Can Large Language Models Think and Experience Emotions and Sensations?

Janusz A. Starzyk, Wiesław L. Galus

Abstract

The rapid development of large language models (LLMs) raises fundamental questions about the extent to which their "thinking" resembles human cognition, emotions, and subjective experiences. Is embodiment – including a body sensory apparatus, and the ability to act in the world - necessary for such thinking to occur? This article addresses the issue from two complementary perspectives. First, we examine practical advances in generative AI, with emphasis on multimodal systems and robotic platforms. Second, we offer a theoretical analysis of the differences between these systems and the cognition and phenomenal consciousness of living organisms, grounded in a new framework: the Motivated Emotional Mind (MEM) model. We identify fundamental differences between artificial and biological systems in terms of knowledge acquisition and structuring; methods of categorization and generalization of sensory inputs; formation of their representations; and associative processes that enable deductive, inductive, and abductive reasoning. We argue that the pursuit of human-like thinking and feeling is only meaningful in embodied robotic systems. Finally, we advance the hypothesis that the emergence of subjective sensations and feelings requires satisfaction of the criteria specified by the MEM model, which we propose as a necessary condition for phenomenal consciousness. Thus, the MEM model may guide AI design, inform philosophy of mind, or shape ethical debates on machine consciousness.

Keywords: Large language models (LLMs)**;** Embodied artificial intelligence**;** Phenomenal consciousness**;** Motivated Emotional Mind (MEM) model**;** Interoception and emotion; Secondary perception; Feeling and emotions;

1. Introduction

The rapid progress of large language models (LLMs) has revived long-standing questions about whether systems built on them can genuinely think and feel in ways comparable to humans. Enthusiastic claims and skeptical rebuttals abound, yet the debate often fragments because participants rely on divergent concepts. What, precisely, counts as thinking? What do we mean by feeling? How should we characterize subjective sensory experience (phenomenal consciousness) and affective states? What properties are necessary for thinking and feeling, and must those properties be constitutive of consciousness rather than merely correlated with it? Do current LLMs (e.g. ChatGPT, Gemini, multimodal robotics) instantiate any of these properties, and might future systems do so?

A coherent answer requires more than ad hoc criteria; it calls for a functional model of a system that could, at least in principle, demonstrate the relevant capacities, or that identifies structural–functional variants capable of partial realizations of them. A tempting strategy is to emulate the organization of human or animal brains. Yet our understanding of phenomenal consciousness remains incomplete. Prominent theories illuminate aspects of experience, but none explain the full complexity of the mental life we report. They do not explain how memories, imaginings, and dreams gain their quasi-perceptual character; what the ongoing stream of thoughts, sounds, and images fundamentally consists of during wakefulness and sleep; and how affect integrates with perception and cognition. To evaluate multimodal, generative AI systems in a principled way, we need a model that is both neurocognitively grounded and explanatorily integrated.

This article uses the Motivated Emotional Mind (MEM) model of phenomenal consciousness to answer the question posed in the title: Can different Large Language Models and intelligent robots think and experience emotions and sensations in a human way?

The MEM model, which aligns with the theoretical frameworks of enactivism and affective neuroscience, represents information integration within systems that conceptualize intelligence as an emergent property. It offers a framework for understanding cognition as a process grounded in the dynamics of real-world interactions. Within MEM, motivation, emotion, and goal-directed behavior function as integral components of dynamic development, contributing to the self-organization and adaptive evolution of its internal structures. MEM proposes a neurocognitively plausible architecture in which motivation and emotion play constitutive roles in shaping conscious contents through reentrant interactions with sensory systems. We argue that MEM is sufficiently **structured** to inform debates on AI consciousness, yet sufficiently **parsimonious** to explain a wide range of psychological phenomena - perceptual qualia, imagery, memory reactivation, and affective experience, without presupposing computational implementations.

The research problem addressed in this paper is the question: *"Can contemporary large language models (LLMs) be considered thinking or conscious systems?"* The aim of the article is to evaluate this question in the light of the MEM model.

In the following sections, we examine whether LLMs can think and whether they can feel images and sounds. Our approach is intentionally cognitive and neurological rather than engineering-computational. Before presenting the MEM framework, we review the main existing theories and relevant empirical findings, both to situate our proposal and to show how MEM reinterprets shared evidence. We then articulate the properties that MEM considers necessary for thinking and feeling, clarify their constitutive versus merely causal status, and assess whether current or future LLM-based systems satisfy any subset of these requirements. In the concluding section, we discuss the prospects of future machines that can think and experience emotions and sensations.

2. The Capacity of LLMs for "Thinking" and Reasoning

Recent advances in large language models have led some researchers to hypothesize that these systems exhibit elements of general intelligence, raising doubts about whether their impressive problem solving should be regarded as a genuine form of thinking or merely as highly skilled pattern exploitation. In this section, however, we use the term thinking in a purely functional sense—to denote sophisticated computational abilities such as inference, planning, and problem solving—without making any claim that such abilities are accompanied by conscious experience.

On this functional reading, state-of-the-art LLMs can be said to "think" insofar as they implement powerful forms of statistical inference over high-dimensional embeddings. They can solve novel, complex problems from diverse domains (mathematics, programming, medicine, law, etc.) at near-human levels. Brubeck et al. (2023) observe that GPT-4 performs remarkably well on such tasks without specialized prompting, suggesting it could be considered an early form of artificial general intelligence (AGI). Proponents claim that such models display "sparks" of general reasoning that extend beyond simple pattern matching.

Studies suggest that mere scaling (increasing parameter counts and more training data) gives rise to emergent abilities, prompting the slogan "*Scale is all you need*". Some researchers anticipate that at sufficiently large scales, especially in multimodal architectures, LLMs could approach human-like intelligence and understanding (Mitchell & Krakauer, 2023). For instance, Google's LaMDA persuaded one engineer that it "in a very real sense understands a wide range of concepts" and is progressing toward self-awareness. Studies suggest that models can assess their own outputs—distinguishing reliable from unreliable answers—implying a primitive form of metacognition or "knowledge about their own knowledge." Kosinski (2023) notes that a theory-of-mind-like ability may have emerged spontaneously in ChatGPT as a byproduct of improving language skills. Some researchers maintain that the latest LLMs demonstrate forms of language understanding and even rudimentary concept modeling and have argued that they might be progressing toward self-awareness and primitive metacognition. In what follows, we critically examine these claims while postponing the question of conscious thinking to later sections.

Other researchers remain skeptical. Despite their impressive performance, these models are

prone to illogical errors and "hallucinations", generating answers that sound convincing yet are false. Critics argue that the inability to learn from experience, as humans do, constrains their reasoning. From a cognitive perspective, these systems rely on statistical correlations in text rather than on grounded world knowledge, raising doubts about whether their problem solving qualifies as genuine thinking or merely skilled pattern exploitation. In short: LLMs undoubtedly solve complex tasks, but it remains contested whether this reflects rational thought or sophisticated mimicry. Please note that in this section, we use 'thinking' in a purely functional sense (problem-solving and reasoning), postponing the question of conscious thinking to Sections 4–5

Critics firmly deny that today's LLMs possess a genuine understanding (Bender & Koller, 2000). They point out that models such as GPT-3 or GPT-4 lack embodied experience and do not construct world models that would confer meaning on words. Trained solely on next-word prediction in text, they master linguistic **form** of language, rather than the semantic **content**. Bender (2020) warns that distinguishing linguistic form from meaning remains central to genuine progress in natural language understanding. As one commentator put it, "*a system trained exclusively on language will never approach human intelligence, even if trained until the heat death of the universe,*" and the resulting understanding will remain shallow, devoid of the "full-blooded" thinking characteristic of humans. According to this view, LLMs should be regarded not as understanding agents but as compressed compendia of human knowledge, akin to vast libraries. Their linguistic proficiency does not stem from a genuine grasp of meanings, but from a statistical reflection of patterns present in training data. Cuskley et al. (2024) argue that the apparent similarity between human and LLM linguistic performance is superficial: the cognitive input of these models differs fundamentally from that of humans—humans learn language from limited but rich, multimodal context, whereas LLMs receive only massive amounts of text. In other words, LLMs lack grounding in the real world—they do not physically experience the scenes or concepts they describe [1].

A human being understands the world by experiencing the value of situations in relation to their own needs, emotions, and goals, while remaining in constant interaction with the surrounding environment. Meaning and understanding are therefore not intrinsic properties of representations but emerge from the dynamic interplay between the cognitive system and the world, constituting the essence of conscious thought.

A system that can adaptively respond to stimuli, anticipate the consequences of its actions, and modify its behavior through feedback exhibits a functional form of cognition, yet its "understanding" is not equivalent to human introspection. A machine that produces an output without introspective experience is not engaged in conscious thinking, even though it may successfully solve problems by executing assigned operations.

In this sense, conscious thought is not merely the processing of information but a form of experiencing meaning grounded in the organism's relation to the world. It is the affective and motivational orientation of the system that gives the world its significance, turning cognition into a process of meaning-making rather than passive representation. A machine may thus imitate the structure of cognition, but without the capacity to experience value, it remains outside the sphere of genuine understanding.

Nonetheless, cognitive scientists continue to study LLMs as potential cognitive models. The review by Niu et al. (2024) review evidence of both similarities and differences between LLMs and human cognition. Interestingly, some studies suggest that even text-only LLMs approximate human perceptual judgments. For example, GPT-3 produced similarity ratings for auditory, color, and taste stimuli that correlated strongly with human psychological data (ibid.).

Even so, current LLMs fall short in tasks requiring "commonsense" reasoning. Their successes in Theory-of-Mind experiments are partial, often explainable by simple heuristics or memorization of

[1] According to Harnad (1990), symbolic representations must be grounded in two nonsymbolic forms: (1) *iconic representations* mirroring sensory projections, and (2) *categorical representations* detecting invariant features of objects and events. LLMs lack both, precluding true grounding.

examples. In sum, the degree to which LLMs understand remains an open question: there is evidence suggesting certain elements of understanding (e.g., knowledge generalization, context-sensitive answering, self-calibration of responses) but equally strong evidence of the absence of deep semantic comprehension.

3. Can LLMs "feel" images and sounds?

Will today's generative AI systems—large language models (LLMs)— remain forever constrained by their architecture and perceptual limitations, much like the metaphorical carp in the San Francisco plaza pond described by Michio Kaku, beings with inherently limited access to the world? (1994).

Classical LLMs (e.g., GPT-3) operated exclusively on text, with no direct access to images or audio. For this reason, they were generally understood to lack perception in the biological sense: they neither see nor hear, but only process descriptions. The last two years, however, have brought a breakthrough in multimodal models that integrate language with additional modalities. OpenAI's GPT-4 and -5, for example, can accept images alongside text and generate textual responses (Achiam et al., 2023). GPT-4 can analyze photographs, charts, or screenshots and produce meaningful answers an instance of what might be called text-mediated visual analysis. In parallel, Microsoft Research's Kosmos-1, has been described as a Multimodal LLM that "can perceive different modalities, learn in context, and follow instructions" (Huang, 2023). Trained jointly on text and images, Kosmos-1 mastered both standard language tasks and perception-language tasks—such as image captioning, visual question answering (VQA), and solving simple visual puzzles. It achieved promising results in image description and could read screenshots (OCR) without a dedicated module—suggesting that it "understood" images via its learned linguistic representations.

How far, however, are these models from genuinely *feeling* stimuli? Here, the debate is polarized. Technically speaking, systems like GPT-4 or Kosmos-1 do process images and sounds (some newer models also process audio), but only by transforming them into inputs for a neural network. They lack conscious sensory experience. Critics argue that even multimodal models still reduce an image to a textualized representation (e.g., internal vectors aligned to words). Absent embodiment and biological senses, they cannot not *feel* as living organisms do. As the work cited above puts it, current multimodal models are, at bottom, mappings from images to text rather than full-blooded perception (Cuskley et al., 2024). Optimists counter that growing model complexity - more modalities, tighter coupling with robotics, larger parameter counts - may gradually narrow this gap. Already, we observe that a language model can display knowledge about the sensory world—for example, GPT can infer from text which sounds or flavors are similar, producing judgments correlated with human data (Niu et al., 2024).

Despite these advances, there is no scientific evidence that LLMs possess subjective experience (qualia) or self-awareness. The prevailing consensus remains that today's generative AI systems do not feel; their references to emotions or sensations reflect learned linguistic associations rather than genuine affect (Mitchell & Krakauer, 2024).

The question then becomes: what do generative language models lack in order to *experience* emotions as living organisms do? Addressing this requires first clarifying what animal emotions are, drawing on advances in neurobiology and affective science. Panksepp (1998) indicated that affective states arise from the intrinsic neurodynamics of self-centred emotional and motivational systems of the brain that current LLM programs lack. This suggests that embodiment (robotic or sensorimotor grounding) is a prerequisite for developments of authentic affective states.

3.1. **Biological inspirations**

One of the most promising developmental path runs through William James's (1884) classical perceptual theory, according to which the physiological changes occurring in the body in response to a stimulus are directly *felt*, and that these feelings themselves constitutes an emotion. As he put it: "*Our whole cubic capacity is sensibly alive; and each morsel of it contributes its pulsations of feeling, dim or sharp, pleasant, painful, or dubious.*" This implies that interoceptive sensations (heartbeat, muscle tension, gastric contractions, tissue oxygenation, etc.) compose the subjective experience of emotion.

By extension, analogues of such mechanisms might be pursued in robotic systems that couple LLMs with multimodal sensor suites, particularly with **interoceptors** monitoring the system's **allostatic** state (Sterling, 2012; Cangelosi & Schlesinger 2015).

These ideas were advanced by Antonio Damasio and Jaak Panksepp**.** Damasio's somatic marker hypothesis explaining how emotions guide decision-making and planned action (Damasio, 1991; Damasio et al., 1991). Somatic markers - physiological signals such as heart palpitations accompanying fear - serve as intuitive valuations of options and simplify action selection. Emotions bias choices toward positive feelings and away from negative ones, thereby enabling the brain to make rapid, adaptive decisions under complexity (Craig, 2014; Rolls, 2014).

Panksepp, in turn, identified evolutionarily primary-process emotional systems rooted deep in subcortical structures of the mammalian brain. Stimulation of these circuits elicits in animals' behaviors and reactions akin to human emotions, and neurochemical modulators (e.g., opioids), can attenuate separation-distress responses in puppies (Davis & Montag, 2019). His "Affective Neuroscience" (Panksepp, 1998) provided a neuroevolutionary framework for emotion, emphasizing that human feelings derive from primordial brain mechanisms shared with other mammals. Together, their work highlights both the evolutionary roots and the decision-making functions of affect.

A pressing question, then, is how, *specifically*, bodily signals give rise to emotional states. Feldman, Bliss-Moreau, and Lindquist (2024) review interoceptive pathways from peripheral receptors through the spinal cord and brainstem, via the thalamus, to cortical regions (notably the insula and orbitofrontal cortex)—showing how these signals are transduced, compressed, and integrated in the human brain to yield affect. They argue that valence (pleasure–displeasure) and arousal arise from characteristic patterns of interoceptive activation along these routes. On this view, emotions are inferences about one's bodily state, foundational for "meaning" and for motivation (Candia-Rivera et al., 2024).

Building on this tradition, Galus and Starzyk proposed their Motivated Emotional Mind **MEM** model of phenomenal consciousness (Galus & Starzyk, 2020; Galus, 2024; Galus, 2025). MEM posit that perceptual representations from sensory to cognitive levels become coupled with emotional representations in the limbic system. These couplings are learned through the co-occurrence of objects and events with interoceptive signals indicating allostatic states in specific organs and body regions. Innate, species-typical couplings between stimuli and affect have also been documented. In line with perceptual theories of emotion, such couplings generate the full range of feelings and affective states.

By contrast, current **LLMs** lack precisely these mechanisms. They have no bodies to generate interoceptive signals, no allostatic states to regulate, and no coupling between perception and emotion rooted in physiology. Their "emotions," when expressed in language, are statistical echoes of human discourse rather than lived affective states. This underscores why embodiment and interoceptive grounding are likely prerequisites for any AI system aspiring to genuine feelings or phenomenal consciousness.

3.2. Robotic systems

If we are to follow nature's lead, we should not expect emotions and feelings to emerge in disembodied LLMs hosted in the computational cloud. Instead, attention should turn to robotic systems, in which an LLM is **embodied** through coupling with a robot chassis that integrates a rich **sensor suite** for perception and **actuators** for motor responses. This embodiment allows for the grounding of language in perception and action.

A first step in this direction was Google's **RT-1** and **RT-2** (Brohan et al., 2022, Zitkovich et al., 2023), which translated web-scale vision–language knowledge into robot actions, thereby linking semantics with physical control. A more ambitious development, **PaLM-E**, integrates a very large model (562B parameters) into a robotic platform that ingests multimodal input streams (text prompts, visual observations, proprioceptive data) and generates both low- and high-level control commands (Driess, 2023). In effect, *PaLM-E* binds "words to perception": its input is a multimodal sentence comprising a textual prompt, visual observations of the environment, and robot state readings. Trained end-to-end**,** the model executes complex tasks such as object-manipulation planning, navigation, and

standard vision–language benchmarks (VQA, image captioning). Crucially, it showed how perception, language, and action can be bound within a single model.

Recent work extends these principles. The **Gemini Robotics VLA line** (Team et al., 2025), integrates spoken commands, vision, and robotic action in a unified architecture, though still at prototype stage. Vision–Language–Action **(VLA)** models combine instruction following, image/video perception, and robot control within a unified pipeline,. The work of Sapokta and colleagues (Sapokta et al., 2025). demonstrate successive improvements in cross-task generalization and the transfer of Internet-derived knowledge into physical manipulation. While this constitutes strong **functional grounding**, it is only functional integration of perception and action and not an evidence of conscious human-like perception. A similar line of work is an open VLA trained on nearly one million robotic episodes, deployable across diverse robot platforms and rapidly fine-tuned to new tasks. Collectively, these efforts provide strong evidence that learned vision–language knowledge can in fact drive real-world motion (Kim et al., 2025).

Yet, despite this progress, current systems remain restricted to **perception–action loops**. They lack *interoceptive sensing* or *motivational states*—the bodily signals and affective mechanisms emphasized by biological theories of emotion (James, 1884; Damasio, 1991; Panksepp, 1998). From the perspective of the **Motivated Emotional Mind (MEM)** framework, embodiment through sensors and actuators is a necessary but not sufficient step. The most significant deficiencies include a lack of emotions, a lack of long-term motivation, and a lack of embodied memory. For machines to approximate emotions and feelings, they must also be equipped with architectures that integrate internal state monitoring, valuation, and affective coupling with perception and action.

While appreciating the importance of progress, it is also important to emphasize the technical limitations that hinder the development and implementation of embodied intelligent robots, such as sensitivity to distributional data shifts (covariate shift, prior probability shift, label shift, concept drift), lack of generalization in new environments, and enormous computational costs.

3.3. **Emotions and feelings**

The robotic models described above possess the capacity to perceive their surroundings, which has led many researchers to claim that they instantiate elements of subjective experience—*qualia*. One may further expect that they include sensors reporting on the internal subsystem states, which are used to optimize responses and maintain homeostasis. Does it follow, then, in line with the perceptual theory of emotion, that these **rudimentary interoceptors** enable robots to *feel* complex affective states?

Survey articles by Pezzulo and colleagues argue that the behavior of living organisms rests on generative models tightly coupled to the body and the world, regulating the sensory consequences of action to maintain homeostasis. For Pezzulo, the basis of meaning emerges from the regulation of internal bodily states. **Passive** models (such as typical LLMs) lack this core (Pezzulo et al., 2024). An increasing number of authors distinguish effective recognition and inference from **human perception** that is **embodied** and **goal-directed**. This is not a claim of "absolute impossibility," but of a missing module: body-coupled regulation and valuation of sensations.

Are there studies, especially those integrating language models with robotics and multimodal sensors, that demonstrate a blurring of the boundary between human and machine (robotic) perception, or, conversely, that justify the claim that such equivalence is impossible? Do they incorporate a role for emotion? Do they indicate that interoceptive reporting of a system's homeostasis is required for the feeling of emotion?

In robotics, interoception-like solutions are beginning to appear: robots monitor internal variables (e.g., motor temperatures, energy levels) and learn homeostatic behaviors, thereby integrating perception and action in real hardware. This is behavior oriented toward homeostasis, not a claim that robots "feel" emotions (Yoshida, Kanazawa, & Kuniiyoshi, 2024). There are also proposals for interoceptive robots in architecture and motion planning; the authors explicitly clarify that the aim is not genuine emotions, but internal representations of state that improve adaptation and cooperation (Zhou, Menassa, & Kamat, 2025). In simulations of behavioral responses, homeostatic and even "prosocial" behaviors can emerge when an agent's internal states are coupled to environmental perception. This is an important conceptual result, yet it remains a **simulation**, not evidence of **actual experience** (Yoshida & Man, 2025).

Complementary work explores *linguistic cognition of emotion* without embodiment. For example, Li and colleagues show that LLMs can reason about emotions purely from linguistic knowledge, without any sensorimotor apparatus, reflecting cognitive representations of emotion, rather than experiential feeling (Li et al., 2024).

In sum**:** contemporary science suggests that interoceptive–homeostatic components (or their functional equivalents) are required for the feeling of emotion. Current systems provide, at most, supplementary interoception (monitoring energy, temperature, loads) that improves control and behavioral prioritization, but does not constitute evidence of phenomenal "feeling."

4. Theoretical model

Seeking to resolve these conflicting conclusions and to bring the discussion into clearer focus, we require a theoretical model that would establish criteria for classifying individual systems with respect to their capacities for autonomous action and thinking**,** feeling, and guidance by their own hierarchy of emotional valuation (i.e., their own good). This constellation of properties accords with Brentano's understanding of intentionality, though it is not equivalent to his definition. It also underpins our understanding of what counts as "human" or natural traits of thinking and feeling.

A useful starting point is David Chalmers's recent essay in *Proceedings and Addresses of the American Philosophical Association* (97:22–45, Chalmers, 2024), where he revisits the problem of symbol grounding and asks whether the capacity to think requires a capacity to feel. He advances the hypothesis that computers and "good old-fashioned AI" can perform operations of thinking, reasoning, pattern recognition, and planning, which include deductive and inductive inference. He does not explicitly extend this to abductive reasoning. He also suggests that multimodal large language models (mLLMs) are capable of a certain form of perception, raising the question of whether multimodality enhances cognitively expressible (propositional) processes. While he suspects that it does he questions the equivalence of propositional information and sensory perception, leaving unsettled the issue of whether **multimodal LLMs** are capable of the same kind of thinking as humans.

Verification of Chalmers's theses within the best-known contemporary models of consciousness proves unavailing. Models such as Global Workspace Theory **(GWT)** (Baars, 1988; Dehaene & Changeux 2011), Predictive Processing Theory **(PPT)** (Friston, 2010; Clark, 2013), and first-order and higher-order representationalism **(HOT/HOP)** (Dretske, 1995; Lycan, 1996; Tye, 2000; Rosenthal, 2005) provide useful explanatory tools but do not offer clear criteria for distinguishing cognitive capacities from affective ones in artificial agents. Likewise, influential approaches such as Integrated Information Theory **(IIT)** (Tononi 2008, Tononi & Koch 2015)) fail to deliver practical classification schemes, and classical functionalism did not gain popularity in the context of embodied multimodal AI.

We set aside metaphysical or scientifically ungrounded accounts such as non-reductive physicalism/property dualism, panpsychism/Russellian monism, eliminativism/illusionism, and quantum theories like Orch-OR.

As the basis for classifying the rapidly proliferating AI models, we adopt the **Motivated Emotional Mind (MEM)**. MEM elaborates upon Recurrent Processing Theory **(RPT)** (Lamme, 2006) by embedding recurrent perceptual processing within a motivational and affective architecture. It emphasizes the integration of perception, cognition, and emotion under a hierarchy of values that guide autonomous action. The MEM framework thus provides both a reductive and operational model of the conscious mind, as well as the most useful criteria suitable for assessing whether AI systems merely simulate cognition or instantiate elements of autonomous, affect-laden agency.

In principle, other embodied predictive or active-inference architectures might realize some aspects of phenomenal consciousness without implementing the full semblion-based hierarchy of MEM. In this article, we argue that MEM provides the most explicit and operationally detailed realization of such an embodied, affect-laden architecture, but we do not claim the logical impossibility of alternative realizations.

4.1. Characterization of the MEM model: core assumptions and theses

The Motivated Emotional Mind (MEM) is a reductive, physicalist model of consciousness which holds that consciousness, especially subjective feelings (qualia), can be explained as the outcome of specific neuronal processes in the brain intimately coupled to the body and the senses. Every conscious sensation is rooted in the activity of sensory receptors or in internal bodily signals. MEM thus emphasizes embodiment: first-person experience arises from the perception of signals arriving from the external world (exteroceptors) or from within the organism (interoceptors and proprioceptors).

MEM extends Lamme's Recurrent Processing Theory (RPT) by integrating motivational and emotional components, an explicit architecture of neuronal representations, and processes by which those representations are formed. In MEM, perception is always coupled with emotion: when a stimulus is experienced, the brain simultaneously activates phenomenal and affective states associated with it (e.g., qualia, pleasure, fear), which are learned from prior experience. Accordingly, each perceptual representation carries an emotional "value"—its significance for the organism (beneficial, threatening, neutral, etc.). The mind is therefore "motivated" (steered by ongoing needs and feelings) and "emotional" (emotions are integral to information processing).

At the core of the theory lies a heterarchical organization of neuronal representations of perceptual and mental states, whose basic unit is the *semblion*: a hypothesized multilayer neuronal structure binding perceptual, mnemonic, and affective representation. Thereby, the hierarchy of semblions reflects the hierarchical structure of world knowledge: subsemblions encode constituent features (e.g., shape, color, features, and attributes of an object), which combine into higher-order semblions representing objects or situations, wholes (e.g., the object itself). At the highest level of generalization and association, this organization creates a model of the world/environment. Associative memory organized in this way enables the mind to recognize the components of the environment and their meaning through a network that connects perception with past feelings and emotions.

Inactive semblions exist physically as synaptic traces; when activated, they instantiate mental content. Thus, the same structure can be described as both a **brain state** and a **mind state**, depending on activation—an important element of MEM's reductive stance.

MEM assumes the brain operates as a multilayer, hierarchical neural network that processes information from receptors up through cognitive and motor levels. This structure resembles deep learning architectures—modern neural networks with dozens of layers (Felleman & Van Essen, 1991). The hierarchy is semi-structured: lower levels topographically map sensory data (e.g., the retinotopic layout across the retina/LGN/V1), whereas higher levels form increasingly abstract representations by integrating features and context.

Signal flow in MEM proceeds not only via vertical connections supporting the Feed-Forward Sweep **FFS** (in the sense of RPT), but also through lateral and heterarchical connections that enable associations among parallel representations in memory. MEM is thus an associative, heterarchical memory, capable both of deep feature representation and of lateral binding of those representations into coherent images and concepts, up to a model of the environment. The resulting representations, *semblions*, are mnemonic traces of perception, engrams that bind multiple aspects of experience.

Semblion representations are both material and mental. When a *semblion* is inactive, it exists potentially, as a structure of synaptic connections (a physical record in the brain); when activated, it instantiates mental content. This property is essential to the model's reductionism: the very same networks of interconnected neurons serve as mind states if most of its neurons are activated or just as brain substructures if they are not active. (Galus, 2024 b)

The concept of the **semblion**, newly defined within the MEM model (Galus & Starzyk 2020, pp 12-16), has a broad and highly specific meaning relative to the classical notion of a neuronal representation—an engram understood as the stored trace of a percept within a neural network. It also extends well beyond Vadakkan's earlier formulation, which first applied this term to describe fragmentary, hierarchical, encoding neural structures. In the MEM model, semblions are characterized by multilevel lateral, ephaptic, and multimodal associations. MEM attributes to them the capacity to compete for access to higher layers during the feedforward sweep **(FFS)** and to drive re-entrant transmission of stimulation in the recurrent process **(RP)**, both processes grounded in identifiable biophysical mechanisms. Collectively, these features yield a credible—though still insufficiently

corroborated—account of various processes underlying thought, cognition, and phenomenal consciousness (Galus 2023a; 2023b; 2025a; 2025b).

In generating consciousness, feedback (re-entry) plays a pivotal role. First, there is an external feedback loop between organism and environment: our actions affect the environment, and the consequences are perceived by the senses, allowing behavior to be corrected (long-range feedback). More crucial for subjective feeling is the internal feedback loop—top-down projections within the brain. Higher (cognitive) levels send backward signals to lower (sensory) levels—secondary perception. Through these recurrent connections, stored patterns can re-excite early sensory areas almost as if the stimulus were present again. Subjectively, this underlies visual imagery, auditory recall, visualization, dreaming, and introspection.

This feedback mechanism explains how conscious thinking and remembering are possible. When we consider a concept or a memory, the corresponding *semblion* becomes active and, via top-down re-entry, generates "internal impressions"—e.g., a mental image, an inner voice, or the feeling of a recalled emotion—contents accessible to consciousness. Galus and Starzyk describe this as "seeing and feeling one's own mental states". A necessary condition is re-entry into early sensory areas: for example, V1 can be driven both by the eyes and by memory (projections from the hippocampus/association cortex), yielding phenomenologically similar visual impressions.

In this way, MEM unifies external and internal perception along a single continuum. At lower levels, there is a continuous stream of world-driven perception (when stimuli arrive via the senses), yet higher areas can concurrently initiate "pseudo-stimuli" that impinge on the same lower areas from above. During wakefulness, the influx of external input typically masks spontaneous internal imagery; in quiescent states (sleep, reverie), however, *semblions* continually generate sequences of mental images and impressions for their own purposes (e.g., hippocampal projections into V1 [Kosslyn, 2005; Slotnick, 2017]). MEM thus accounts for both conscious dreaming and the familiar "stream of thought" (Singer, 1978) as natural consequences of a recurrent architecture—the brain is constantly simulating perceptually.

In MEM, emotions and affective states are tightly linked to bodily signals. When homeostasis is perturbed—for instance, by hunger, pain, threat, or, conversely, by need satisfaction—interoceptors and proprioceptors convey this information to the brain, where it is processed analogously to external stimulus. We then attach emotion labels to the resulting feelings: a drop in blood glucose is felt as unpleasant tension (which we call anxiety or hunger); elevated heart rate and muscle tension are felt as fear, and so on. In MEM, an emotion is conceptualized as "semblion of feeling"—a neuronal representation of a recognized bodily state. By identifying patterns in interoceptive signals, the brain forms emotional *semblions* (analogous to perceptual *semblions)* triggered by interoceptive inputs. Put differently: physiological feeling + its neuronal representation = emotion. This view aligns with the perceptual theory of emotion à la James–Lange (and its modernization by Damasio's somatic-marker hypothesis): bodily reaction precedes conscious emotion. Galus and Starzyk state explicitly that emotional experience is impossible without embodiment and bodily receptors. Thus, affective states just are bodily sensations, requiring no additional "mystical" quality. Their primary role is motivation and the valuation of action scenarios.

MEM proposes that an organism can plan and act intentionally because it can visualize and feel the consequences of potential decisions before executing them. Via feedback-based imagery, brain activates semblions of possible actions and virtually feels the anticipated emotions or outcomes (secondary perception). This internal simulation functions as a **trial run**, allowing the system to compare alternatives without direct environmental risk.

The decision that ultimately prevails is the one that best promotes allostatic balance —the long-term maintenance of well-being in a changing environment (Sterling, 2012). In this manner, MEM not only explains intentional action but also grounds **conscious will** and **teleology** in neurobiological mechanisms. The authors explicitly invoke Brentano's concept of intentionality (Mulligan, 2004): mental states are always *about* something. In MEM, intentionality arises because the organism harbors internal states (needs, drives, feelings) that orient it toward goals and regulate its behavior accordingly. Further details are presented in the monograph *Reductive Model of the Conscious Mind* and related work (Galus & Starzyk 2020; Galus 2022; 2023; 2024; 2025).

In sum, MEM offers a coherent picture in which phenomenal consciousness is perception of sensory states (external or internal) by a richly structured memory network; feelings are perceptions of bodily states and their representations (emotions); and decision-making emerges from a mechanism that selects, among conscious representations, those most advantageous to the organism, on the basis of the felt emotional valuation of those representations.

A comparison of perception, affective states, and decision-making in LLM, robotic, and biological systems operating according to the MEM model is presented in the Comparative Table at the end of the article (see Appendix A).

5. Application of the MEM Model to the Question of Thinking and Consciousness

The MEM model provides a framework for addressing questions posed by Chalmers about the relationship between thinking and perception. Specifically, it illuminates whether a system devoid of sensory grounding—such as a text-only LLM—can genuinely think or be conscious.

5.1. **Chalmers's Questions and the MEM Perspective**

Chalmers poses two key questions (Chalmers, 2024):

1. the **causal question** – whether sensory experience is necessary for thinking to arise; and
2. the **constitutive question** – whether perceptual elements are intrinsic components of thought.

He distinguishes three possibilities regarding the grounding of thought:

A. Thinking can exist without prior sensory experience, (**causal grounding)**;
B. It may necessarily include sensory components, (**constitutive grounding)**;
C. It may require neither, leading to the thesis that there is such a thing as "purely abstract" thinking, detached from sensory experience. Systems of this sort he calls *pure thinkers*.

Consistently, he infers that if thinking without senses is possible, then it is not constitutively grounded in sensibility. Even if, in humans, the senses are practically indispensible for the development of thought (causal grounding), in a logical–metaphysical sense, one can conceive of beings that think without the sensory participation. Hence he conjectures that at least text-only LLMs, which lack senses, might qualify as *pure thinkers*.

Chalmers wrights: "*If we devised a 'pure' AI system with no input/output connections to the world, its lack of connections to the world would not by itself prevent it from being able to think and understand a good deal—from mathematics to philosophy to speculative scientific hypotheses about reality.*" This claim that a hypothetical "pure thinker" AI system could think and understand "from mathematics to philosophy" without connections to the world should be regarded as unproven. Even if a machine (called here a pure thinker) were able to perform logical operations, it would possess no awareness of doing so, just as a computer solving a differential equation has no understanding of the process it executes.

One may, of course, with considerable terminological flexibility, speak of "thinking" in LLMs in virtue of their impressive computational abilities. If thinking is understood in this loose, purely computational sense—as the manipulation of internal symbolic or sub-symbolic states according to learned rules—then it is not unreasonable to describe current LLMs as engaging in a kind of computational thinking. To avoid confusion, we will refer to this as pseudo-thinking: a coherent, often useful form of inference and problem solving that remains entirely at the level of symbol manipulation within the model's architecture, without any implication of phenomenal awareness.

Moreover, LLMs do not "learn" through their own experiences appraised by their own goals. Their training data are only "indirect," derived from human perception that has been preprocessed and delivered in ready-made form. Consequently, the resulting representations are, by any strict standard, neither constitutively nor causally grounded in perception. Since LLMs lack such grounding in actual perception, interoception, and homeostatic regulation, there can be no question of feelings, phenomenal experiences, or consciousness in their current implementations, even if their pseudo-thinking sometimes superficially resembles human reasoning.

By contrast, conscious thinking, in the human or animal sense, presupposes more than pseudo-thinking. On the MEM account, it requires that neuronal (and corresponding mental) representations be grounded in recurrent sensory and interoceptive processes, such that thoughts are accompanied by secondary perception and affective evaluation of their content. Conscious thinking thus involves a first-personally experienced "grasp" of meaning, rooted in re-entrant activation of sensory maps and in emotionally valenced bodily states.

Grounding in perception, both constitutive and causal, is a prerequisite for consciousness. Causal grounding denotes direct sensory acquaintance with the world, which enables the construction of a model of reality. Constitutive grounding denotes satisfaction of the identity condition between the mental and the material domains (Galus, 2023b). Both are conditions of consciousness. To reiterate: conscious thinking, in the human or animal sense, requires grounding neuronal and mental representations (symbol grounding) in the sensory experience. Since LLMs lack such grounding, there can be no question of feelings, phenomenal experiences, or consciousness.

5.2. Thinking in the MEM Model

Let us examine whether an intelligent system operating under the MEM model satisfies the criteria formulated by Chalmers and is thereby capable of conscious thinking, both in the case of biological organisms and potential artificial systems.

Constitutive grounding.

In MEM, thinking is constitutively grounded in perception, and in particular in perceptual structures called *semblions*. These structures link cognitive areas, where abstract symbolic representations are encoded, with sensory layers responsible for exteroception and interoception. Their compact hierarchical architecture enables multidirectional stimulation through the processes of feedforward sweep (FFS) and recurrent processing (RP). The FFS process serves to generate representations, whereas RP enables the visualization of the contents of thought within lower sensory layers. This secondary perception - the reentry of thought into sensory areas - produces awareness of the content of thought. In this sense, thinking can be treated as the reactivation of perception and interoception.

Consider an example of the perceptual grounding of abstraction: when you think of "freedom," your mind does not operate on a "pure, abstract concept," but activates semblions and subsemblions associated with a sense of space, movement, absence of pressure, and often an interoceptive state (relief, breathing, bodily ease). These are not decorative "add-ons" to thinking, memory, and imagination; they are its substrate—the very essence of thought.

Causal grounding.

As presented in Section 4, all knowledge stored in memory originates in perceptual processes, and all felt emotions derived from interoceptive activity (Galus, 2023b). In addition to the bottom–up integration mentioned above, the MEM model exhibits functional integration. The network of semblions simultaneously constitutes a powerful associative memory - where patterns of objects and phenomena are consolidated - and as a processor that transforms perceived information through the FFS process. Within neuronal structures, these processes form representations of percepts (engrams/semblions)—fragments of sensorimotor experience generalized for recognition and action.

This process exploits bottom–up associations between successive layers. The biological/biophysical mechanisms of these associations were described in the book cited above (Galus & Starzyk 2020). Among the mechanisms relevant here is the original concept of Neuro-Electro-Dynamics (NED) by Aur and Jog (Aur & Jog 2010; Aur 2025a, 2025b). Alongside vertical associations, lateral and intermodal associations play an equally fundamental role, including synaptogenesis and neurogenesis, synaptic reinforcement, ephaptic coupling of dendritic spikes, dendritic spine aggregation and structural modulation, tripartite synapses, and epigenetic modifications (Galus 2025b).

Within MEM, thinking is not an operation on abstract symbols divorced from perception, but an operation on reactivated, transformed, internally simulated perceptual schemata—what, in the theory of

embodied linguistics, are called "image schemas". Abstract concepts and ideas emerge from the recursive integration of semblions, often generalized and assembled into new configurations; for example, metaphors, cross-modal analogies, and higher-order image schemas (Galus 2018).

In summary:

- Perception/interoception gives rise to *semblions*—grounded, quasi-sensory, motivationally evaluated schemata that constitute representations of percepts.
- Thinking consists of operating on reactivated semblions, with reentrant (top-down) flow into sensory layers that "refreshes" and corrects these representations. The absence of recurrent processing **(RP)** would impair the ability to recognize objects and events whose sensory image is distorted, incomplete, or ambiguous. The recursive process supports error correction in recognition and resistance to illusions and perceptual deceits.
- Secondary perception of reentrant activity results in the awareness of thought.
- Interoception of unmet needs generates motivations. Interoceptive signals arising from motivation and disrupted homeostasis assign weight (salience), select, and stabilize representations, and are felt as affective states accompanying the contents of thought.
- This renders grounding both constitutive (thought contains perceptual components) and causal (without development on sensory and motivational data, the proper repertoire of semblions will not form). Together, they constitute a necessary condition for awareness of thought's content.

The MEM model encompasses all forms of thinking while accounting for phenomenal feelings, emotions, empathy, moral attitudes, prosocial dispositions, imagery, daydreams, memories (autobiographical self-awareness), and aesthetic impressions. How can such rich manifestations of mind arise in human and animal organisms?

Associations between perceptions and emotional states yield sensitivity to characteristic, recurrent features and configurations of the matter interacting with the system. This may cultivate responsiveness to harmony and the beauty of nature**,** which can in turn, be projected onto features of social relations and organizations**.** The animal world offers, numerous examples, both positive and negative of such affective projection. Consequently, sensitivity to harmony in the natural world may extend to sensitivity to harmony in social relations**,** giving rise to empathy, reciprocal altruism, higher forms of morality, and an internal imperative to abide by ethical and social norms.

6. Empirical Tests for Machines that Think and Experience Emotions and Sensations

As established in the preceding sections, widely used LLMs cannot think or feel in any meaningful sense. Robotic systems studied to date lack of such capacities (see the comparative table). What, then, would be required for a system to acquire the ability to think and feel consciously?

To achieve this, one must follow the principles and design implications of the MEM model. This entails embodying the system; equipping it with a broad array of interoceptors and motor control neurons to perform actions and to report all states of homeostasis and allostasis; designing bottom-up transmission of stimulation through the feedforward sweep (FFS) and building vertical associative networks; enabling maximally flexible and dense lateral and multimodal associations; providing feedback from cognitive to sensory layers, so that higher-level activity will undergo secondary perception; organizing **competition** among parallel streams of stimulation elicited by inputs from different sources; and, more generally, implementing the principles of information processing described in MEM.

These include:

- providing a stimulating environment for the development of perceptions and action control;
- the organization of associative memory in a mixed heterarchical–hierarchical form;
- a hedonic center (pleasure/pain) coupled to mechanisms for attention switching, curiosity, and novelty detection, etc.

In short, the task is to create an architecture and environment of semblions operating in accordance with the rules of the MEM model.

A critical advantage of the MEM model is that it yields **testable and falsifiable predictions** about the internal dynamics required for conscious perception, imagery, affect, and decision-making. Contemporary neurotechnology and robotic sensing allow these predictions to be operationalized in controlled experiments. In this section, we outline two complementary classes of tests—(1) perceptual–imagery reactivation tests and (2) interoceptive–affective simulation tests—each capable of empirically distinguishing systems that merely *compute* from those capable, in principle, of *experiencing*.

6.1. Perceptual Reactivation Test: Does Imagery Re-Activate Sensory Maps?

According to MEM, conscious imagery, recollection, and "internal simulation" all require **top-down re-entry** into early sensory cortices (Lamme, 2006; Kosslyn, 2005; Slotnick, 2017). If a system *imagines* a scene without reactivating the sensory layers from which the original percept was formed, MEM predicts that no genuine quasi-perceptual experience will occur: such a system performs symbolic reasoning only.

This leads to a straightforward experimental test.

Test rationale:

1. **Text-only LLMs (pure thinkers)** operate exclusively on linguistic-symbolic embeddings.
    - "Imagining" a scene consists of generating additional symbolic sequences.
    - They lack low-level sensory maps; thus **no measurable reactivation** of early sensory layers is possible.
    - LLM therefore predicts *zero* quasi-perceptual signatures.
2. **Multimodal LLMs / VLA robots** possess vision encoders trained on images.
    - During training they may exhibit classifier-level illusions, but
    - During later *imagery* tasks, generation remains confined to associative or language layers.
    - Early visual maps should remain **silent** or **weakly/randomly activated**.
3. **MEM-based humanoids**, in contrast, require strong **secondary perception**:
    - During imagery and illusion tasks, they should **spontaneously reactivate low-level sensory patterns** that closely match the original percept.
    - Even after factual correction (e.g., "this is an illusion"), the *phenomenal impression* persists because secondary perception continues to stimulate early maps—exactly as in humans experiencing Müller–Lyer illusions (consistent with Lamme & Roelfsema, 2000; Galus 2023a; 2025a).

Predictions:

A MEM-consistent system must satisfy:

- **Pattern similarity**: The reactivated sensory maps during imagery should correlate highly with maps from original perception.
- **Persistence of phenomenal error**: After cognitive correction, the system should *still* produce low-level sensory activations corresponding to the illusion, while only the **decision output** reflects the correction.
- **Bidirectional coupling** between cognitive semblions and sensory layers.

These requirements follow directly from MEM's claim that imagery is perception-like because it is genuinely perceptual at the level of neural implementation (secondary perception).

6.2. Affective Simulation Test: Can the System "Feel" Its Own Internal Counterfactuals?

MEM's second strong prediction concerns **emotion and motivated decision-making**, which emerge from reentrant activation of **interoceptive semblions** tied to real bodily states (James, 1884; Panksepp, 1998; Damasio, 1991; Galus & Starzyk 2020).

Thus, a system capable of genuine affect must, when *simulating* a scenario, partially reproduce the **interoceptive activation patterns** associated with the real bodily events.

Experimental paradigm:

1. Subject the robot to real events affecting its internal bodily state:
    - mechanical impact,
    - overheating or load stress,
    - battery depletion,
    - internal strain.

    Record its interoceptive maps in real time.

2. Ask the robot to imagine two contrasting action plans:
    - a "rapid" but risky plan that would normally induce damage or pain,
    - a "cautious" plan that avoids negative consequences.
3. Measure whether **the same interoceptive maps** activated during real aversive states reappear during *imagined* risky scenarios.

    Predictions:

- **LLMs and multimodal LLM robots:**
    - No physiologically structured reactivation should occur.
    - Their costing of scenarios remains symbolic, not somatic.
    - No quantitative relationship between "imagined pain" and choice behavior.
- **MEM humanoid:**
    - Imagining a harmful scenario should **reproduce interoceptive semblion patterns** nearly identical to those in real painful events.
    - These patterns must *not* appear during cautious planning.
    - A **graded relationship** should exist between the strength of reactivated "pain" semblions and the likelihood of selecting a safer action.
    - This produces what Damasio called a *somatic-marker-like* valuation process, but implemented within MEM's semblion architecture (Damasio, 1996; Rolls, 2014; Galus 2025b).

Thus MEM uniquely predicts that a conscious robot will "feel" counterfactuals—through quantifiable re-entry into interoceptive layers.

6.3. Why These Tests Matter

These tests provide **operational criteria** for the emergence of:

- **quasi-perceptual imagery** (via reactivation of sensory maps),
- **felt motivation and valuation** (via reactivation of interoceptive maps),
- **phenomenal consciousness** understood as reentrant perception of internal states.

Crucially, the MEM model predicts **sharp and measurable distinctions** from LLM:

System	Sensory reactivation	Interoceptive reactivation	Phenomenal consequences
LLM (pure thinker)	none	none	none
Multimodal robot / VLA	weak, random, classifier-level	none	none
MEM humanoid	strong, structured, illusion-persistent	structured, valenced	quasi-perceptual imagery + felt affect

Because these predictions are falsifiable, they provide a scientific path for determining whether an artificial system is capable of **thinking with awareness** and **feeling with affect**, as defined by MEM.

Outline of a minimal architecture sufficient to create a thinking and feeling humanoid:

- Embodiment: an integrated housing combining the humanoid's "brain" with its sensory receptors and motor effectors.
- Effectors: actuators enabling autonomous reactions driven by neural control signals.
- Sensors: exteroceptive (vision, audition, touch), proprioceptive, and interoceptive.
- World model: generative, with re-entrant loops projecting back to sensory layers (for secondary perception).
- Memory: a network of semblions (episodic/semantic), indexed by affective contexts.
- Valuation: homeostatic regulation, risk assessment, and cost-benefit calculus.
- Planning: model-based reinforcement or motivated learning with counterfactual prediction.
- Action controller: selection and monitoring of outcomes (with an error logging and correction).
- Language (optional): a referential layer bound to semblions and actions, rather thannot merely text↔text associations.

6.4. Minimal computational formalization of MEM and semblions

The MEM framework was originally articulated in neurobiological and conceptual terms. Here we sketch a minimal computational formalization that makes the core commitments explicit and testable, while remaining agnostic about many low-level biophysical details. This formalization is intended as a scaffold for future implementations rather than a complete neural model.

6.4.1. System state and architecture

We model a cognitive system (biological or artificial) as a dynamical architecture with the following components:

- **Exteroceptive state**
  $\mathbf{x}(t) \in \mathbb{R}^{n_x}$: at each discrete time step t, the system receives a vector of sensory signals from the external environment (e.g., visual, auditory, tactile streams).
- **Proprioceptive state**
  $\mathbf{p}(t) \in \mathbb{R}^{n_p}$: internal signals about posture, joint angles, and movement of effectors.
- **Interoceptive state**
  $\mathbf{b}(t) \in \mathbb{R}^{n_b}$: bodily signals related to homeostasis and integrity (e.g., energy resources, temperature, mechanical strain, internal damage indicators). In the biological case these

correspond to classical interoceptive modalities; in an artificial agent they would be realized by sensors monitoring internal physical variables.

- **Neural (or neural-like) activations**
  The system contains a multilayer network of units $N = \bigcup_{\ell=0}^{L} N_\ell$, partitioned into:
  - early sensory maps N_0, N_1receiving $\mathbf{x}(t)$and $\mathbf{p}(t)$,
  - associative and hippocampal layers $N_2, \dots, N_{L-2}$,
  - affective/interoceptive layers N_{aff}receiving $\mathbf{b}(t)$,
  - motor/output layers N_{mot}projecting to effectors.

  Each unit $i \in N$has an activation $h_i(t)$updated by:

$$h_i(t+1) = f\left(\sum_j w_{ij}^{ff}\, h_j(t) + \sum_k w_{ik}^{lat}\, h_k(t) + \sum_l w_{il}^{fb}\, h_l(t) + u_i(\mathbf{x}(t), \mathbf{p}(t), \mathbf{b}(t)) + \theta_i\right),$$

  where w^{ff}, w^{lat}, w^{fb}denote feedforward, lateral, and feedback weights; $u_i$are sensory/interoceptive inputs; and $f$is a non-linear activation function.

- **Needs and value variables**
  The system maintains a vector of needs $\mathbf{n}(t)$(e.g., energy, integrity, social contact for animals; energy, structural health for robots) with target setpoints $\mathbf{n}^*$. Deviations $\Delta\mathbf{n}(t) = \mathbf{n}(t) - \mathbf{n}^*$generate global modulatory signals (discomfort/relief, negative/positive valence).
- **Action policy**
  Actions $\mathbf{a}(t)$are chosen from a set $\mathcal{A}$ by a policy $\pi(\mathbf{h}(t), \mathbf{b}(t), \mathbf{n}(t))$, which may be implemented via model-free or model-based reinforcement learning, but must be modulated by the affective value system.

  This architecture can be depicted as a block diagram with (Galus 2023b):

1. **Body and sensors** feeding exteroceptive, proprioceptive, and interoceptive channels into
2. **Primary sensory/interoceptive maps**, which project to
3. **Associative and hippocampal layers** forming a heterarchical memory, tightly coupled to
4. **Affective/valuation circuitry** that encodes needs and allostatic variables, and
5. **Motor and planning modules** that send efferent commands to the body and receive re-entrant feedback from the environment.

Recurrent connections from associative/hippocampal and affective layers back to early sensory and interoceptive maps implement **secondary perception**—the core of conscious imagery and feeling in MEM.

6.4.2. Semblions as distributed representational units

Within this architecture, a **semblion** is a recurrently co-activated subnetwork linking sensory, mnemonic, and affective elements. Formally, for each semblion $S_m$we define:

$$S_m = (V_m, E_m, \boldsymbol{\mu}_m, \boldsymbol{\rho}_m, v_m),$$

where:

- $V_m \subset N$: a set of units spanning multiple layers (sensory, associative, affective) that co-activate during particular experiences;
- E_m: the subset of connections among units in V_m;
- $\boldsymbol{\mu}_m$: a prototypical pattern of activation over exteroceptive/proprioceptive inputs and sensory units associated with this semblion (its **perceptual code**);

- $\boldsymbol{\rho}_m$: a prototypical pattern of interoceptive and affective activation (its **feeling code**);
- v_m: a scalar or low-dimensional vector summarizing long-term value (e.g., expected impact on needs, valence).

Given the current network state $\mathbf{h}(t)$and bodily state $\mathbf{b}(t)$, the **activation level** of semblion $S_m$can be approximated as:

$$a_m(t) = \sigma\big(\alpha\, \mathrm{sim}(\mathbf{h}_{V_m}(t), \boldsymbol{\mu}_m) + \beta\, \mathrm{sim}(\mathbf{b}(t), \boldsymbol{\rho}_m) + \gamma\, v_m\big),$$

where:

- $\mathbf{h}_{V_m}(t)$is the restriction of $\mathbf{h}(t)$to units in V_m,
- sim(·,·) is a similarity function (e.g., dot product, cosine similarity),
- α, β, γ are scaling parameters,
- $\sigma$is a squashing nonlinearity.

A semblion is **inactive** when $a_m$is below threshold, existing only as a physical connectivity pattern E_m; it becomes a **mental state** when $a_m$crosses threshold and its recurrent projections drive secondary perception in sensory and interoceptive maps. In this way, the same physical structure has two descriptions: as a brain subnetwork and as a momentary content of experience.

6.4.3. Affective coupling

Affective coupling in MEM connects interoceptive deviations to both learning and ongoing semblion activation. We introduce a global **affect signal**:

$$A(t) = g(\Delta\mathbf{n}(t)) \in \mathbb{R},$$

where g maps deviations from homeostatic setpoints to a signed scalar (negative for threat or damage, **positive for restoration or satisfaction).**

- $n(t)$ – vector of needs/allostatic variables (energy, integrity, temperature, mechanical stress, etc.)
- $n^*(t)$ – the corresponding vector of target values (setpoints), but *allostatic*, that is:
- are not fixed “homeostatic” values, that is :
    - are not fixed “homeostatic” values,
    - can shift depending on context, time of day, anticipated effort, etc. (Sterling, 2012).
- Δn(t)=n(t)-n^* (t) – vector of deviations from the current setpoints (already after taking allostatic adjustments into account).This signal influences both:

1. **Semblion activation** – e.g., via the term γv_m above, where v_m is updated to reflect the long-term association between semblion $S_m$and positive/negative changes in $A(t)$;
2. **Plasticity** – potentiation or depression of connections involving active semblions is gated by $A(t)$.

Intuitively, episodes that improve bodily balance strengthen the associations among currently active semblions and interoceptive patterns and increase their long-term value v_m; episodes that worsen bodily balance weaken them or assign negative valence.

Box 1. refines the informal definition of $g$in Section 6.4.3 by proposing a concrete, MEM-compatible class of functions that can be implemented directly in simulations of embodied MEM-style agents.

Box 1. From allostatic deviation to global affect in MEM

In Section 6.4.3 we introduced the global affect signal as

$$A(t) = g(\Delta n(t)) \in \mathbb{R},$$

where $\Delta n(t) = n(t) - n^*(t)$ denotes the vector of deviations of current needs $n(t)$from their allostatic setpoints $n^*(t)$. Here we spell out a concrete MEM-consistent family of functions for g and clarify the role of the gain parameter κ.

Needs, allostatic setpoints, and deviations

Let

- $n(t) = \big(n_1(t), \dots, n_K(t)\big)$ be the current values of Kneeds or internal variables (e.g., energy level, tissue integrity, temperature, social contact, structural health in a robot),
- $n^*(t) = (n_1^*(t), \dots, n_K^*(t))$ be their context-dependent allostatic setpoints (which may drift slowly with time, task and developmental stage),
- $\Delta n(t) = n(t) - n^*(t)$ be the signed deviation of each need from its current target.

For each component $\Delta n_i(t)$, negative values mean "too low" (deficit, damage, unmet need) and positive values "too high" (excess, overload, potentially harmful surplus).

To make different needs comparable, we first define a normalized deviation

$$\widetilde{\Delta n}_i(t) = \frac{n_i(t) - n_i^*(t)}{s_i},$$

where $s_i > 0$ is a scaling constant that converts physical units of the i-th variable (e.g., Joules, °C, mechanical strain) into a dimensionless quantity of roughly comparable magnitude across needs.

Valence contributions of individual needs

MEM assumes that not all needs contribute equally to global affect, and that deficits and surpluses of the same physical variable may have different affective impact. We therefore introduce separate weights for negative and positive deviations:

- $w_i^- \geq 0$: affective weight of deficits of need i,
- $w_i^+ \geq 0$: affective weight of surpluses of need i.

We then define the signed contribution of each need $i$as

$$h_i(\widetilde{\Delta n}_i(t)) = -\, w_i^- \max(0, -\widetilde{\Delta n}_i(t)) \; + \; w_i^+ \max(0, \widetilde{\Delta n}_i(t)).$$

Intuitively:

- If $n_i(t)$ falls below its setpoint $n_i^*(t)$, then $\widetilde{\Delta n}_i(t) < 0$ and $h_i(\widetilde{\Delta n}_i(t)) \approx -w_i^- \mid \widetilde{\Delta n}_i(t) \mid$, which contributes negative valence (pain, threat, discomfort).
- If $n_i(t)$ rises above its setpoint, $\widetilde{\Delta n}_i(t) > 0$ and $h_i(\widetilde{\Delta n}_i(t)) \approx +w_i^+ \mid \widetilde{\Delta n}_i(t) \mid$, which contributes positive valence (relief, satisfaction, safety) or, for extreme overload, can be re-interpreted as risk.

The **net allostatic pressure** at time $t$is then

$$z(t) \;=\; \sum_{i=1}^{K} h_i\,(\widetilde{\Delta n}_i(t)).$$

Defining the affect function g and the gain κ

At the level of global affect, MEM uses a saturating nonlinearity to keep affective values within a bounded range and to capture the empirical fact that feelings do not grow without limit. A natural choice is: $A(t) \;=\; g(\Delta n(t)) \;=\; \kappa \tanh(z(t))$,

where:

- $\tanh(\cdot)$ is a smooth odd function that is approximately linear near zero and saturates for large $|\,z\,|$,
- $z(t)$ is the net allostatic pressure defined above,
- $\kappa > 0$ is a **global affective gain parameter**.

In this formulation:

- For small deviations from allostatic setpoints, $\tanh(z) \approx z$, so $A(t) \approx \kappa\, z(t)$ is approximately proportional to the weighted sum of normalized deviations.
- For large deviations, $|\,A(t)\,|$ saturates at κ, which can be interpreted as the maximal intensity of global affect the system can express or utilize.

Thus g is not an arbitrary mapping but a **structured aggregator** that:

- Normalizes each deviation by s_i,
- Weights each need by w_i^-, w_i^+,
- Sums these contributions into $z(t)$,
- Passes the result through a saturating nonlinearity scaled by κ.

This makes the allostatic status of the whole organism (or robot) available to the rest of the architecture as a single scalar signal $A(t)$ that modulates semblion activation and learning.

Relation to allostasis and slow variables

In an allostatic system, setpoints $n^*(t)$ and weights $w_i^{\pm}$ can themselves depend on longer-term context and chronic load. For example, we may define a slow "allostatic load" variable

$$L(t) \;=\; \sum_{i=1}^{K} \gamma_i \int_0^t e^{-(t-\tau)/\tau_{\text{allo}}}\,|\,\widetilde{\Delta n}_i(\tau)\,|\;d\tau,$$

which accumulates the history of deviations over a time-scale τ_{allo}, with sensitivity coefficients γ_i.

The gain parameter κ can then be made a function of $L(t)$, e.g.

$$\kappa(L) = \kappa_0\,(1 + \beta L) \text{ or } \kappa(L) = \frac{\kappa_0}{1 + \beta L},$$

depending on whether chronic allostatic load is assumed to **sensitize** (larger κ: hyper-reactivity) or **blunt** (smaller κ: affective flattening) the affective system. In this way, the same formalism covers both momentary homeostatic deviations and long-term allostatic dysregulation.

Connection to semblions and learning

In the MEM architecture, the global affect signal $A(t)$ couples back to semblion dynamics and plasticity. For example:

- In the semblion activation equation, $A(t)$ can multiplicatively gate the contribution of interoceptive components to semblion activation, boosting or suppressing emotional semblions depending on affect.

- In the learning rules, $A(t)$ can modulate Hebbian or STDP-like updates, strengthening connections during episodes of strong positive or negative affect and leaving them relatively unchanged when affect is near zero.

The gain κ therefore controls **how strongly allostatic state shapes** both the momentary conscious content (which semblions are active) and long-term memory structure (which semblions are consolidated or pruned).

6.4.4. Learning rule for forming and updating semblions

A minimal learning scheme for semblions can be formulated as follows. At each time step t:

1. Compute neural activations $h_i(t)$and semblion activations $a_m(t)$.
2. Identify a set of **winner** semblions $S_m$with $a_m(t) > \Theta$.
3. For each such semblion, update its parameters according to:
   - **Hebbian association within semblion**

$$\Delta w_{ij} = \eta\, A(t)\, h_i(t)\, h_j(t), \text{for } i,j \in V_m,$$

   where η is a learning rate and the sign/magnitude of $A(t)$modulates plasticity.

   - **Updating perceptual and feeling prototypes**

$$\boldsymbol{\mu}_m \leftarrow (1-\lambda)\,\boldsymbol{\mu}_m + \lambda\,\mathbf{h}_{V_m}^{sens}(t), \boldsymbol{\rho}_m \leftarrow (1-\lambda)\,\boldsymbol{\rho}_m + \lambda\,\mathbf{b}(t),$$

   where $\mathbf{h}_{V_m}^{sens}(t)$ denotes the sensory-dominant components of $\mathbf{h}_{V_m}(t)$ and $\lambda$is a slow consolidation rate.

   - **Updating long-term value**

$$v_m \leftarrow v_m + \kappa\, A(t),$$

   with κ a value-learning rate [2].

4. Periodically, perform a **structural update**: cluster co-activation patterns across time and split or merge semblions whose patterns have diverged or converged significantly.

This minimal rule expresses the core MEM intuition: semblions are formed and reshaped when sensory, mnemonic, and interoceptive patterns repeatedly co-occur under conditions of strong affect. They then serve as units of recognition, memory, and evaluation.

---

[2] If, for the implementation of MEM, we use von Mises/Gaussian-like kernels to model the projection of semblions onto sensory maps, then the coefficient **κ** specifies whether the re-entry targets a sharply defined pattern (high κ) or rather a more general, diffuse template (lower κ), (see Krause, Compte, Rademaker, 2025)

6.4.5. Pseudo-code for an online MEM-style agent

A simplified online cycle for an artificial MEM-style agent can be written as:

```
initialize semblion set {S_m} with random micro-structures
initialize needs n*, current n = n*

loop over time t:
    # 1. Sense
    x(t) ← exteroceptive sensors
    p(t) ← proprioceptive sensors
    b(t) ← interoceptive sensors
    n(t) ← update_needs(b(t))
    A(t) ← g(n(t) - n*)

    # 2. Perceive and recall
    h(t) ← forward_recurrent_pass(x(t), p(t), b(t), h(t-1))
    a_m(t) ← semblion_activations(h(t), b(t), S_m)
    H_conscious(t) ← reentrant_projection(active_semblions, sensory_maps)

    # 3. Plan and act
    world_state_estimate ← decode(H_conscious(t))
    candidate_actions ← generate_plans(world_state_estimate)
    predicted_consequences ← simulate_plans(candidate_actions, semblion_network)
    chosen_action ← select_action(predicted_consequences, A(t), needs)
    execute(chosen_action)

    # 4. Learn
    for each S_m with a_m(t) > Θ:
        update_weights_and_prototypes(S_m, h(t), b(t), A(t))
    maintain_semblance_structure({S_m})
end loop
```

This pseudo-code abstracts away from specific learning algorithms, but it highlights the key MEM cycle: perception of the world and body → activation of semblions → secondary perception and affect → action planning → affect-modulated motivated learning.

The MEM is a reductive model and therefore compatible with multiple realizability. Indeed, many research groups are developing biologically inspired architectures capable of forming semblions and of backward propagation of stimulation via the feed-forward sweep (FFS) and recurrent processing (RP).

Concepts of artificial associative-memory systems that exhibit some of the features required by the MEM model have been proposed, based on a hierarchical, multilayer connectionist structure with linear combinatorial complexity designed for parallel processing (Horzyk 2017). Adrian Horzyk presented this idea as a hierarchical neural network structure composed of specialized spiking neurons, whose interactions are modeled using graphs (Horzyk, 2013).

Neurons, with their presynaptic fields, dendritic inputs, and axons, correspond to neural memory fields at each layer of a semblion. Their principal function is to detect similarity relations. Associative relations map onto the heuristics of the neural memory field that regulate the propagation of stimulation to higher or adjacent layers, where contextual relations are established.

Attention should also be given to a new paradigm proposed by Dorian Aur, termed **Electrodynamic Intelligence (EDI)**. According to this view, the formation of semblions precludes any principled separation of information processing from storage in the brain. Signal transmission, stored patterns, and encoded meaning are co-located in the same physical substrate. Information is not kept in an external register or in discrete functional modules but is instead embedded in the evolving structure of the system's electric field (Aur, 2025a; 2025b). Parallel processing is assumed: processing and storage arise from the self-consistent interactions of the electric-field configurations of perceived stimulations, ensuring complementarity with the spatial distribution of micro-electric fields of proteins

in dendritic spines and synaptic channels of neurons. Conformational changes in protein chains under the influence of signal fields guarantee memory plasticity.

The mode of these interactions is described by **NeuroElectroDynamics (NED)** developed by Aur and Jog (Aur & Jog, 2010). He cites examples of memristive structures capable of supporting ElectroDynamic Intelligence (EDI) (e.g., Caravelli et al., 2021; Maheshwari et al., 2022; Brivio et al., 2022; Kim et al., 2024). In such systems, logic circuits and the neurons of a digital network are replaced by an electrodynamic processing core composed of spatially arranged elements that behave analogously to neurons and synapses, coupled ephaptically and operating through the propagation and interference of electric fields in memristive channels. These systems provide vast capacity, high operating speed, and extremely low power consumption.

At present, however, such systems still lack the capacity for lateral and multimodal associations, backward stimulation via recurrent processing (RP), and secondary perception. Nor are they embodied in a humanoid form. Nevertheless, the rapid pace of technological progress gives reason to expect that, in the not-too-distant future, we will see embodied AGI systems equipped with interoceptors and diverse exteroceptors, capable of rich behavioral responses and able to serve as assistant-companions offering guidance and genuinely empathetic care, arising from sincere intentions and supported by robust feelings of concern, understanding, and warmth.

7. LLM reasoning, world models, and the MEM perspective on "thinking"

Recent work has intensified the debate about whether next-token prediction, when scaled sufficiently and combined with appropriate prompting, can give rise to something that deserves to be called "reasoning" or even "world modelling". Proponents of the **"LLMs as emerging reasoners"** view point to several phenomena:

- **Chain-of-thought prompting and self-consistency**: when encouraged to "think step by step", large models often produce multi-step derivations that approximate human-like problem solving in mathematics, logic puzzles, and scientific question-answering.
- **Tool-augmented reasoning**: when integrated with external tools—calculators, code interpreters, web browsers—LLMs can write and debug programs, design and test hypotheses, and iteratively improve their own outputs, exhibiting a kind of meta-level problem solving.
- **Latent conceptual structure**: linear probes and representation analysis suggest that LLM embedding spaces reflect semantic categories, syntactic roles, and even some aspects of physical and social structure; this has prompted the suggestion that "world models" are implicitly encoded in the statistics of language.

On this optimistic picture, next-token prediction serves as a universal learning objective: because language compresses regularities of the world and social interaction, a sufficiently large model trained on enough text might acquire a usable internal model of the world, including logical and causal relations. Skeptics respond that these demonstrations, while impressive, may still reflect **sophisticated pattern completion** rather than genuine understanding. The reliability of LLM "reasoning" degrades under distribution shift, adversarial examples, or slight changes in problem presentation. Models can produce elaborate but nonsensical derivations, revealing that the underlying process is not constrained by an independent grasp of truth conditions or causal structure, but by surface regularities of the training corpus. They also lack the capacity to learn from *their own* embodied experience; any "world knowledge" is inherited passively from human-generated text.

From the perspective of the MEM framework, this controversy can be clarified by explicitly distinguishing the two senses of "thinking" already invoked in our article: *functional thinking* and *phenomenally grounded thinking*, as introduced in Section 2 and elaborated in Section 5.1.

The MEM model is neutral about how powerful functional thinking can become in purely disembodied architectures. It is compatible with the idea that, given enough parameters and training data, LLM-based systems could outperform humans on many forms of abstract reasoning, scientific discovery, or software design. It also acknowledges that multimodal models trained on images and text acquire richer exteroceptive structure than purely textual models.

Where MEM takes a principled stance is on **the conditions under which such thinking becomes conscious and affect-laden**. In particular:

- Next-token prediction over text (even when extended to images) does not in itself create **constitutive grounding**: token embeddings are not directly tied to the system's own sensory and bodily states. They are grounded, at best, in *other people's* experiences, distilled into language.
- Functional reasoning in LLMs lacks **causal grounding** in the sense that its representational repertoire does not depend on a history of embodied interaction and interoceptive regulation. Changing the body or environment of an LLM-agent does not alter its internal representations unless this change is reflected in its training corpus.
- The "world models" encoded in LLM weights are therefore **vicarious**: they reflect an implicit summary of how humans talk about the world, not how the system itself experiences it.

From a MEM perspective, this means that LLMs and current VLAs instantiate a powerful **simulation of thinking**, but not thinking-with-awareness. Their internal states do not pass through the secondary-perception loop in which semblions, enriched by affect and bodily stakes, are re-installed in sensory and interoceptive maps. As a result, their reasoning, however sophisticated, lacks the felt aspect of understanding that characterizes human cognition.

At the same time, the recent advances in LLM reasoning and world modelling are highly relevant for MEM-inspired AI. They show that:

- High-capacity predictive architectures are capable of forming rich, compressed latent spaces that could serve as the **associative core** of a semblion network.
- When coupled to robotic bodies and interoceptive sensors, these architectures could be repurposed as **semantic engines**, binding linguistic and perceptual labels to semblions that are genuinely grounded in the agent's own experience.

In this sense, MEM does not dismiss the importance of next-token prediction. Rather, it argues that **prediction must be embedded in the right kind of body-brain loop**—with interoception, needs, and re-entrant perception—before we should talk about machines that not only think functionally but also **experience** their thinking.

8. Limitations of MEM and alternative pathways for conscious AI

Throughout this article, we have used MEM as a systematic framework for distinguishing between mere simulations of cognition and architectures that might one day support conscious thinking and feeling. It is important, however, to emphasize that MEM is offered as a **hypothesis and research program**, not as a final theory.

Several limitations deserve explicit mention:

1. **Lack of a full-scale implementation**
   While we have outlined key mechanisms (semblions, secondary perception, interoceptive feelings, motivational hierarchies), MEM has not yet been realized in a large-scale, functioning artificial agent. The formalization sketched above provides a starting point, but substantial

engineering work is required to demonstrate that a semblion-based architecture can support the breadth and robustness of human-like cognition.

2. **Biophysical commitments and open questions**
   MEM is deeply rooted in specific neurobiological mechanisms, including dendritic integration, ephaptic coupling, and the NeuroElectroDynamics / ElectroDynamic Intelligence paradigm. These commitments are scientifically grounded but not yet universally accepted. Some elements of the hypothesis—for example, the role of complex electric fields in memory and representation—remain controversial and require further empirical validation.
3. **Embodiment and interoception as necessary conditions**
   We have argued that embodiment with interoception is necessary for genuine feeling and, by extension, for conscious thinking. This claim is strongly motivated by comparative neurobiology and by the perceptual theory of emotion, but it is not a logical truth. Alternative theories—such as purely information-theoretic approaches (IIT), global broadcasting theories (GWT), or some forms of predictive processing—suggest that certain highly integrated, functionally rich systems might support consciousness even in the absence of traditional bodily interoception. MEM takes a more conservative stance: it posits that, for systems like us, consciousness is realized in an embodied, affect-laden architecture, and that any artificial system claiming similar phenomenology will likely need analogous components. Whether radically different architectures could support subjectivity remains an open question.
4. **Scope and species-relative character of the criteria**
   The criteria we have derived for evaluating AI systems are tailored to mammalian-like organisms and MEM-inspired robots. They may not apply straightforwardly to alien minds or radically different artificial architectures. For instance, a future theory might identify a distinct but functionally equivalent substrate that plays the role of interoception or needs. MEM can be generalized to some extent (e.g., by abstracting "body" to any physically extended system with self-maintenance requirements), but its true scope remains to be tested.
5. **Phenomenology and underdetermination**
   Even if a MEM humanoid were built and passed all the tests we propose (secondary perception, affective simulation, value-modulated planning), this would not logically prove that it is conscious. Competing theories could reinterpret the same data in their own terms. In this respect, the empirical situation for consciousness will likely remain underdetermined by any one behavioral or neural marker. What MEM offers is a coherent package of **mechanisms and predictions** that can be compared with the predictions of rival approaches, not a metaphysical guarantee.

In light of these limitations, our central claim should be read as follows:

MEM provides a detailed, biologically motivated, and operationally rich account of how conscious thinking and feeling arise in organisms like us. It yields a set of concrete criteria and tests for assessing artificial systems. We propose, as a working hypothesis, that architectures satisfying these conditions are strong candidates for conscious agency, whereas architectures lacking them are not.

Alternative paths to conscious-like AI remain conceivable. For example, a sufficiently complex, self-organizing predictive-processing system might discover internal variables that function analogously to interoception, even if they are not directly tied to bodily organs, or a future electrodynamic processor might realize semblion-like structures in a non-neuronal substrate. The MEM framework does not exclude such possibilities a priori; instead, it offers a benchmark: any viable alternative must show how it can reproduce the key explanatory work done by semblions, secondary perception, and affective grounding.

9. Conclusion

The analysis presented in this paper leads to a clear distinction between the functional intelligence of current large language models and the embodied, affective cognition of biological organisms. LLMs, though capable of impressive reasoning and multimodal processing, operate without grounding in sensory or interoceptive experience. Their "understanding" is confined to correlations within linguistic and perceptual data, lacking the constitutive and causal grounding in perception and emotion that the Motivated Emotional Mind (MEM) model identifies as prerequisites for conscious thought.

The condition of constitutive and causal grounding, as articulated by Chalmers, is necessary for the emergence of conscious thinking and feeling. Systems operating in accordance with the MEM model satisfy these conditions. However, they are not sufficient conditions, even when jointly met. According to MEM, genuine thinking and feeling arise from the integration of perception, emotion, and motivation within a recurrent and embodied system. Conscious thought is not merely information processing but the experience of meaning generated through the organism's affective relation to the world. This requires embodiment, interoception, and feedback loops between cognitive and sensory layers—features absent from disembodied AI systems.

Future progress toward machines capable of authentic thinking and feeling must therefore proceed through embodied architectures that replicate these principles: systems endowed with sensory and interoceptive mechanisms, associative memory structures (semblance networks), and affective valuation loops. Only such architectures, developed in accordance with the MEM model, could eventually give rise to artificial agents that not only act intelligently but also *experience* their actions and internal states.

At the same time, we do not claim that MEM offers the only conceivable pathway to artificial consciousness, nor that embodiment and interoception are logically necessary for any possible conscious system. Rather, we put forward MEM as a biologically inspired hypothesis about how consciousness is realized in creatures like us and as a practical framework for evaluating future AI. The criteria and tests derived from MEM—semblance-based memory, secondary perception, interoceptive feelings, and value-modulated planning—can be applied to forthcoming generations of embodied AI, including systems that integrate powerful LLM-style predictors with rich sensorimotor and interoceptive substrates. If such systems begin to satisfy these criteria in increasingly robust ways, MEM will gain empirical support as a theory of conscious thought; if they do not, the theory will need to be revised or replaced. Either way, MEM provides a concrete research program for moving from speculative talk about "thinking machines" toward experimentally grounded assessments of when and how artificial systems might come to both think and feel.

Appendix A

A MEM-based evaluation matrix for artificial systems

To make the comparison between contemporary AI systems and MEM more systematic, we derive a set of observable criteria directly from the MEM architecture and use them to evaluate three classes of systems: (i) text-only LLMs, (ii) multimodal models and robots (e.g., PaLM-E-style VLAs), and (iii) a hypothetical MEM-consistent humanoid.

We group the criteria into four domains: embodiment and sensing, representational architecture, affect and motivation, and planning and counterfactual simulation. Each criterion can be scored on a coarse 0–2 scale:

- **0 – absent** (no implementation or only trivial hard-coding),
- **1 – partial** (rudimentary or indirect implementation),
- **2 – substantial** (architecture implements the MEM requirement in a clear and non-trivial way).

Criteria

1. **E1: Rich exteroceptive and proprioceptive sensing**
   Does the system possess high-bandwidth sensory channels and a representation of its own posture/movement?
2. **E2: Interoceptive sensing of internal variables**
   Does the system monitor internal physical states relevant to its survival or long-term functioning (e.g., energy, temperature, strain, damage)?
3. **R1: Heterarchical associative memory (semblance-like coding)**
   Does it employ a multi-layer, recurrent, associative architecture in which distributed patterns function as units of memory and recognition, rather than purely symbolic tables or static embeddings?
4. **R2: Secondary perception (re-entrant activation of sensory maps)**
   Can internal processes (recall, imagination, planning) re-activate early sensory or sensory-like representations in a way that is structurally similar to perception?
5. **A1: Needs and allostatic variables**
   Does the system maintain internal "needs" or goals that evolve autonomously and have to be managed over extended periods, rather than only responding to externally provided task rewards?
6. **A2: Learned coupling between interoception and value**
   Are internal physical states learned, over time, as predictors of positive or negative outcomes, and do they shape subsequent behavior (somatic-marker-like effects)?
7. **P1: Counterfactual simulation in a perception-like format**
   Does the system simulate alternative futures by re-activating sensory/interoceptive patterns, rather than only manipulating symbols or scalar costs?
8. **P2: Value-modulated policy selection**
   Are action policies selected on the basis of affective evaluations grounded in internal bodily variables, as opposed to externally defined reward signals alone?
9. **L1: Language grounded in non-linguistic representations**
   When using language, does the system bind words and sentences to underlying perceptual and interoceptive semblions, or are linguistic operations confined to text-text or image-caption mappings?

Scoring

The table below illustrates how these criteria might be applied to three representative system types. The scores are indicative rather than definitive and can be refined as architectures evolve.

Criterion	Text-only LLM (GPT-style)	Multimodal VLA / robot (PaLM-E-style)	MEM-consistent humanoid (hypothetical)
E1. Exteroception & proprioception	0 – no direct sensing; only text tokens	1 – cameras, joint encoders, sometimes tactile	2 – full exteroceptive and proprioceptive suite
E2. Interoception	0 – no bodily sensors	0–1 – battery/temperature occasionally read, but not richly represented	2 – multi-dimensional internal sensing of energy, strain, damage, etc.

Criterion	Text-only LLM (GPT-style)	Multimodal VLA / robot (PaLM-E-style)	MEM-consistent humanoid (hypothetical)
R1. Heterarchical associative memory	1 – deep network of activations but no explicit semblion organization, no episodic grounding	1 – similar, with additional perceptual embeddings; some episodic traces	2 – explicit semblion network spanning sensory, mnemonic and affective layers
R2. Secondary perception	0 – no re-entry into sensory maps; "imagery" is linguistic	1 – internal visual layers may be modulated, but generally not perception-like in planning	2 – recurrent feedback from memory into early sensory/interoceptive maps for imagery and feeling
A1. Needs & allostasis	0 – goals externally specified per prompt or fine-tuning	1 – simple internal costs (battery, safety margins)	2 – explicit needs and allostatic variables steering behavior over long horizons
A2. Learned interoception–value coupling	0 – no bodily substrate; "valence" is symbolic	0–1 – limited; some cost functions reflect hardware limits but rarely learned	2 – interoceptive patterns learned as predictors of harm/benefit and reused in planning ("pain" semblions)
P1. Perception-like counterfactual simulation	0 – planning is symbolic/textual or via external tools	1 – some internal forward modelling, mostly symbolic or low-dimensional	2 – imagined scenarios re-instantiate sensory and interoceptive semblions
P2. Value-modulated policy selection	1 – policies optimized for external objectives but not bodily value	1 – optimized for task rewards and some hardware constraints	2 – policy selection shaped by affective reactivation of bodily states
L1. Language grounding	0–1 – grounded only in patterns of text and pre-processed images	1 – partially grounded in perception and action; still weakly tied to interoception	2 – language bound to semblions that integrate perception, action and feeling

This evaluation matrix serves two purposes. First, it makes the notion of "implementation of MEM components" more precise and falsifiable: any proposed conscious AI must be evaluated against these dimensions, not by vague appeals to "intelligence". Second, it shows that current LLMs and VLAs, while remarkably capable in certain dimensions (R1 and partially P2), are systematically lacking in interoception, affective coupling, and perception-like counterfactual simulation—the very elements MEM regards as crucial for genuine feeling and conscious thought.

References:

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., … & Zoph, B. (2023). GPT-4 Technical Report. *ArXiv*. https://arxiv.org/abs/2303.08774

Aur, D. (2025a) When Matter Thinks: The Physics of Experience in Electrodynamic Intelligence Systems. *Zendo* DOI: 10.5281/zenodo.17087828.
Aur, D., (2025b) The Second Wave: From Neuroelectrodynamics to Electrodynamic Intelligence. *TechRxiv.* DOI: 10.36227/techrxiv.175623132.24668882/v1
Aur, D., & Jog, M. S. (2010). *Neuroelectrodynamics - Understanding The Brain Language.* IOS Press.
Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.
Barrett, L. F. (2017). How emotions are made: The secret life of the brain. Houghton Mifflin Harcourt.
Bender, E. M., & Koller, A. (2020, July). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 5185-5198).
Brivio, S., Spiga, S., & Ielmini, D. (2022). HfO2-based resistive switching memory devices for neuromorphic computing. *Neuromorphic Computing and Engineering*, *2*(4), 042001.
Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., ... & Zitkovich, B. (2022). Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817.
Brubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.
Candia-Rivera D., Engelen T., Babo-Rebelo M., Salamone P.C. (2024) Interoception, network physiology and the emergence of bodily self-awareness, *Neuroscience & Biobehavioral Reviews,* Vol.165, 105864, DOI: https://doi.org/10.1016/j.neubiorev.2024.105864.
Cangelosi, A., & Schlesinger, M. (2015). *Developmental robotics: From babies to robots*. MIT press.
Caravelli, F., Sheldon, F. C., & Traversa, F. L. (2021). Global minimization via classical tunneling assisted by collective force field formation. *Science Advances*, *7*(52), eabh1542
Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181-204.
Chalmers, D. J. (1996). *The conscious mind: in search of a theory of conscious experience*. New York: Oxford University Press.
Chalmers, D. J. (2024). Does thought require sensory grounding? From pure thinkers to large language models. *arXiv preprint arXiv:2408.09605*.
Craig, A. D. (2014). How do you feel?: an interoceptive moment with your neurobiological self. In *How Do You Feel?*. Princeton University Press.
Cuskley, C., Woods,R., Flaherty, M.; (2024) The Limitations of Large Language Models for Understanding Human Language and Cognition. *Open Mind*; 8 1058–1083. doi: https://doi.org/10.1162/opmi_a_00160
Damasio, A. (1991). *Somatic Markers and the Guidance of Behavior.* New York: Oxford University Press. pp. *217*–299.
Damasio, A.R., Tranel, D., & Damasio, H. (1991). *Somatic markers and the guidance of behavior: theory and preliminary testing*, in: *Frontal Lobe Function and Dysfunction*, Oxford University Press.
Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *351*(1346), 1413-1420.
Davis, K. L., & Montag, C. (2019). Selected Principles of Panksepp ian Affective Neuroscience. *Frontiers in Neuroscience*, *12*, 427897. https://doi.org/10.3389/fnins.2018.01025
Dehaene, S., & Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70*(2), 200-227.Dretske, F. (1995). *Naturalizing the Mind*. MIT Press.
Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Wahid, A., ... & Florence, P. (2023). Palm-e: An embodied multimodal language model. *ArXiv*. https://arxiv.org/abs/2303.03378
Feldman MJ, Bliss-Moreau E, Lindquist KA. (2024) The neurobiology of interoception and affect. Trends Cogn Sci;28(7):643-661. doi: 10.1016/j.tics.2024.01.009.
Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, *1*(1), 1-47.
Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature reviews neuroscience*, *11*(2), 127-138.

Galus, W.L, (2018) Semblions of Words. The Language of Natural and Artificial Neural Networks. *Qeios* 1ATS9M. DOI: https://www.qeios.com/read/1ATS9M.2

Galus, W.L. (2023a) Different Aspects of Consciousness Explained by Distinct Biophysical Processes. *Journal of Theoretical and Philosophical Psychology.*, Advance online publication: https://doi.org/10.1037/teo0000236

Galus, W. (2023b) Mind–brain identity theory confirmed? *Cogn Neurodyn* **18**, 1467–1487. DOI: https://doi.org/10.1007/s11571-023-09992-6

Galus, W.L. (2025a) Perception as the Essence of Phenomenal Consciousness. Available at *SSRN*: http://dx.doi.org/10.2139/ssrn.5207381

Galus, W.L., (2025b) A new physicalist model of consciousness: a proposal. Available at *SSRN*: DOI: http://dx.doi.org/10.2139/ssrn.5207389

Galus W.L., Starzyk J.A., (2020) *Reductive Model of the Conscious Mind*. IGI Global, ISBN13: 9781799856535; DOI: 10.4018/978-1-7998-5653-5

Grillner, S. (2006). Biological pattern generation: the cellular and computational logic of networks in motion. *Neuron*, *52*(5), 751-766.

Horzyk A. (2013). *Sztuczne systemy skojarzeniowe i asocjacyjna sztuczna inteligencja*. Akademicka Oficyna Wydawnicza. ISBN-13: 978-83-7837-525-8

Horzyk, A. (2017). *Deep Associative Semantic Neural Graphs for Knowledge Representation and Fast Data Exploration.* Proc. of KEOD 2017, 67-79. DOI: 10.5220/0006504100670079

James W (1884) What is an emotion? *Mind* 9*,* 188–205.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335-346.

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., ... & Wei, F. (2023). Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, *36*, 72096-72109. *ArXiv*. https://arxiv.org/abs/2302.14045

Kim, K., Song, M. S., Hwang, H., Hwang, S., & Kim, H. (2024). A comprehensive review of advanced trends: from artificial synapses to neuromorphic systems with consideration of non-ideal effects. *Frontiers in Neuroscience*, *18*, 1279708.

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., ... & Finn, C. (2024). Openvla: An open-source vision-language-action model.*arXiv preprint arXiv:2406.09246*.

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, *4*, 169.

Kosslyn, S. M. (2005). Mental images and the brain. *Cognitive neuropsychology*, *22*(3-4), 333-347.

Krause, N. N., Compte, A., Rademaker, R. L., (2025) Top-down feedback can explain the existence of working memory traces in early visual cortex. *bioRxiv* 2025.11.27.690959; doi: https://doi.org/10.1101/2025.11.27.690959

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, 10(11), 494–501.

Lamme, V. A. F., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11), 571–579.

LeDoux, J. E., & Brown, R. (2017). A higher-order theory of emotional consciousness. *Proceedings of the National Academy of Sciences*, *114*(10), E2016-E2025.

Li, M., Su Y., Huang H-Y., Cheng J., Hu X., Zhang X., Wang H., Qin Y, Wang X., Lindquist K.A., Liu Z., Zhang D.,(2024) Language-specific representation of emotion-concept knowledge causally supports emotion inference, *iScience,* Vol. 27, 12: 111401

Lycan, W. G. (1996). *Consciousness and Experience*. MIT Press.

Maheshwari, S., Serb, A., Papavassiliou, C., & Prodromakis, T. (2022). An adiabatic capacitive artificial neuron with RRAM-based threshold detection for energy-efficient neuromorphic computing. *IEEE Transactions on Circuits and Systems I: Regular Papers*, *69*(9), 3512-3525.

Mitchell, M. & Krakauer, D.C. (2023) The debate over understanding in AI's large language models, *Proc. Natl. Acad. Sci. U.S.A.* 120 (13) e2215907120, https://doi.org/10.1073/pnas.2215907120.

Mulligan, K. (2004). Brentano on the Mind. *The Cambridge Companion to Brentano*, 66-97.

Niu, Q., Liu, J., Bi, Z., Feng, P., Peng, B., Chen, K., ... & Liu, M. (2024). Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. *arXiv preprint arXiv:2409.02387*.

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions.* Oxford University Press.

Pezzulo G, Parr T, Cisek P, Clark A, Friston K. (2024) Generating meaning: active inference and the scope and limits of passive AI. *Trends Cogn* Sci;28(2):97-112. doi: 10.1016/j.tics.2023.10.002.

Rolls, E. T. (2014). Emotion and decision-making explained: a précis. *Cortex*, *59*, 185-193.

Rosenthal, D. M. (2005). *Consciousness and Mind*. Oxford University Press.

Sapkota R., Cao Y., Roumeliotis K.I., Karkee M. (2025) Vision-Language-Action Models: Concepts, Progress, Applications and Challenges, *arXiv*:2505.04769v1 [cs.CV] https://arxiv.org/html/2505.04769v1

Scherer, K. R. (2005). What are emotions? And how can they be measured?. *Social science information*, *44*(4), 695-729.

Singer, J. L. (1978). Experimental studies of daydreaming and the stream of thought. In *The stream of consciousness: Scientific investigations into the flow of human experience* (pp. 187-223). Boston, MA: Springer US.

Slotnick, S. D. (2017). *Cognitive neuroscience of memory*. Cambridge University Press.

Shanahan, M. (2010). *Embodiment and the inner life: Cognition and Consciousness in the Space of Possible Minds*. Oxford University Press.

Sterling, P. (2012). Allostasis: a model of predictive regulation. *Physiology & behavior*, *106*(1), 5-15.

Team, G. R., Abeyruwan, S., Ainslie, J., Alayrac, J. B., Arenas, M. G., Armstrong, T., ... & Zhou, Y. (2025). Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*.

Tononi, G. (2008). "Consciousness as Integrated Information: a Provisional Manifesto." *The Biological Bulletin*, 215(3): 216–242.

Tononi, G., & Koch, C. (2015). Consciousness: here, there and everywhere?. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1668), 20140167.

Tye, M., (2000). *Consciousness, Color, and Content*. MIT Press.

Wang, J., Jiang, H., Liu, Y., Ma, C., Zhang, X., Pan, Y., ... & Zhang, S. (2024). A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *arXiv preprint arXiv:2408.01319*.

Xu, C., & McAuley, J. (2023, June). A survey on model compression and acceleration for pretrained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 9, pp. 10566-10575).

Yang, Y., Zhou, T., Li, K., Tao, D., Li, L., Shen, L., ... & Shi, Y. (2024). Embodied multi-modal agent trained by an llm from a parallel textworld. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 26275-26285).

Yoshida N., Kanazawa H., Yasuo Kuniyoshi Y. (2024) Synthesising integrated robot behaviour through reinforcement learning for homeostasis. *bioRxiv* 2024.06.03.597087; doi: https://doi.org/10.1101/2024.06.03.597087

Yoshida, N., & Man, K. (2025). Homeostatic Coupling for Prosocial Behavior. *ArXiv*. https://arxiv.org/abs/2506.12894

Zitkovich, B., et.al. (2023) RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. *Proceedings of The 7th Conference on Robot Learning, in Proceedings of Machine Learning Research*: 229:2165-2183 Available from https://proceedings.mlr.press/v229/zitkovich23a.html.

Zhou X., Menassa C.C., Kamat V.R., (2025) Interoceptive Robots for Convergent Shared Control in Collaborative Construction Work, *arXiv:2501.09290* [cs.RO] https://doi.org/10.48550/arXiv.2501.09290