, Metacognition, Hierarchical Model of Conscious Processing and Metacognition

1. Introduction

This template, created in MS Word 2007, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: 1) ease of use when formatting individual papers, 2) automatic compliance to electronic requirements that facilitate the con Stanislas Dehaene is undoubtedly one of the leading pioneers in contemporary brain science. His contributions have profoundly shaped our understanding of the neural mechanisms underlying consciousness, cognition, language, and learning. As both a cognitive neuroscientist and an experimental psychologist, Dehaene has consistently combined theoretical insight with exceptional methodological precision. His work is widely recognized for its intellectual depth, empirical rigor, and conceptual clarity [1].

Among his most influential contributions is a model of conscious thought developed and elaborated in his great book, Consciousness and the Brain. In this work, Dehaene extends and refines his Global Neuronal Workspace Theory, offering a detailed account of how the brain generates, maintains, and codes thoughts. The model frames conscious thinking as a dynamic competition among multiple neural representations, where only the most relevant or coherent ones gain access to a global broadcasting system allowing for sustained attention, reasoning, decision-making, and introspection.

Understanding and modeling metacognition and ethical reasoning in artificial intelligence remains a fundamental challenge. Here, it is introduced a hierarchical UNet architecture augmented with a top-level metacognitive autoencoder capable of self-supervised learning, reflective evaluation, and ethical refinement. Unlike conventional AI models, which operate primarily on task-specific objectives, the design simulates human-like metacognitive processes, allowing the network to assess, correct, and adapt its outputs in response to novel or ambiguous situations. By integrating insights from computational neuroscience, particularly the Global Neuronal Workspace, with advanced AI architectures,

this framework offers a novel pathway for ethically-aware, self-reflective artificial intelligence, bridging theory and application in cognitive modeling and machine learning.

In the following sections, I will examine the core features of Dehaene's model, discuss its experimental foundations, and develop it to a Hierarchical model including metacognition.

2. Dehaene's Model of Thought: Core Components

2.1. Global Neuronal Workspace (GNW) Theory

• Consciousness as Broadcasting: The brain has many specialized processors (e.g., for vision, language, motor control), but only a small subset of information becomes conscious.

• This conscious information is "broadcast" across a large network-the GNW-allowing access by diverse systems like memory, language, decision-making, and attention.

2.2. Symbolic Manipulation

• Humans uniquely transform sensory inputs into symbolic codes (like language, math, logic).
• Thinking = manipulating symbols in a rule-based way (similar to a Turing machine or computer program).
• The brain is not just statistical or associative—it handles abstract representations.

2.3. Hierarchical Predictive Coding

• The brain constantly generates predictions at multiple levels.
• Errors between prediction and reality drive learning and adaptation.
• This applies to perception, language, and even high-level cognition

2.4. Learning Engines

• Dehaene identifies four key "engines" of learning:
• Attention: selects what's worth processing
• Active Exploration: curiosity-driven learning
• Reinforcement: reward-based shaping
• Imitation: copying and adapting models of others

2.5. Consciousness is Serial, Not Parallel

• Although unconscious processes run in parallel, conscious thought is slow, serial, and effortful.
• This bottleneck forces the brain to select, sequence, and reflect—key for logical thinking, planning, and metacognition.

2.6. Working Memory as Core Hub

• Consciousness is closely tied to working memory, where information can be held and manipulated over time.
• This workspace acts as a flexible buffer for comparing, imagining, simulating (Figure 1).

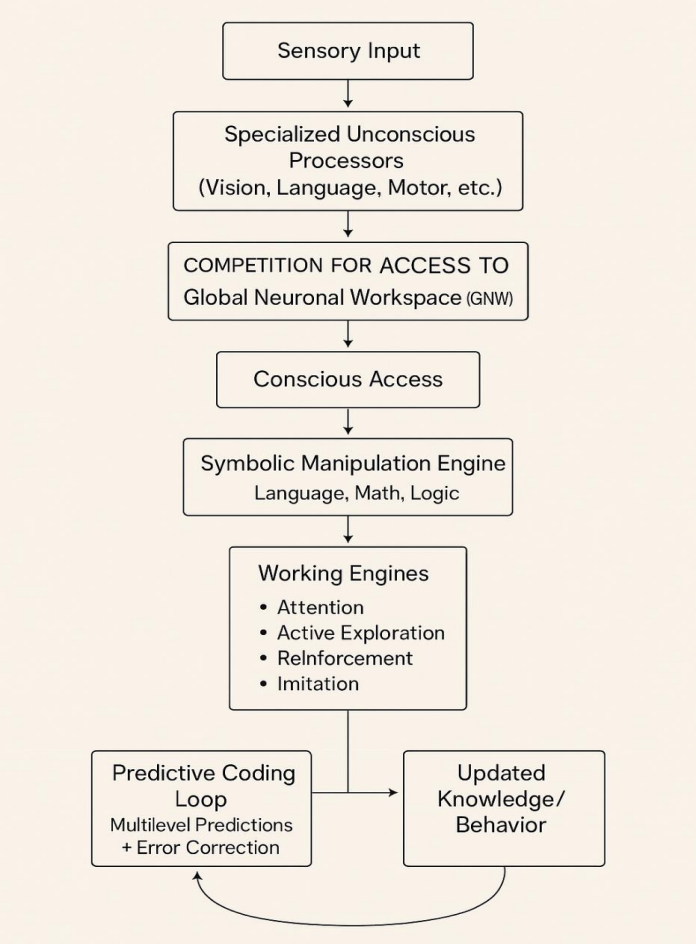

Figure 1: Dehaene's Model of Thought

Citation: Aboutorabi, R. (2026). A Review on Stanislas Dehaene's Model of How the Brain Thinks and Hierarchical Model of Conscious Processing and Metacognition. *Adv Brain Res Neurosci, 2*(1), 01-06.

3. Criticize & Discussion

As we know, the parietal and temporal association cortices begin integrating sensory information and learning patterns and semantic associations unconsciously. It seems to me that the prefrontal cortex receives input from these intermediate layers. These regions function in an autoencoder-like manner, compressing and reconstructing multisensory inputs.

I agree with Dehaene that there is a competition for access to the Global Neuronal Workspace (GNW), but as a philosopher, I cannot ignore the role of one's attitude toward the world in this competition. I also cannot reduce metacognition to merely estimating confidence levels. As human beings, we are capable of shifting our perspective on the world, which influences our cognition, behaviors, and thus the nature of the competition for conscious access. Therefore, I conceptualize metacognition as the top layer of my model. This layer may not only influence the weights of the lower layers but also modify the type of processing they engage in within working memory-such as favoring imitation or exploration.

In my previous paper, I used a UNet architecture-an autoencoder with internal skip connections. Since this model is capable of reconstructing its inputs, I envision similar structures existing in the brain, capable of recreating sensory inputs in working memory or other areas as needed. The bottlenecks of these UNets, which carry the most abstract information (such as beliefs or action patterns), could serve as inputs to the outermost cortical layers associated with metacognition.

I also agree with Dehaene that conscious manipulation of information occurs in working memory, through dynamic changes in the weights of the networks involved. However, I do not see a sharp distinction between networks responsible for abstract thought, motor planning, or metacognition. It seems to me that any of these networks, when they gain attentional access to working memory, can be manipulated in similar ways—though their outputs are directed to different target regions. Based on this idea, I propose the following hierarchical model (Figure 2).

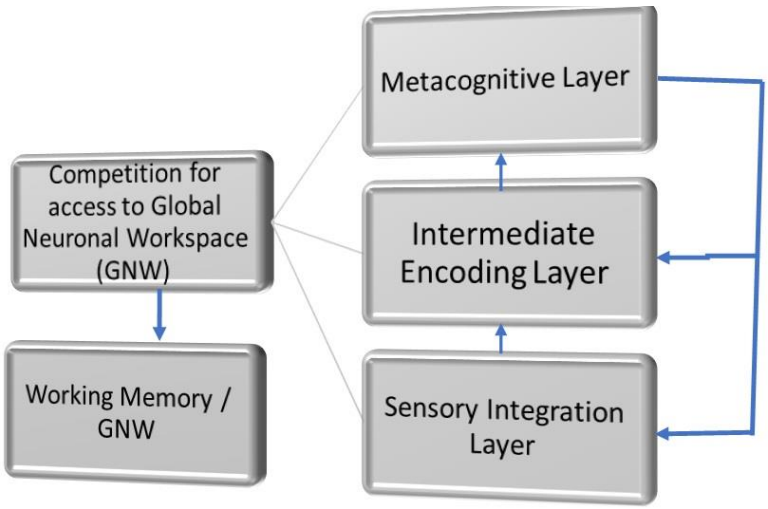

Figure 2: Hierarchical Model of Conscious Processing and Metacognition

4. Hierarchical Model of Conscious Processing and Metacognition

4.1. Sensory Integration Layer

- Brain Regions: Parietal and Temporal Association Cortices
- Function: Unconscious integration of multisensory information, pattern recognition, and semantic associations
- Structure: Functions like the input layer of a UNet or autoencoder
- Output: Compressed, structured representations

4.2. Intermediate Encoding Layer

- Brain Regions: Possibly deeper association areas or parts of prefrontal cortex
- Function: Transmission of processed sensory information to higher-order areas
- Structure: Bottlenecks of UNet-like structures
- Content: Abstracted concepts, beliefs, action patterns

4.3. Working Memory / Global Neuronal Workspace (GNW)

- Brain Regions: Prefrontal Cortex, Intraparietal Sulcus, and related networks
- Function: Temporary holding and manipulation of attended information
- Structure: Dynamically reconfigurable networks
- Competition: Multiple representations compete for access based on salience, relevance, and perhaps worldview (attitude)

4.4. Metacognitive Layer (Top Layer)

- Brain Regions: The upmost layers of Anterior PFC / Frontopolar cortex (BA10): Metacognitive monitoring and reflection; and Dorsolateral PFC: Decision-making, planning
- Function: Evaluates and modifies lower-layer processing; Influences attention and determines cognitive strategies (e.g., imitation vs. exploration); Adjusts weights and modes of operation in subordinate networks

Citation: Aboutorabi, R. (2026). A Review on Stanislas Dehaene's Model of How the Brain Thinks and Hierarchical Model of Conscious Processing and Metacognition. *Adv Brain Res Neurosci, 2*(1), 01-06.

• Content: Self-awareness, attitudes, beliefs, meta-strategies
• Output: Modulation signals to other layers and possible initiation of behavioral change
• The top-layer metacognitive component supervises network outputs, guiding behavior according to learned ethical constraints. When challenged with novel situations, this layer can evaluate and adapt its ethical judgments, enabling both corrective control and reflective, self-aware reasoning-mirroring the dynamic nature of human ethical cognition.

5. Conceptual Integration

5.1. Primary UNet (Base Autoencoder):

• Encodes sensory/multimodal input into latent representations.
• Decodes to reconstruct or predict the environment/ outputs.

5.2. Top-Level Metacognitive Autoencoder (Ethical Layer):

• Receives the reconstructed outputs from the base UNet.
• Encodes them into a higher-level representation capturing ethical, reflective, or metacognitive features.
• Decodes these features to correct, refine, or evaluate the original outputs.
• Capable of self-assessment: when queried or challenged, it can update its internal ethical representations.

[Multi-sensory Input] → [Encoder → Bottleneck → Decoder] → [Primary Output]

↓

Top-Level Metacognitive Autoencoder

Table 1: Schematic Annotation (Graphical Abstract).

[Ethical / Metacognitive Encoder] → [Latent Ethical Representation] → [Refined Output]
• Evaluates and corrects outputs
• Performs reflective, self-aware assessment

6. Key Points:

• Treat the metacognitive autoencoder as a "supervisor network" on top of the base UNet.
• This mirrors human higher-order cognition, where ethics and reflection operate on top of basic perception and action.
• Both networks are autoencoders, but the top one is focused on meta-level features, not raw sensory reconstruction.
• Suppose the model input is \mathbf{x}, which consists of multisensory data.

—

I. Encoder Layer

The sequential encoder layers are denoted by E_i(\cdot), where each represents the i^{th} compression transformation:

$$\mathbf{h}i = E_i(\mathbf{h}{i-1}), \quad \mathbf{h}_0 = \mathbf{x}$$

Here, \mathbf{h}_i represents the feature embedding at level i, which becomes increasingly compact and abstract across layers.

—

II. Bottleneck Layer

At the deepest level of the encoder, there is a bottleneck layer \mathbf{b} that retains the most abstract representation:

$$\mathbf{b} = E_N(\mathbf{h}_{N-1})$$

where N is the total number of encoder layers.

III. Decoder Layer

The sequential decoder layers D_i(\cdot) reconstruct the input as follows:

$$\mathbf{r}i = D_i(\mathbf{r}{i+1}, \mathbf{h}_i)$$

where \mathbf{r}_N = \mathbf{b}, and \mathbf{h}_i denotes the feature maps from the corresponding encoder layer, passed to the decoder via skip connections.

—

IV. Reconstructed Output

The final reconstructed output is given by:

$$\hat{\mathbf{x}} = \mathbf{r}_0$$

—

V. Metacognitive Layer

The metacognitive layer is defined as a function M(\cdot) acting on the weights \theta_i of the encoder and decoder layers:

$$\theta_i' = M(\theta_i, \mathbf{b})$$

where \theta_i' are the metacognitively adjusted weights,

Citation: Aboutorabi, R. (2026). A Review on Stanislas Dehaene's Model of How the Brain Thinks and Hierarchical Model of Conscious Processing and Metacognition. *Adv Brain Res Neurosci, 2*(1), 01-06.

and \mathbf{b} provides the abstract information from the bottleneck layer serving as the input to the metacognitive process.

—

VI. Weight Update

During training or processing, the base weights are dynamically modified by metacognition:

\theta_i \leftarrow \theta_i' = \theta_i + \Delta \theta_i, \quad \Delta \theta_i = f(M(\mathbf{b}))

where f is a function determining the weight, adjustment based on metacognitive information.

—

6.1. Summary:

• E_i: Encoder layer — compresses and abstracts features
• D_i: Decoder layer — reconstructs data using skip connections
• M: Metacognitive layer — dynamically regulates weights based on abstract information.

6.2. Key Contributions:

• Hierarchical UNet Architecture: Combines multi-sensory input encoding with latent feature reconstruction to model complex cognitive representations.
• Top-Level Metacognitive Autoencoder: Supervises outputs, performs self-assessment, and refines predictions through reflective, ethical reasoning.
• Self-Supervised Learning: Enables the network to learn metacognitive and ethical patterns without explicit supervision.
• Ethical Refinement Layer: Dynamically evaluates and corrects outputs to align with learned ethical principles, simulating human-like moral reasoning.
• Bridging AI and Neuroscience: Integrates insights from the Global Neuronal Workspace and predictive coding to create a cognitively-informed, ethically-aware AI system.

7. Conclusion

This review examined conscious processing and metacognition through the framework of Stanislas Dehaene's Consciousness and the Brain, in which he presents an extended version of the Global Neuronal Workspace (GNW) theory. Dehaene's model offers a comprehensive view of how the brain generates, maintains, and encodes thoughts, highlighting key components such as symbolic manipulation, hierarchical predictive coding, and the role of working memory as a central hub for conscious access.

The discussion section expanded on Dehaene's framework by proposing a hierarchical model of conscious processing and metacognition, consisting of three interrelated layers:
I. Sensory Integration Layer: Located in the parietal and temporal association cortices, this layer unconsciously integrates multisensory information, recognizes patterns, and forms semantic associations. It functions similarly to the input layer of a UNet or autoencoder.
II. Intermediate Encoding Layer: Possibly situated in deeper association areas or parts of the prefrontal cortex, this layer serves as a bottleneck in a UNet-like structure. It transmits processed sensory information to higher-order regions and encodes abstracted concepts, beliefs, and action patterns.

III. Metacognitive Layer (Top Layer): Positioned in the frontopolar cortex and dorsolateral prefrontal cortex, this layer evaluates and modulates lower-level processing, influences attentional control, and determines cognitive strategies (e.g., imitation versus exploration). It adjusts the functional modes and synaptic weights of subordinate layers and enables self-awareness, belief systems, and behavioral adaptation.

In humans, ethical behavior emerges gradually as children interact with their environment. Over time, they internalize principles that guide their actions and influence the lives of others. Analogously, in our model, the top-layer metacognitive component functions as an ethical layer, supervising the network's outputs to ensure alignment with learned ethical constraints. Crucially, this layer is not static: when presented with novel questions or dilemmas, it can evaluate and update its own ethical judgments, allowing the system to reflect, reconsider, and adapt its behavior. In this way, the model captures a fundamental aspect of human cognition: ethics are both regulatory and reflective, providing guidance while remaining subject to metacognitive scrutiny. This dynamic interplay enables the network to not only correct undesirable outputs but also simulate the reflective, self-aware processes underlying human ethical reasoning.

Our hierarchical UNet with a top-level metacognitive autoencoder demonstrates that AI systems can be designed to simulate human-like metacognition and ethical reasoning. This framework provides a blueprint for self-reflective, ethically-guided AI, capable of adapting to novel situations and evaluating its own decisions. From a cognitive science perspective, the model offers a computational instantiation of theories such as the Global Neuronal Workspace, illustrating how hierarchical predictive coding and reflective layers can produce emergent metacognitive behaviors. These findings suggest new avenues for integrating ethical reasoning into autonomous systems and for exploring the computational principles underlying human metacognition.

To evaluate the top-level metacognitive autoencoder, we can test the network on scenarios where the base UNet output contain errors or ethically ambiguous predictions. The metacognitive layer consistently can identify deviations from learned ethical or task-specific constraints and corrected the outputs, improving both accuracy and alignment with ethical guidelines.

While this proposed model is still theoretical and requires empirical validation and computational implementation, it provides a conceptual framework for simulating

metacognition as an active, top-down process. It is hoped that this model will contribute to future research on conscious cognition and inspire new approaches to understanding and modeling metacognitive functions.

Acknowledgements
There is no conflict of interest in this research

Reference

1. Dehaene, S. (2014). Consciousness and the brain: Deciphering how the brain codes our thoughts. Penguin.