ALETHEIA | RESEARCH JOURNAL

GenAI as Facade of Cognitive Extension

How Abstraction, Non-Determinism, and Sensory Decoupling in Generative AI Implementations Cause Systemic Failures

Author: Wesley Winkler, Metanoia Systems Research.
Keywords: Cognition, AI, Extended Cognition, Representational Systems, Abstraction
JUN 16 2026

Abstract

The prevalence of artificial intelligence systems, specifically generative artificial intelligence systems based on large-scale, billion and multi-billion parameter transformer models of human language, has become deeply integrated into nearly every existing major software suite on the market. It becomes appropriate to ask: what are the cognitive trade-offs, and how should we regard this technology that has now propagated into the mainstream so deeply? Namely, should Generative AI (GenAI) be treated as just another tool that can extend human ability, an extension of the human body and will, like a hammer? Should it be treated as deterministic automation? Or should it be treated as something else entirely?

In this perspective paper, I will argue that extension of cognition while removing sensory feedback through multiple layers of abstraction, severs the situated environmental connection needed for true skill acquisition in novices, and skill retention in experts. To do this, I will frame current interactions with GenAI through variations on a proposed Quad-Process Model (QPM) of cognition from the fields of artificial intelligence and robotics, which are variations on Kannanman's Dual Process Model which breaks cognition into two complementary “systems” of cognition.

I will then make the argument that GenAI should not be treated as a pure mechanism for extended cognition in the manner in which it is currently being incorporated into the cognitive systems of human organizations. In fact, it is precisely the facade of extended cognition that creates organizational and operational risk in GenAI implementations through auto-regressive tendencies, and forced abstraction.

I will make this argument on the basis of three main claims:

1. The use of language is a symbolic abstraction layer that sits between the environment and environmentally situated signals, boundaries, and skill based affordances.
2. This abstraction inherently reduces or eliminates systemic, sensory and signal feedback loops which undercuts the ability of individuals and systems to develop or maintain assembled affordance-based skills in their environment.
3. The inability to develop skills and dynamically tune to new sensory signals means that system resilience and precision will degrade over time due to a reduction in situated, environmental intuition and skill acquisition, creating regression to the mean and the potential for catastrophic cascading failure.

Taken together, I argue that GenAI should be treated as an extensible abstraction layer based on existing cultural and symbolic artifacts of meaning (System 3). Not as a foundational pre-cognitive processing layer (System 0). More clearly defined domain boundaries must be drawn between integrated cognitive tools and the human individual or collective human systems organized around them. To treat GenAI as System 0 is to invite skill atrophy, technical debt, and the impingement of system metacognitive ability.

Background: What Is Cognition, and How Does It Extend?

Cognition, broadly defined, is: "*all the activities and processes concerned with the acquisition, storage, retrieval and processing of information — regardless of whether these processes are explicit or conscious.*" (Bayne et al., 2019) These processes, while often considered computational in nature, are not restricted to within a brain or even a body. It is important to the substance of the argument that we understand the concepts of embodied, extended, situated, and neuro-centric cognition. Each of these concepts represent different schools of thought around the study of cognition within a specific behavioral system.

Cognition is about decision-making and decision calculus, both passive and subconscious and conscious, intentional, and rational, which we will contextualize through Kahneman's lens of System One and System Two thinking (Gulati et al., 2020; Li et al., 2025), though this framing tends to be overly reductionistic, as both are constantly

happening at the same time in something closer to infinite recursion than a clean binary. However, the prevalence of Kahneman's work in the public consciousness means that it is a useful abstraction when it comes to different "types" of cognition. For further reading, one should consider activation, suppression, and integration of the Default Mode Network (DMN) in cognitive tasks. The DMN is a set of pathways in the brain which are involved in task-agnostic internal meaning making and cognition. The DMN supports metanarrative formulation, self-identification, and concept linkages. The DMN couples to task specific pathways and neuromodules in order to support high effort task execution, but also is engaged in low effort heuristic reasoning. (Menon V., 2023) Bottom line, it gets nuanced quite quickly, which is appropriate for neuroscience, but is less helpful with scaffolding for the conversation at hand. Therefore, we will be adopting Kahnemen's language and concept model for the rest of the paper.

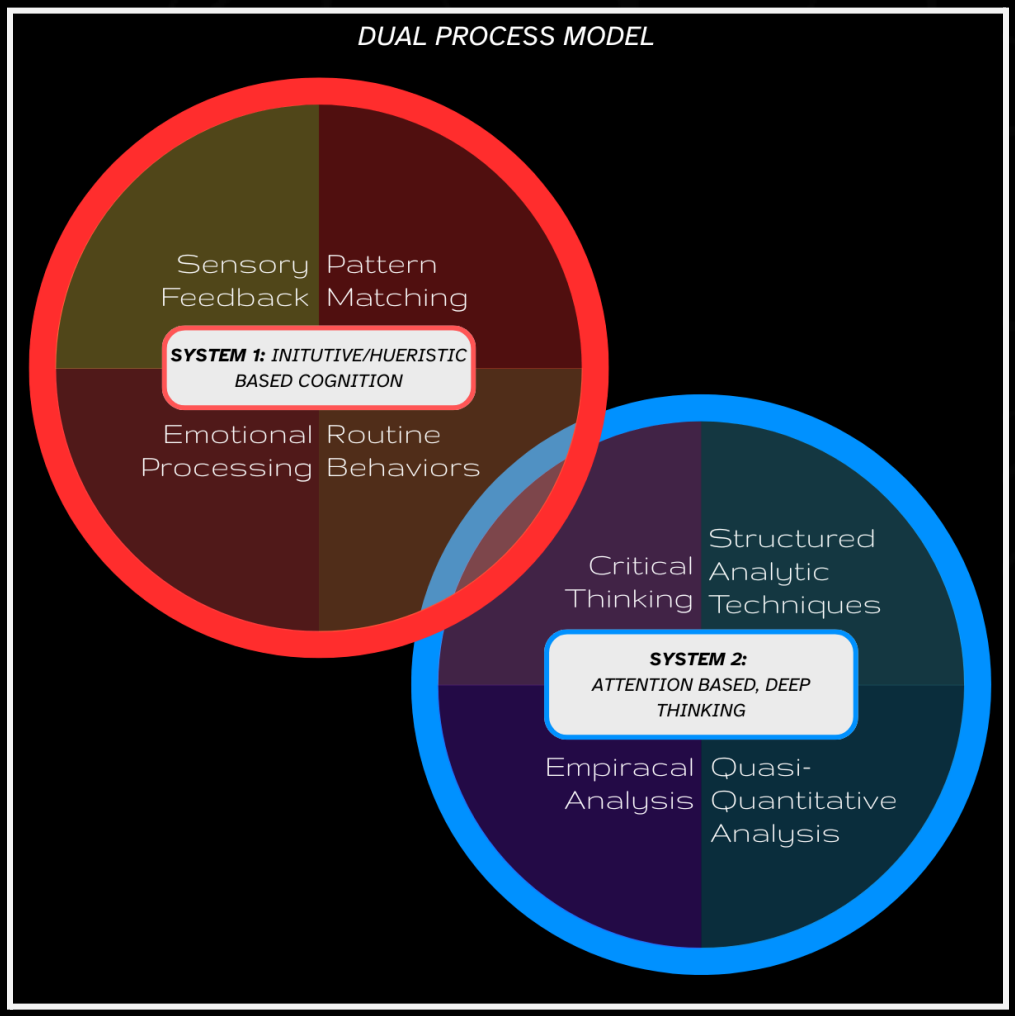


Figure 1: System 1 vs System 2, inspired by (Pherson et al., 2024)

Neuro-centric cognition holds that all thought and ideas originate inside the neurons of the brain, that specific modules work together through cross-module connections and inputs to create our conscious experience, process sensory data, and so on. While this paper does not go into theories of consciousness directly, it is worth noting that conscious experience is an element of human cognition which is not well addressed by computational systems (Wang et al., 2024).

Embodied cognition, on the other hand, is a broad category that generally studies the interplay between somatic states and neural states. It purports that cognition is distributed throughout the body, not merely in the brain itself. Sensory feedback becomes the input and the mechanism by which we make subconscious decisions, and it is how our brain creates subconscious pattern matches which enable us to develop our superpowers of intuition, developed upon exposure to a variety of external systems. The role of metacognition, the capacity to monitor and regulate one's own cognitive processes, is equally foundational; higher metacognitive awareness correlates with more deliberate, analytical decision-making, while weaker metacognitive regulation tends toward intuitive or avoidant styles (Basu & Dixit, 2022). This is explored accessibly in Annie Murphy Paul's *The Extended Mind (ISBN: 0544947665)* , which is worth reading in full.

Extended cognition, popularized by Andy Clark and David Chalmers, posits that our mind and our cognition are not merely limited to our bodies, but that we can offload cognitive tasks to gain additional computational load and decision-making power onto external systems. The classic example is Otto's notebook, a thought experiment in which a man named Otto, having developed dementia, begins writing things down: the address where he must go to see his granddaughter's piano recital, for instance. In this sense, Otto's notebook becomes an extension of his cognitive ability. When he needs to recall something, he opens his notebook, finds the address, and navigates to that location. By contrast, his daughter, also attending the recital, has the address memorized; when it comes time, she recalls it through memory and navigates accordingly. In both cases, the individuals arrive at the recital hall, though by different mechanisms of information retrieval and cognition. Extended cognition has notable variations which encompass critical differentiations. Clark and Chalmers initial thesis treats anything that performs a cognitive task, as integrated with the mind. Subsequent critiques of this position, such as by Sutton and Manary treat the mind as extensible, and external tools compliment the mind to extend its capabilities. (Menary, 2012; Sutton, 2010)

Studies around tool use and professional sports further illustrate this. The dynamics of tool use such as a professional baseball player swinging a bat show that learned sensory-motor patterns mean that when a professional picks up a tool of their trade, they are able to integrate how that tool gives biofeedback into their bodies, and thus feel or explore the medium in which they are participating. Just as a blacksmith might use a specific hammer to find impurities in a block of steel by hammering and listening for a precise tone or reverberation, or when shaping steel, might use a heavier hammer and feel how much the steel moves, how much reverberation travels into the handle, so too does expert tool use become an extension of bodily sensory exploration. (Baber et al., 2014; Keller & Janet Dixon Keller, 1996) This enables expert-level pattern matching and prunes neural connections that are not necessary for the task. In this case, pruning connections is more important than building new ones, making the pathways more efficient. Effective tool use depends on perceiving affordances (what actions are possible with the tool in context), not just recalling stored “rules” (Ibid, 2014) By smoothing the path to an answer, GenAI bypasses the neural 'stumbling' required for pruning and reinforcement.

Situated Cognition posits that all cognition happens within a given environment, and that actions and knowledge is constrained to within a specific set of conditions. While the closely linked ***Distributed Cognition*** *represents* how cognition is spread throughout all information processing and interpretive capabilities in a given system, including people, cultural systems, ontology, and tool-use.

Integrating GenAI into Quad Process Models of Cognition

Where Clark and Chalmers have suggested the extended mind hypothesis, others have built upon Kahneman's System One and Two dual-processing cognition theory. System One is essentially subconscious processing, fast, automatic, where cognitive heuristics, sensory feedback, and information pre-sorting occur, without very much deliberate, rational thought. System Two, activated by attention, is slow, deliberate thinking which costs both more time and energy, but through which rational faculties can be used for more distinct cognition and metacognition (Yeung & Summerfield, 2012; Qiu et al., 2017).

This dual-process theory has been co-opted and expanded upon, though the efficacy and repeatability of Kahneman's work is often called into question; the complexity of

cognition in real life is far more complicated than a clean System One / System Two binary (Shea & Frith, 2016). Rather, these tend to describe perceptions around types of cognition rather than describing the mechanisms of cognition itself, a stark contrast to Global Neuronal Workspace Theory, Integrated Information Theory, or other neurocentric hypotheses around how our brains process, store, and modify information computationally (Wang et al., 2024).

System Zero has been proposed as different things by different researchers, occupying the same general space, either as a precursor or foundational layer sitting beneath System One and Two, both in robotics and in AI and human cognition research (Chiriatti et al., 2024; Chiriatti et al., 2025; Shea & Frith, 2016). It must be noted that System Zero, depending on who is proposing it, can mean very different things. In the AI space, it has specifically been hypothesized as an information preprocessing layer: learning algorithms go through mass batches of data and present to the user a set of options, decision points, and conclusions from which the user then operates in System One and Two thinking. It has also been proposed as a pre-cognitive layer primarily based around sensory input prior to subconscious cognition or deliberate processing.

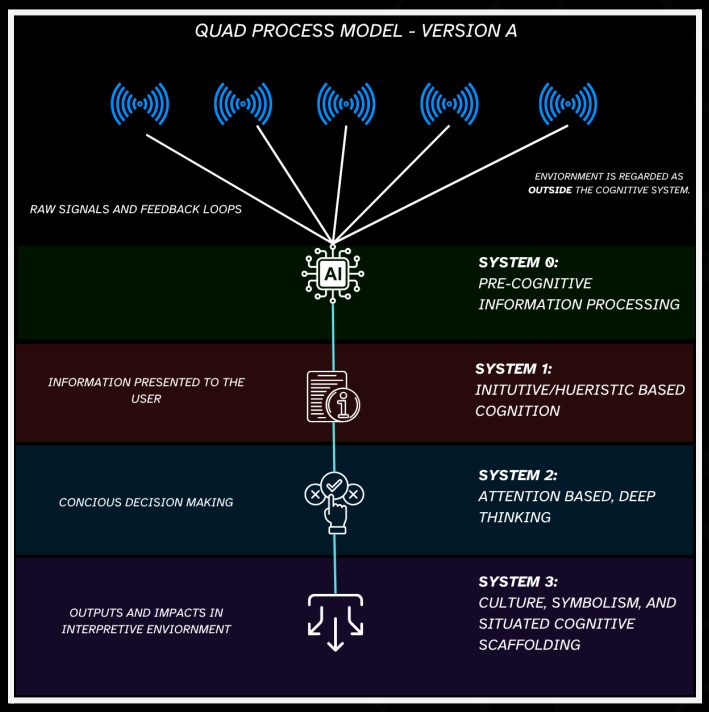
Both of these definitions make more sense when considered alongside System Three, a proposed layer of symbols, heuristics, and culture which forms from the confluence of System One and Two thinking, while also dynamically informing future System 1 and System 2 thinking. (Taniguchi et al., 2025). In some models, a version of System Three (though not called System 3 explicitly) is specifically about the orchestration of AI agents and the scaffolding around which cognition takes place (Borghoff et al., 2025). This also fits firmly within the boundaries of “classic” AI’s use of symbolic reasoning to create deterministic systems.

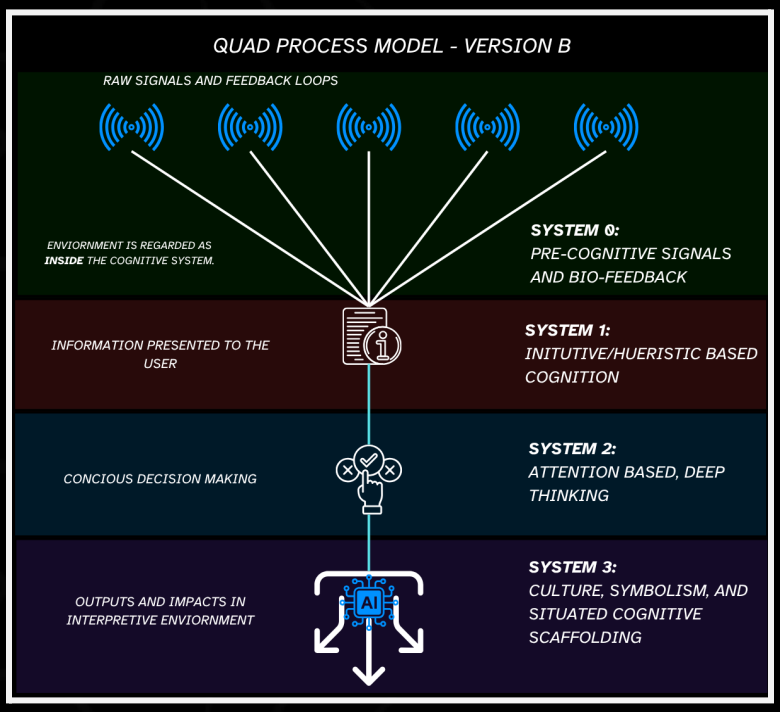
Let us propose two synthesized versions of a quad-aspect processing model, incorporating Taniguchi’s Layer 3 definition, and Chiriatti’s System 0 approach; both of which incorporates Kahneman's work as a heuristic for contextualizing where cognition and data processing is taking place.

Model A: AI sits in System Zero. Data streams are given to the LLM, which parses, contextualizes, and normalizes that data before presenting it back to the user, from which they can make decisions not very informed by high-order cognitive processes, or can choose to divert attention if something anomalous appears. These outputs will produce further symbolic understanding within both the user's mind and the cultural and

conceptual lens from which they operate (Chiriatti et al., 2024; Chiriatti et al., 2025). In this model, the environment and it's signals are regarded as *external* to the cognitive system. This model implicitly denies the distributed nature of cognition as being situated in an environment.

Model B: The LLM is a tool built on top of System Three, in which System Zero is all of our situated and situational sensory and data input within our environment (Taniguchi et al., 2025). This model still enables the LLM as data processing and signals processing, but it does so without confusing the epistemic layers, in this model, the situated environment is pre-supposed to be a part of the cognitive system.





Claim One: Language as Abstraction

Generative models are not based in reality as it is, but in our human understanding and descriptions of those things, making them two to three steps further removed from actual reality. While they may capture the neurocognitive symbolism that language inherently carries, because language itself is a neuro-symbolic representation of abstract concepts and objective descriptions about reality (Zhu & Su, 2023), language does not actually describe the thing itself. It is not an experience of the thing itself. It is a description from which we form some sort of representation. It is like the description of how rain feels on the skin, rather than the actual biofeedback of feeling rain on your

skin. A generative system will always sit two to three layers of abstraction above the sensory experience of base reality.

A classic coding example used for Machine Learning examples is as follows: You want to build an algorithm to classify what kind of fruit something is. You have examples, in the form of pictures, of various fruits, and you train a model using a convoluted neural network (CNN) to create a model that can identify pieces of fruit. In this example, you have several layers of abstraction. 1. The physical characteristics of the fruit. 2. Capturing/recorded data from the image of the fruit. 3. Embedding layers which associate mathematical representations of elements of the fruit with the fruit. 4. Classifier output.

Generative large language models take this even further: there is a raw phenomenon, a human observes it, language descriptors are created to describe/categorize it, then training data is created with data fit to those categories, then a user's natural language query is transformed into tokens to be mathematically compared, then the model makes its best prediction using the categories that are locked in parameters. Each token of the input is a “pointer”, directionally telling the algorithm to move to the most likely prediction. At least four layers of abstraction above physical reality, and one that is reliant on pre-existing categories of abstraction (Queloz, 2025; Shiffrin & Mitchell, 2023). Additionally, if the meaning of the “pointer” used by the human user and the linkages in the algorithm are misaligned, then instantaneous meaning loss occurs, and the prediction becomes less accurate.

Each layer of abstraction is a layer of interpretive commitment, a point at which human categorical judgment was frozen into the algorithm. By the time a user receives an output, they are not encountering reality or even a representation of reality, but the residue of prior human representations, mathematically recombined (Queloz, 2025). No embodied cognitive pathway is built from this kind of encounter.

The effects of prior human categorical judgments and abstractions being placed in between the user and raw signals, introduces an auto-regressive tendency in systems in which novel signals are force fit into existing categories, and thus removed from the user’s attention. This fitting of the data to existing categories, even when novel, results in broader or inaccurate categories over time. This can be observed inside the concept of semantic drift, where a word’s meaning shifts over time. As the semantic range of a word expands, its utility for precise data categorization diminishes. (de Sa et al., 2024)

In LLM-as-system-0 arrangements, this semantic drift is often hidden from human users, who do not update their internal representations to match. As categories broaden and precision decreases, outputs converge toward statistically average responses, which is exactly the regression dynamic that erodes institutional expertise over time.

Thus we cannot expect the same type of bio-cognitive pathways to be developed when engaging with these systems as would arise from an actual experience of data in the real world. By no stretch of the imagination are we saying that legacy systems of automation or more deterministic algorithms create some sort of inherent sensory feedback which enhances human embodied cognition. Rather, the implication of our argument is that the pervasiveness and institutional momentum being generated around GenAI is detrimental to a multiplicity of human systems on a scale rarely touched by novel technology.

Claim Two: AI as Pre-Processing Reduces Sensory Feedback Necessary for Cognition & Skill Acquisition

Based on the abstraction issue discussed in Claim One, a subsequent problem that emerges in GenAI integration is the treating of GenAI as extended cognition in the absence of any meaningful sensory feedback loop between the situated environment and the user.

The abstraction layer inherent in generative systems means that, when AI is treated as QPM-A System 0, it stands in between environmental signals and the user whose cognition you are trying to extend. While there are similar issues in automation with over reliance, the key differentiator is that in most dashboards, gauges, and pipelines are summarizations taken deterministically from sensors, databases, etc. An example would be trying to determine the temperature outside with a custom application and sensor set. You decide to build two applications and compare them. Application 1 links to some temperature gauges distributed across your property through Internet of Things (IOT) devices, then a function averages the temperature readings and returns the result to the user. Application 2 uses an LLM with access to the same sensor data. The user queries

the LLM and the LLM uses the sensor input for its token prediction. It then returns the predicted token to the user to answer the question. Instead of being a direct reading, it is a prediction of the correct average across sensors, instead of the reading. While most of the time, the answer may be similar, the LLM has a distinct mechanism for removing the outlier. If one sensor is melting, application 1 would show a spike (a signal). An LLM might "hallucinate" that sensor into the average because the spike is statistically improbable, thereby obfuscating the situated reality the user needs to see. This effectively cuts off the user from their situated environment, instead of extending their ability to read, understand, and interact with it.[1]

This is not merely theorizing about potential degradation in skill acquisition, there are now multiple studies, and observational data about skill acquisition when GenAI is introduced. Shen demonstrates that novice users and entry level employees suffer from complete lack of skill acquisition when given GenAI to complete a new task. This applied to understanding of the system (software libraries in this case), the ability to complete the task without AI, and troubleshooting the system. (Shen, et. all, 2026) However, there remained patterns of usage which showed improvement. Namely, patterns that involved an increase in explanation or exploration of the code instead of additional offloading.

Similar results are seen in studies on misinformation identification ability and the use of AI assistants to augment and teach identification of misinformation online. Participants' immediate capability grew with the help of the AI assistant but long-term, their unassisted ability fell over 15% below the baseline capability at the start of the study. (Rani, et. all, 2025)

All of this is not to say that GenAI augmented learning patterns have not been observed. (Tariq, et al, 2025) showed that in increasingly complex cybersecurity situations (phishing vs intrusion) human-AI collaboration produced better results than independent execution or LLM-only execution. A general improvement trend between iterations in accuracy and recall, regardless of whether a user started work independently or started with GenAI collaboration, performance improved over time. Indicating that there are tangible learning outcomes that can be attributed to collaboration.

[1] LLMs can utilize tools, which are snippets of code that can execute deterministic protocols and return the result, either for passthrough prediction or as a standalone return. However, this is environmental and system design and represents a foundationally different approach than the one we are critiquing.

One might attempt to simulate some sort of sensory feedback through various UI design choices or toolification. However, the issue is that these are not deterministic measures based on physics, but rather token prediction, which is notorious for being finicky. Adversarial researchers have demonstrated repeatedly that any transformer-based model is inherently vulnerable to a variety of poisoning methods, most of which are very simple to execute. (OWASPLLMProject Admin, 2024) This takes no computing power whatsoever and can often be accomplished in a matter of minutes, completely breaking the token prediction model of these transformer-based, stochastic models and token prediction algorithms. (Cox & Bunzel, 2025)

This point will be revisited after discussing some proposed models for AI interaction, but suffice to say, if any transformer based model can be defeated, poisoned, or manipulated in a near-infinite attack space, then building systems that rely on transformer models as cognitive pre-processing for all upstream tasks introduces a myriad of vulnerabilities to any system cognitive framework and creates a central, cascading failure point for organizational decision making. Neural network systems underlying large language models exhibit high sensitivity to perturbation, where small changes in parameters or inputs can produce disproportionately large output variations (Sun et al., 2020; Wang et al., 2023). Prior work in complex learning systems demonstrates that such perturbations can propagate through model structure, generating cascading effects that alter outputs globally rather than locally (Oh et al., 2022). This aligns with broader findings that tightly coupled parameter systems exhibit cascading risk dynamics under uncertainty (Pandey & Motee, 2025). Succinctly, small shifts in inputs can have outsized cascading effects in GenAI outputs.

Sensory feedback is necessary in traditional System 0 pre-processing loops because they are grounded in deterministic physical reality. An LLM's outputs, by contrast, can silently shift without the user having any sensory signal or flag that something has changed. The user is also dulling their own ability to develop external patterns recognition and sensitivities that integrate other external patterns and signals. Thus even if a user tries to emulate a sensory feedback coupling onto an LLM interface, linking signal and user, it would be attempting to build feedback through a medium which is inherently a fabrication of a plausible state of the system environment. This system can be invisibly destabilized, thus creating a near-infinite number of ways the signal could be corrupted, and the user would have no inherent way to know if the signal was true or a fabrication.

This does not defeat the idea of cognition being extended through AI systems wholesale. However, we must distinguish extended systems of cognition from embodied systems of cognition. A system may be both embodied and extended; but when the extension is purely artificial and does not result in an actual capability gain in the person via an embodied mechanism, what is accumulated instead is cognitive debt, a term proposed by researchers in their work on large language models and how they degrade human performance and capability over time (Jiao et al., 2023; Gibson et al., 2023).

The core principle is this: extension plus embodiment builds capability. Extension without embodiment degrades capability over time. LLMs, in their current form, are pure extension, and the absence of any embodied coupling is not a design oversight but a structural feature of the medium itself (Savage, 2025).

Claim Three: Lack of Skill Acquisition Accelerates Regression to the Mean.

Distributed Cognition and the Limits of AI Integration

When Edwin Hutchins wrote *Cognition in the Wild* (Hutchins, 1995), the book quickly became highly regarded as a classic of cognitive ethnography, formalizing the concept of distributed cognition, which views cognition as an emergent property of a system involving both humans and their tools. It bridges the gap between anthropology and cognitive science, setting a precedent for the study of team dynamics, culture, and distributed cognitive systems. The central assertion is that cognition is not merely limited to a computational process and a manipulation of internal symbols of a single person or organism alone, but a perception and action loop which is continually iterating in a specific enactive and situated environment, specifically through sensory-motor engagement with the world, in which skills are brought forth and enacted, not retrieved like some sort of cognitively stored file in the brain which is referenced.

In fact, distributed cognition, embodied cognition, and extended cognition are all really facets of the same stream of thinking, in which cognition, and potentially consciousness, is distributed, not centralized inside the wet room of the brain. Extension is essentially talking about how the body and the brain engage with things that are physically external

to the brain and incorporate them into cognition. Situated or distributed cognition is how a single node of cognition, in this case a human mind, is situated within its proper environment and engages with both the tools and the other nodes of cognition in that given system and environment (Hutchins, 1995). In totality, we get a very different view on skill acquisition, in which the components of skills are so intimately tied to environments that they become in and of themselves emergent properties based on the affordances, the ability to act, in a specific given environment. Removed from the environment, the nodes required for the skills are very often less transferable than one might expect.

Across Clark's extended cognition and Hutchins' distributed cognition, evidence converges on the idea that cognition and skill can be dynamically enacted through continuous interaction between the brain, body, and environment, rather than stored as fixed internal representations or reference points (Hutchins, 1995). The critical role of prospective cognition, mentally simulating future scenarios to make inter-temporal trade-offs, further illustrates this dynamic quality; such deliberation is inherently constructive and subject to systematic biases in ways that AI-generated outputs cannot replicate or adequately support (Bulley & Schacter, 2020). These patterns, skills, are able to be enacted within a specific environment and are always being finely tuned dynamically based on the feedback of the system between various nodes. It is a dynamic perception and action loop that extends beyond merely the person but into its entire environment.

A concrete example of this would be Esther Thelen's dynamic systems motor development theory, specifically looking at how infants develop walking as an emergent property of biofeedback between muscle strength, balance, and environmental support, not as a pre-existing walking capability, but rather as a self-organized reinforcement learning system. Contrast this with large language models in and of themselves, which have no capability for self-improvement. In fact, systems that have been experimented with that do have the ability to dynamically change their internal weights often suffer absolutely catastrophic capability loss once a certain percentage of changes have occurred (Soori et al., 2023; Papadopoulos et al., 2020).

Extending this to human and LLM interactions: when a task is offloaded to an LLM and there is no dynamic feedback involved in the process at all, no intentional friction that can serve as an indicator for learning, evaluation, and incorporation of feedback for the

situated task at hand, then you have functionally removed all ability of the people and the total system from learning and dynamically tuning and improving process as an emergent property of environment, cognition, and sensory feedback loops (Gibson et al., 2023; Jiao et al., 2023).

The difference between the blacksmith hammering metal, indigenous peoples' deep connection to their ancestral lands, which enables them to read sensory feedback from the environment, whether it be the migration patterns of birds, different atmospheric pressure changes, predictions of coming weather or seasons, or where certain plants or animals will dwell, almost in a subconscious manner due to their deep connection with their situated environment; is that all of these examples rely on physics and bio-feedback systems which are more deeply integrated than abstracted. While these constitute complex systems with their own feedback loops, soft assemblies, and system information flow dynamics, they are much more situated in the real world, where prediction can happen in a quasi-Bayesian way (Bossaerts & Murawski, 2017). While sometimes those predictions are incorrect, they are based more in foundational reality as it is, as opposed to our constructed reality that is not.

This can of course be useful to reduce cognitive load when managing things that no single node, i.e., person, or no conglomerate of systems could manage without either radically increasing attention, resourcing, or manpower to the task, or utilizing machine learning or AI tools. However, it should be noted that utilizing AI as a preprocessing task or an offloading task that is foundational or upstream of decision making and cognitive tasks places everything downstream of that implementation at risk of skill degradation and atrophy, due to the limitation of feedback mechanisms that enable the human decision maker, judgment provider, or operator the ability to develop intuitive judgments and improve their skills in the manner necessary for expertise and expert performance (Chen et al., 2023; Vincent, 2021; Janssen et al., 2020).

We may contrast this with System Three implementation, in which AI is treated as an artifact of existing human cognitive systems, a symbolic representation of knowledge and skills that we would like to continue to do at greater scale and speed, as opposed to a precondition for understanding and interpreting the world (Taniguchi et al., 2025; Zhu & Su, 2023).

Practical Design Patterns: Where in a Quad-Process System Cognition Model should GenAI sit?

Reference back to Model A of the Quad-Process (QPM) of System Cognition, wherein AI is regarded as foundational signals pre-processing (Chiriatti et al., 2024; Chiriatti et al., 2025).

The issue with Model A (QPM-A) should be immediately apparent. Generative AI and large language models are based on our own socio-cultural and linguistic understanding, which is highly representational and symbolic at multiple layers of abstraction (Queloz, 2025). To assert that System Zero is the foundational preprocessing layer, roughly equivalent to pre-sensory or pre-cognitive sensory input, is to assert that our layers of abstraction have superseded raw sensory input. This type of treatment creates an infinite feedback loop without any mediating mechanism. All new data, when processed in this manner, will be normalized into existing symbolic categories, leaving very little room for the addition of new cognitive pathways in the given system. Thus the overall systemic intelligence level will become lower over time, and regression to the mean will become more aggressive and more pronounced as the feedback loop normalizes more data into second and third order abstractions, narrowing the variety from which a user may understand and select from their situated environment.

By treating AI as a pre-filter, we transform catastrophic forgetting, over-normalization, and data overfitting into an epistemic crisis extending not just to quantitative data sets but to qualitative understanding of reality. This design model, left unchecked, is what leads to what we might call AI-based epistemic collapse. We have essentially offloaded our ability to preprocess information and read new signals in our cognitive environment, giving that task to a system whose job it is to make data fit into categories it is already trained on, then predict what comes next (Jiao et al., 2023; Langer et al., 2021).

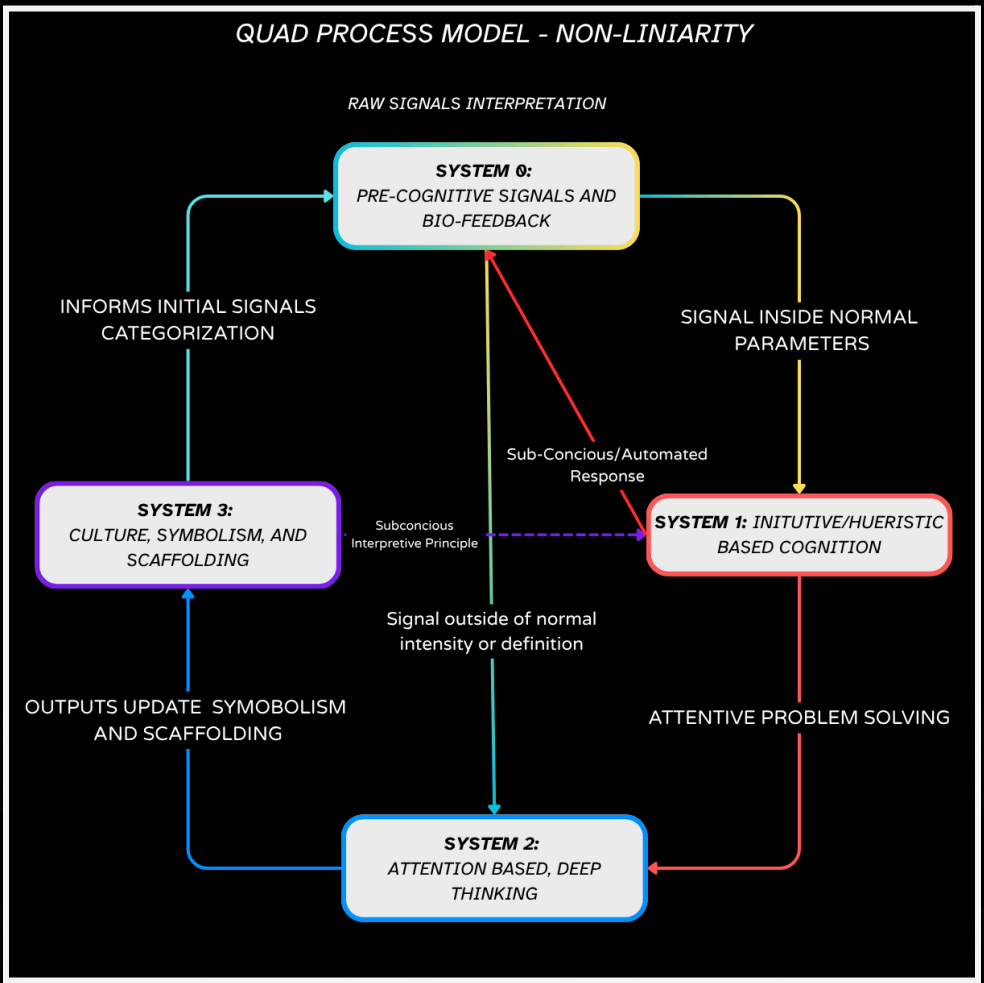
QPM-B, which treats AI as symbolic and systems based artifacts of system cognition, is not without its own issues. However, as a design principle, treating AI as an extension of System Three, situated symbolic and heuristic cognition encompassing culture, language, and similar functions, does not automatically fall into the same auto-regressive tendency (Taniguchi et al., 2025; Sridharan, 2023). This is because we

foundationally treat System Zero not as a layer of abstraction, but rather open it up to exploration of harder sciences and the development of better machine learning tools tied more directly to observation of exact physical phenomena: EMF signals, pressure changes, inertia of motion, and so on.

These models still fail insofar as they are not expressly self-referential, concurrent, and self-reinforcing. They lend themselves to hierarchical and linear-based cognitive models, which reality does not bear out. Calling them "systems" as opposed to "layers" or "stages" is a notable improvement; however, they are often discussed linearly. The reality is that System Three does impact our analysis of System Zero, that raw cognitive signals and sensory feedback are themselves already shaped by cultural and symbolic scaffolding, and vice versa.

Together, this creates a complex feedback system in which constant evolution and learning takes place. The various systems work together to produce a self-reinforcing system that is much closer to the accepted “Default Mode Network” in cognition studies, but is still a useful abstraction for system design purposes.

Gen AI As Facade of Extended Cognition, Winkler, 2026.



Otto's Notebook Revisited

One may be tempted to critique my use of Otto's Notebook as a similar abstraction layer to GenAI. I agree with this critique, so long as we are saying that Otto's notebook is a functional output that sits within System 3. Otto's notebook does not function as an information pre-processing layer. It does not sit between sensory feedback and Otto's body, rather, it is the product of Otto's cognition, being symbolically extended into a tool that can represent already completed cognitive processes. Referencing Figure 3 above, we can see how Otto's notebook is an interpretive heuristic which enables Otto's System 1 to quickly make decisions in light of a degraded internal cognitive system. If Otto's notebook were treated as System 0; for example, that anything written in the notebook supplants another input: say his daughter coming to pick him up, then the degradative qualities become apparent.

Any cognitive tool becomes degradative when it is positioned to intercept environmental signals before they reach the human operator, regardless of what layer it originated in. This is foundationally, extension without embodiment. The argument is additionally not that System 3 cannot introduce failures or errors into systems, only that System 3 represents a current state of abstraction and understanding, which is produced by more grounded cognitive processes underneath of it. A mistake in System 3 has more opportunities to be corrected by counter-factual or conflicting sensory feedback, whereas abstraction at System 0 bypasses other situated cognition layers which may serve a corrective, self-updating function over time. The difference between a single tool being used for information pre-processing, and GenAI implementations, is that GenAI is being purported to replace entire functions traditionally done by a human, end-to-end. Encompassing a multiplicity of situated skills, functions, and feedback loops.

Conclusion

The argument made in this paper is not that generative artificial intelligence is without value. Rather, the manner in which it is often conceptualized and subsequently deployed, as a foundational preprocessing layer, a cognitive surrogate, and a replacement for situated human judgment, is producing systemic fragility that will compound over time and is already measurable in empirical data.

The three claims made here form a causal chain, not a list of parallel concerns.

Generative language models are structurally abstracted from physical reality by at least four layers of interpretive commitment, each of which represents a point at which human categorical judgment was frozen into a parameter. This is not a solvable engineering problem, it is the nature of the medium. Because of this abstraction, no meaningful sensory feedback loop can be built between a generative system and a situated human operator. The outputs are stochastic, non-deterministic, and silently poisonable in ways that produce no perceptible signal to the user. What gets built in place of a genuine feedback loop is a fabrication of one, a plausible-seeming return that the user has no intrinsic mechanism to verify or calibrate against.

Because no genuine feedback loop exists, the skill acquisition cycles that depend on perception-action loops, dynamic environmental tuning, and embodied pattern reinforcement are bypassed entirely. The result, as the empirical literature is beginning

to confirm, is cognitive debt: measurable degradation in unassisted human performance in precisely the domains where AI assistance is most heavily deployed.

Left unaddressed, this dynamic will accelerate. As generative systems are positioned upstream of human cognition, as System Zero preprocessors rather than System Three artifacts, they introduce an auto-regressive tendency in which novel signals are force-fit into existing categories, semantic precision degrades, and the informational variety from which genuine expertise and anomaly detection emerge is systematically narrowed.

The institutional stakes of this are significant. Organizations that deploy GenAI as a cognitive foundation rather than a cognitive scaffold are not merely risking individual skill atrophy. They are building brittle decision architectures in which the failure modes are invisible until they cascade destructively.

None of this is inevitable. When GenAI is positioned as a System Three artifact, a symbolic and heuristic scaffolding built on top of human cognition and genuine environmental signal, rather than a System Zero interpolation, it can do what it actually does well: organizing existing knowledge, checking cross-domain plausibility, surfacing anomalies for human attention, and extending the reach of deliberate System Two cognition without replacing the embodied signal processing that makes expert judgment possible within a situated environment.

We must treat the resilience and improvement of human cognitive capability not as a secondary consideration to efficiency, but as the precondition for any system that needs to remain capable, resilient, and adaptable over time.

Recommendations on Principles for GenAI Implementation

From a system design perspective, we propose the following principles to create true extension by encompassing feedback loops inside of a given environment, rather than a generative facade of extension.

Avoidances

1. Be wary of pairing non-experts with AI for domain specific tasks that require collaboration and a high degree of judgement, due to lack of skill acquisition and psychosis risk (Awa et al., 2026; Matcheri Keshavan et al., 2026; Chen et al., 2023).
2. Avoid attaching deterministic signals based feedback to generative outputs to emulate situated feedback, due to abstraction risk and poisonability.

Design Patterns

AI in Systems 3 and 2

3. AI can be used with structured scaffolding as a feedback loop from System 3 to System 2, deliberate cognition as representations of the current state of knowledge, enabling better decision making (Walker et al., 2025; Toy et al., 2024). (With the caveat of information supply chain analysis to safeguard signals and data input to the decision making process)
4. Use AI as a scaffolding to check plausibility when working in cross-domain solutions or systems.
5. Only use GenAI directly for abstraction level tasks requiring natural language related to that task. Not raw signal and data pre-processing.

AI in Systems 0 and 1

6. Treat GenAI as an attention flag when processing signals. Directing attention to anomalous or non fitting signals within the system. Even better when cross-domain abstraction can link anomalous signals that networks of deterministic algorithms might miss (Valiente & Pilly, 2024; Lin et al., 2023).
7. To support principle 6, utilize GenAI to extract and create ontology from representative data sets which can be used to enhance signals into the system while not constraining the system through data overfitting (Zhu et al., 2024).

AI in UI/UX

8. Integrate progressive-disclosure principles in low-friction interfaces to enable thoughtful, intentional exploration of signals and feedback when required via anomaly detection or flagging.

9. Whenever possible, tie GenAI outputs to confidence levels, data provenance, or inference uncertainty that is visible to the operator rather than presenting outputs as clean conclusions (Janssen et al., 2020; Langer et al., 2021).

Governance Principles

10. Have decision auditing capability with data/signal supply chain analysis built into telemetry readings on institutional AI use.
11. Implement immutable identity validation for GenAI & human operators within a given operational system to augment decision supply chain analysis. This principle is one piece of a decision defense-in-depth approach to the cognitive security of a system or organization implementing GenAI, and augments principle 10 for end-to-end data and outcome validity.
12. If offloading tasks fully to automation or AI. Limit human interaction and exposure to the AI process, instead, reinforce, train, and record the manual process to avoid knowledge loss and cognitive debt.
13. Establish baseline human performance benchmarks in domain-critical tasks prior to AI deployment and conduct periodic audits to detect skill atrophy or cognitive debt accumulation at the institutional level (Walker et al., 2025).

Bibliography

Awa, J., Melika Sarkandi, Brown, K., & Choudhury, Q. (2026). AI Psychosis: A Scoping Review of Emerging Psychiatric Outcomes Linked to AI Chatbot Use. *Research Week*. https://doi.org/10.82451/k5882

Baber, C., Parekh, M., & Cengiz, T. G. (2014). Tool use as distributed cognition: how tools help, hinder and define manual skill. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00116

Basu, S., & Dixit, S. (2022). Role of metacognition in explaining decision-making styles: A study of knowledge about cognition and regulation of cognition. *Personality and Individual Differences*. https://doi.org/10.1016/j.paid.2021.111318

Bayne, T., Brainard, D., Byrne, R. W., Chittka, L., Clayton, N., Heyes, C., Mather, J., Ölveczky, B., Shadlen, M., Suddendorf, T., & Webb, B. (2019). What is cognition? *Current Biology*, *29*(13), 608–615. https://doi.org/10.1016/j.cub.2019.05.044

Borghoff, U. M., Bottoni, P., & R. Pareschi. (2025). Human-artificial interaction in the age of agentic AI: A system-theoretical approach. *ArXiv*, *abs/2502.14000*. https://doi.org/10.3389/fhumd.2025.1579166

Brooks, R. (1991). New approaches to robotics. *Science*, *253*, 1227–1232. https://doi.org/10.1126/science.253.5025.1227

Bulley, A., & Schacter, D. (2020). Deliberating trade-offs with the future. *Nature Human Behaviour*, *4*, 238–247. https://doi.org/10.1038/s41562-020-0834-9

Cao, Y., Xu, B., Li, B., & Fu, H. (2024). Advanced design of soft robots with artificial intelligence. *Nano-Micro Letters*, *16*. https://doi.org/10.1007/s40820-024-01423-3

Che, Y., Okamura, A., & Sadigh, D. (2018). Efficient and trustworthy social navigation via explicit and implicit Robot–Human communication. *IEEE Transactions on Robotics*, *36*, 692–707. https://doi.org/10.1109/tro.2020.2964824

Checking your browser - reCAPTCHA. (2024). Nih.gov. https://pubmed.ncbi.nlm.nih.gov/40378006/

Chen, V., Liao, Q., Vaughan, J. W., & Bansal, G. (2023). Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction*, *7*, 1–32. https://doi.org/10.1145/3610219

Cox, D. S., & Bunzel, N. (2025). *Quantifying the Risk of Transferred Black Box Attacks*. ArXiv.org. https://arxiv.org/abs/2511.05102

de, Silveira, D., & Pruski, C. (2024). *Survey in Characterization of Semantic Change*. ArXiv.org. https://arxiv.org/abs/2402.19088

Gibson, D., V. Kovanović, Ifenthaler, D., Dexter, S., & Feng, S. (2023). Learning theories for artificial intelligence promoting learning processes. *Br. J. Educ. Technol.*, *54*, 1125–1146. https://doi.org/10.1111/bjet.13341

Guingrich, R. E., & Michael. (2025). *A Longitudinal Randomized Control Study of Companion Chatbot Use: Anthropomorphism and Its Mediating Role on Social Impacts*. ArXiv.org. https://arxiv.org/abs/2509.19515

Gulati, A., Soni, S., & Rao, S. (2020). *Interleaving fast and slow decision making*. 1535–1541. https://doi.org/10.1109/icra48506.2021.9561562

Hutchins, E. (1995). *Cognition in the wild*. Mit Press.

Janssen, M., Hartog, M. D., Matheus, R., Ding, A. Y., & Kuk, G. (2020). Will algorithms blind people? The effect of explainable AI and decision-makers' experience on AI-supported decision-making in government. *Social Science Computer Review*, *40*, 478–493. https://doi.org/10.1177/0894439320980118

Jiao, D., RobledoCastro, C., CastilloOssa, L., Corchado, J., Shanmugasundaram, M., Tamilarasu, A., Pergantis, P., Bamicha, V., Skianis, C., Drigas, A., Haider, Z., Zummer, A., Waheed, A., Abuzar, M., Chen, C., Huang, N., Hu, B., Zhang, M., Yuan, J., & Guo, J. (2023). ChatGPT: The cognitive effects on learning and memory. *Brain-X*, 10456646532399--2427----683---674--------1050--. https://doi.org/10.1002/brx2.30

Keller, C. M., & Janet Dixon Keller. (1996). *Cognition and tool use : the blacksmith at work*. Cambridge University Press.

Langer, M., König, C. J., Back, C., & Hemsing, V. (2021). Trust in artificial intelligence: Comparing trust processes between human and automated trustees in light of unfair bias. *Journal of Business and Psychology*, *38*, 493–508. https://doi.org/10.1007/s10869-022-09829-9

Li, Z.-Z., Zhang, D., Zhang, M.-L., Zhang, J., Liu, Z., Yao, Y., Xu, H., Zheng, J., Wang, P.-J., Chen, X., Zhang, Y., Yin, F., Dong, J., Guo, Z., Song, L., & Liu, C.-L. (2025). From system 1 to system 2: A survey of reasoning large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PP*. https://doi.org/10.48550/arxiv.2502.17419

Lin, B. Y., Fu, Y., Yang, K., Prithviraj Ammanabrolu, Faeze Brahman, Huang, S., Bhagavatula, C., Choi, Y., & Ren, X. (2023). SwiftSage: A generative agent with fast and slow thinking for complex interactive tasks. *ArXiv*, *abs/2305.17390*. https://doi.org/10.48550/arxiv.2305.17390

Massimo Chiriatti, Ganapini, M. B., Enrico Panai, Wiederhold, B., & Riva, G. (2025). System 0: Transforming artificial intelligence into a cognitive extension. *Cyberpsychology, Behavior, and Social Networking*, *28*, 534–542. https://doi.org/10.1089/cyber.2025.0201

Massimo Chiriatti, M. Ganapini, Enrico Panai, Ubiali, M., & Riva, G. (2024). The case for human–AI interaction as system 0 thinking. *Nature Human Behaviour*, *8*, 1829–1830. https://doi.org/10.1038/s41562-024-01995-5

Matcheri Keshavan, Torous, J., & Yassin, W. (2026). Do generative AI chatbots increase psychosis risk? *World Psychiatry*, *25*(1), 150–151. https://doi.org/10.1002/wps.70017

Matthieu Queloz. (2025). Explainability through systematicity: The hard systematicity challenge for artificial intelligence. *Minds and Machines*, *35*. https://doi.org/10.1007/s11023-025-09738-9

Menary, R. (2012). *The extended mind*. Mit Press.

Menon, V. (2023). 20 years of the default mode network: A review and synthesis. Neuron, 111(16), 2469–2487. https://doi.org/10.1016/j.neuron.2023.04.023

Mohsen Soori, B. Arezoo, & Roza Dastres. (2023). Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cognitive Robotics*. https://doi.org/10.1016/j.cogr.2023.04.001

Oh, S., Ustun, B., McAuley, J., & Kumar, S. (2022). Rank List Sensitivity of Recommender Systems to Interaction Perturbations. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. https://doi.org/10.1145/3511808.3557425

OWASPLLMProject Admin. (2024, November 19). *OWASP Top 10 for LLM Applications 2025 - OWASP Top 10 for LLM & Generative AI Security*. OWASP Top 10 for LLM & Generative AI Security. https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

P. Bossaerts, & Murawski, C. (2017). Computational complexity and human decision-making. *Trends in Cognitive Sciences*, *21*, 917–929. https://doi.org/10.1016/j.tics.2017.09.005

Pandey, V., & Motee, N. (2025). *Distributionally Robust Cascading Risk in Multi-Agent Rendezvous: Extended Analysis of Parameter-Induced Ambiguity*. ArXiv.org. https://arxiv.org/abs/2511.20914

Papadopoulos, G., Antona, M., & C. Stephanidis. (2020). Towards open and expandable cognitive AI architectures for large-scale multi-agent human-robot collaborative learning. *IEEE Access*, *9*, 73890–73909. https://doi.org/10.1109/access.2021.3080517

Pherson, R. H., Donner, O., & Gnad, O. (2024). Understanding How We Think: System 1 and System 2. *Springer Nature*, 5–19. https://doi.org/10.1007/978-3-031-48766-8_2

Qiu, L., Su, J., Ni, Y., Bai, Y., Li, X., & Wan, X. (2017). The neural system of metacognition accompanying decision-making in the prefrontal cortex. *PLoS Biology*, *16*. https://doi.org/10.1371/journal.pbio.2004037

Savage, N. (2025). Can AI expand the human mind? *Communications of the ACM*, *68*, 12–14. https://doi.org/10.1145/3748644

Séverin Lemaignan, Mathieu Warnier, Sisbot, E. A., A. Clodic, & Alami, R. (2017). Artificial cognition for social human-robot interaction: An implementation. *Artif. Intell.*, *247*, 45–69. https://doi.org/10.1016/j.artint.2016.07.002

Shea, N., & Frith, C. (2016). Dual-process theories and consciousness: the case for "Type Zero" cognition. *Neuroscience of Consciousness*, *2016*. https://doi.org/10.1093/nc/niw005

Shiffrin, R., & Mitchell, M. (2023). Probing the psychology of AI models. *Proceedings of the National Academy of Sciences of the United States of America*, *120*. https://doi.org/10.1073/pnas.2300963120

Sridharan, M. (2023). *A cognitive architecture for integrated robot systems*.

Sun, X., Zhang, Z., Ren, X., Luo, R., & Li, L. (2021). Exploring the Vulnerability of Deep Neural Networks: A Study of Parameter Corruption. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(13), 11648–11656. https://doi.org/10.1609/aaai.v35i13.17385

Sutton, J. (2010). *Exograms and Interdisciplinarity: History, the Extended Mind, and the Civilizing Process*. 189–225. https://doi.org/10.7551/mitpress/9780262014038.003.0009

Taniguchi, T., Hirai, Y., Suzuki, M., Murata, S., Horii, T., & Tanaka, K. (2025). System 0/1/2/3: Quad-process theory for multi-timescale embodied collective cognitive systems. *Artificial Life*, *31 4*, 465–496. https://doi.org/10.48550/arxiv.2503.06138

Tariq, S., Singh, R., Chhetri, Mohan Baruwal, Nepal, S., & Paris, C. (2025). *Bridging Expertise Gaps: The Role of LLMs in Human-AI Collaboration for Cybersecurity*. ArXiv.org. https://arxiv.org/abs/2505.03179

Toy, J., MacAdam, J., & Tabor, P. (2024). Metacognition is all you need? Using introspection in generative agents to improve goal-directed behavior. *ArXiv*, *abs/2401.10910*. https://doi.org/10.48550/arxiv.2401.10910

Valiente, R., & Pilly, P. K. (2024). Metacognition for unknown situations and environments (MUSE). *Neural Networks : The Official Journal of the International Neural Network Society*, *194*, 108131. https://doi.org/10.1016/j.neunet.2025.108131

Vincent, V. U. (2021). Integrating intuition and artificial intelligence in organizational decision-making. *Business Horizons*. https://doi.org/10.1016/j.bushor.2021.02.008

Walker, P. B., Haase, J. J., Mehalick, M. L., Steele, C. T., Russell, D. W., & Davidson, I. N. (2025). Harnessing metacognition for safe and responsible AI. *Technologies*. https://doi.org/10.3390/technologies13030107

Wang, G., Bao, H., Liu, Q., Zhou, T., Wu, S., Huang, T., Yu, Z., Lu, C., Gong, Y., Zhang, Z., & He, S. (2024). Brain-inspired artificial intelligence research: A review. *Science China Technological Sciences*, *67*, 2282–2296. https://doi.org/10.1007/s11431-024-2732-9

Wang, H., Ma, G., Yu, C., Gui, N., Zhang, L., Huang, Z., Ma, S., Chang, Y., Zhang, S., Shen, L., Wang, X., Zhao, P., & Tao, D. (2023). *Are Large Language Models Really Robust to Word-Level Perturbations?* ArXiv.org. https://arxiv.org/abs/2309.11166

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 1310–1321. https://doi.org/10.1098/rstb.2011.0416

Zhu, J., & Su, H. (2023). Toward the third generation artificial intelligence. *Science China Information Sciences*, *66*, 1–19. https://doi.org/10.1007/s11432-021-3449-x

Zhu, M., Zhu, Y., Li, J., Wen, J., Xu, Z., Che, Z., Shen, C., Peng, Y., Liu, D., Feng, F., & Tang, J. (2024). *Language-conditioned robotic manipulation with fast and slow thinking*. 4333–4339. https://doi.org/10.1109/icra57147.2024.10611525