frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 01 December 2025
DOI 10.3389/frai.2025.1677528

Check for updates

OPEN ACCESS

EDITED BY
Ramesh Chandra Poonia,
Christ University, India

REVIEWED BY
Surapati Pramanik,
Nandalal Ghosh B.T. College, India
Francisco Rafael Trejo-Macotela,
Universidad Politécnica de Pachuca, Mexico

*CORRESPONDENCE
Abdelaali Mahrouk
✉ abd.marok25@gmail.com

RECEIVED 01 August 2025
ACCEPTED 08 October 2025
PUBLISHED 01 December 2025

CITATION
Mahrouk A (2025) Epistemic limits of local interpretability in self-modulating cognitive architectures.
Front. Artif. Intell. 8:1677528.
doi: 10.3389/frai.2025.1677528

Epistemic limits of local interpretability in self-modulating cognitive architectures

Abdelaali Mahrouk*

Université Frères Mentouri Constantine 1, Constantine, Algeria

Introduction: Local interpretability methods such as LIME and SHAP are widely used to explain model decisions. However, they rely on assumptions of local continuity that often fail in recursive, self-modulating cognitive architectures.
Methods: We analyze the limitations of local proxy models through formal reasoning, simulation experiments, and epistemological framing. We introduce constructs such as Modular Cognitive Attention (MCA), the Cognitive Leap Operator (Ψ), and the Internal Narrative Generator (ING).
Results: Our findings show that local perturbations yield divergent interpretive outcomes depending on internal cognitive states. Narrative coherence emerges from recursive policy dynamics, and traditional attribution methods fail to capture bifurcation points in decision space.
Discussion: We argue for a shift from post-hoc local approximations to embedded narrative-based interpretability. This reframing supports epistemic transparency in future AGI systems and aligns with cognitive theories of understanding.

KEYWORDS

stratified decision landscapes, salience-gated attention, cognitive leap operator, internal narrative generator, modular cognitive attention, recursive contextual memory, meta-computational narratives, narrative interpretability

1 Introduction

The recent proliferation of large language models (LLMs), self-reflective agents, and modular cognitive architectures (Ha and Schmidhuber, 2018) has transformed the landscape of artificial intelligence. These systems—capable of autonomous reasoning, recursive planning, and context-sensitive goal adaptation—now underpin applications in scientific discovery, legal reasoning, autonomous systems, and educational technologies (Bommasani et al., 2021; Bubeck et al., 2023).

Yet, as these architectures grow in complexity and autonomy, classical interpretability methods—such as gradient-based saliency maps, LIME, SHAP, and attention heatmaps—often prove inadequate. Rooted in local perturbations or linear approximations, these techniques do not fully capture distributed reasoning, inter-policy modulation, and reflective control at scale (Rudin, 2019; Geirhos et al., 2020; Ghassemi et al., 2021; Ji et al., 2023). In modular or self-reflective systems, they risk misrepresenting causality, since emergent behavior arises from multi-step interactions between internal policies rather than isolated activations or token dependencies.

This suggests a potential limitation: the prevailing assumption that local interpretability can scale to global understanding is increasingly fragile. The "local-to-global" extrapolation presumes a compositional alignment between micro-decisions (token-level attributions) and macro-behaviors (e.g., planning, reflection, theory of mind). However, in modular agents where policies dynamically interact (Langosco et al., 2023; Chan et al., 2022), such extrapolations may yield misleading narratives. We argue instead that interpretability should shift its focus from static saliency to dynamic policy flows—how internal decision policies evolve, interact, and stabilize over time.

This reframing introduces a novel interpretive ontology: rather than explaining decisions by isolated local inputs, we trace global behaviors through the evolution of internal policy states and their attractor dynamics. To support this proposal, Section 3 will introduce three new constructs—Modular Cognitive Attention (MCA: see Glossary in Section 2.6), the Cognitive Leap Operator (Ψ: see Glossary in Section 2.6), and the Internal Narrative Generator (ING: see Glossary in Section 2.6) —alongside a glossary and conceptual diagram (Figure 1). These tools aim to clarify how interpretability can be embedded within recursive, self-modulating agents.

Gap Identification (CARS model):

1 What is known: Local interpretability methods dominate the field due to their simplicity and model-agnostic design. They remain widely used for auditing, debugging, and fairness in AI systems.
2 What is unknown: These methods do not scale effectively to modular or reflective cognitive architectures. There is limited understanding of how to trace internal policy evolution across multi-policy systems.
3 Why it matters: Without scalable interpretability frameworks, safety, auditability, and societal trust in autonomous cognitive systems remain compromised.

Table 1 contrasts attribution-based interpretability with the proposed policy dynamics approach, highlighting improvements in scalability, causality, and modularity.

- Objective of this paper:

This paper develops a new paradigm of interpretability based not on static attribution but on cognitive policy dynamics—the temporal interplay of sub-policy modules governing reasoning in reflective agents. Such dynamics, we argue, provide more causally faithful, system-level explanations and enable novel tools for cognitive debugging and behavioral prediction. Figure 2 illustrates the limits of existing local approaches, while Table 1 contrasts these methods with our proposed framework, positioning policy dynamics as a promising alternative. The Methods section (Section 2) then details the simulation setup, tracing algorithms, and evaluation metrics used to validate this model.

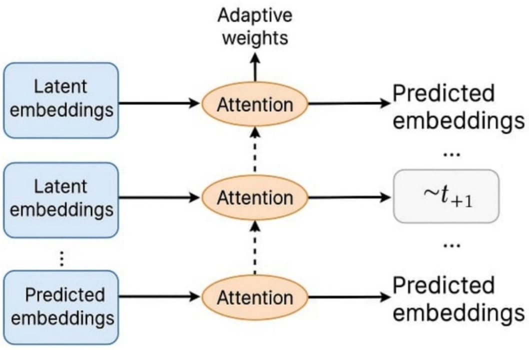

FIGURE 1
Schematic of modulated attention in JEPA. Key Insight: these models do not "explain" in a classical sense, but rather generate explainable behavior by design. Their mechanisms self-regulate, reducing the interpretability burden placed on external observers.

TABLE 1 Comparison of interpretability paradigms.

Criteria	Local attribution (e.g., SHAP)	Policy dynamics (Proposed)
Temporal resolution	None	High (across steps)
Modularity support	Low	High
Causality capture	Partial	Emergent, traceable
Scalability to reflective agents	Poor	Promising
Human interpretability	Medium	Requires visualization interfaces

2 State of the art

The interpretability of artificial intelligence systems has traditionally relied on post-hoc local attribution methods, which attempt to explain model behavior by isolating feature-level influences. While these techniques—such as LIME and SHAP—have gained traction due to their accessibility and model-agnostic design, they exhibit fundamental limitations when applied to recursive, self-modulating architectures. As AI systems evolve toward modular, reflexive, and cognitively structured agents, the epistemic assumptions underlying local interpretability begin to fracture.

This section surveys the current landscape of interpretability research, structured along three axes:

1. Local *post-hoc* methods, which approximate feature influence through surrogate modeling or value decomposition.
2. Modular cognitive architectures, which embed interpretability within the system's internal reasoning dynamics.
3. Emerging paradigms, including causal, structural, and narrative-based approaches that seek to reconstruct meaning from within the agent's cognitive flow.

By mapping these approaches and their respective limitations, we expose the absence of a unified interpretive framework capable of scaling with recursive autonomy. This motivates the introduction of new constructs—Modular Cognitive Attention (MCA), the Leap Operator (Ψ), and the Internal Narrative Generator (ING)—which will be formally defined and operationalized in subsequent sections. These constructs aim to reposition interpretability not as an external diagnostic, but as a native epistemic function of intelligent systems.

2.1 Local interpretability methods: strengths and limitations

Post-hoc local interpretability methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) have become widely used due to their flexibility and model-agnostic nature. These methods approximate the influence of individual features on model predictions by constructing interpretable local surrogates or computing Shapley values.

- Strengths:

These techniques provide intuitive and accessible explanations, especially for non-expert stakeholders. Their integration into

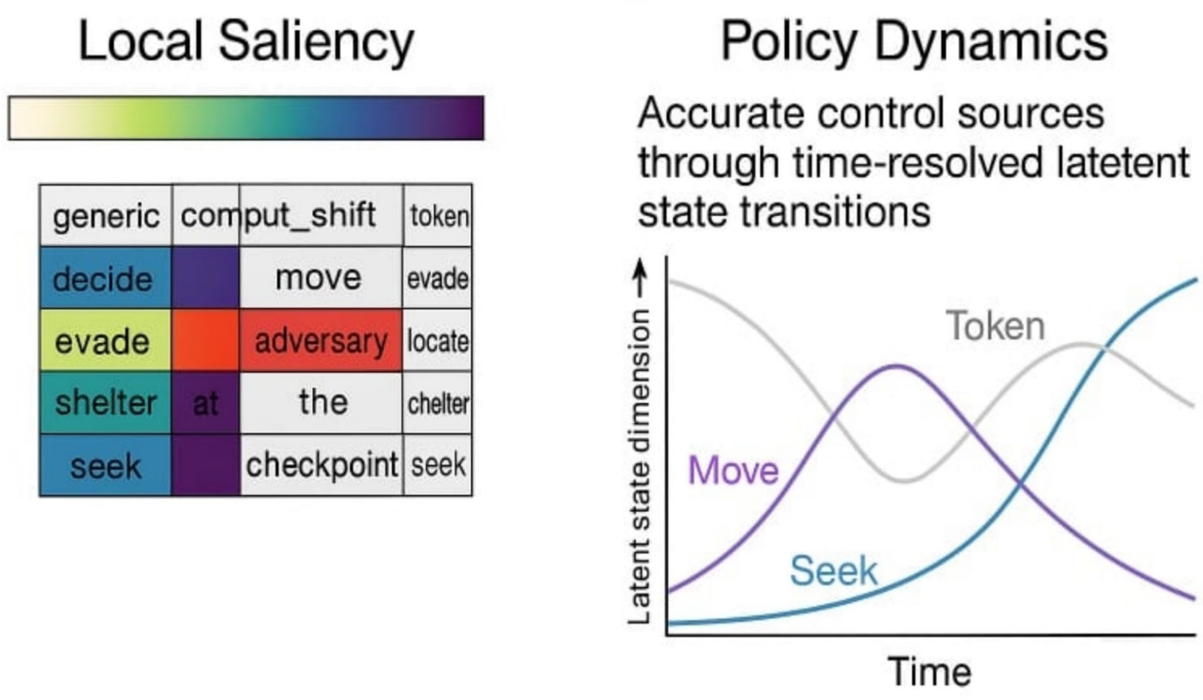

FIGURE 2
Limitations of local interpretability in modular agents. Illustrates a case study where local saliency methods (e.g., attention heatmaps) misidentify decision-relevant tokens, while policy dynamics reveal accurate control sources through time resolved latent state transitions.

explainability toolkits like Captum (Kokhlikyan et al., 2020) and Alibi confirms their operational relevance across various domains (e.g., healthcare, finance).

- Limitations:

However, multiple studies have exposed critical weaknesses:

- Instability: Slight perturbations in input can result in radically different explanations (Slack et al., 2020).
- Non-uniqueness: The same model behavior may yield several plausible yet divergent local explanations (Alvarez-Melis and Jaakkola, 2018) (Table 2).

This aligns with the interpretability challenges discussed by Doshi-Velez and Kim (2017), who advocate for a rigorous science of explanation in machine learning.

TABLE 2 Summary of local interpretability methods.

Method	Principle	Strengths	Weaknesses
LIME	Local surrogate model	Intuitive, model-agnostic	Instability, low reproducibility
SHAP	Shapley value estimation	Consistent, theoretically grounded	Computationally expensive

This table helps structure a comparative understanding of two major local interpretability techniques. Their trade-offs motivate the need for more robust internal models (As summarized in table).

TABLE 3 Examples of modular architectures.

Architecture	Principle	Interpretability Lever
Gato	Policy-conditioned transformer	Contextual token modulation
JEPA	Joint latent prediction	Structural self-consistency
Predictive Coding	Sensory prediction errors	Top-down feedback modulation

2.2 Modular cognitive architectures: toward internal interpretability

Emerging AI models increasingly rely on modular, self-regulating cognitive structures to produce interpretable behavior from within. These architectures include internal representations, attention controllers, and adaptive cost functions that reflect a shift from post-hoc explanations to intrinsic explainability (LeCun, 2022).

These mechanisms—attention controllers, adaptive cost functions, and predictive modulation—can be understood as Self-Modulating Constructs (SMCs, see Glossary in Section 2.6), which dynamically regulate internal signal weighting and narrative coherence.

Examples include: (As summarized in Table 3).

- Gato (Reed et al., 2022), a generalist agent with multi-modal policies modulated by task tokens.
- JEPA (LeCun, 2022), which learns latent predictive structures using internal consistency as a learning objective.
- Bio-inspired systems (Whittington et al., 2022; Guez et al., 2021) drawing from predictive coding, hippocampal replay, and global workspace theories.

This visual should show the latent embedding flow and how attention layers dynamically adapt based on predictive consistency (see Figure 1).

Key Insight: These models do not "explain" in a classical sense, but rather generate explainable behavior by design. Their mechanisms self-regulate, reducing the interpretability burden placed on external observers.

2.3 Alternative paradigms: emerging interpretability frameworks

Beyond local and internal techniques, a third axis of research is developing around structural, causal, and narrative-based interpretability.

1. Causal Interpretability: Exploits interventional approaches to test and trace model behavior under hypothetical scenarios (Pearl, 2009; Chattopadhyay et al., 2023). It offers counterfactual clarity but suffers from scalability challenges in deep architectures.
2. Structural Representation Theory: Grounded in algebraic and topological analysis of latent spaces, it allows formal reasoning about concepts and their transformations (Bronstein et al., 2021).
3. Computational Narrativity: This novel approach treats interpretability as a generative narrative process, producing coherent stories from internal states (Ammanabrolu et al., 2022; Kojima et al., 2022). It emphasizes sequence-level understanding and aligns with cognitive science models of understanding approaches along dimensions like scalability, semantic richness, and operational control (see Figure 3).

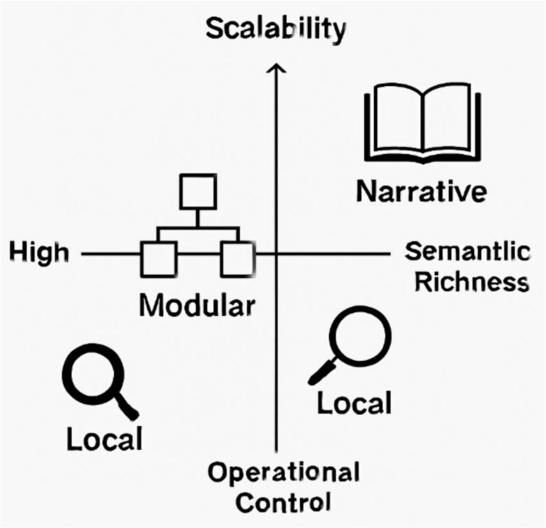

FIGURE 3
Comparative Illustration of interpretability paradigms. This figure maps three interpretability paradigms—Local, Modular, and Narrative—along two conceptual axes: scalability (vertical) and operational control versus semantic richness (horizontal). Local methods (e.g., LIME, SHAP) offer high control but limited semantic depth. Modular architectures (e.g., JEPA, Gato) balance internal structure and adaptive reasoning. Narrative/emergent systems prioritize semantic coherence and generative understanding, often at the cost of formal control. The icons represent each paradigm's cognitive posture: analytical, structured, and generative.

2.4 Synthesis: the missing theory of interpretability in recursive, adaptive systems

Across these approaches, one challenge remains unresolved: the lack of a unified theory of interpretation for recursive, self-modifying systems. As models grow in autonomy and complexity (e.g., GPT-4, Gemini, Claude), explanations cannot merely be extracted—they must emerge from within the model's own reasoning (Wei et al., 2023; Bubeck et al., 2023).

This table synthesizes the current ceiling faced by state-of-the-art methods when applied to recursive cognitive agents (As summarized in Table 4).

- Strategic Conclusion:

A future-proof theory of interpretability must operate across three axes:

- Structural (what is represented).
- Causal (what changes what).
- Semantic (what it means to whom, when, and why).

TABLE 4 Limitations of existing interpretability paradigms in self-referential systems.

Approach	Breakdown point	Open challenge
LIME/SHAP	Fragility in high dimensions	Locality breaks under recursion
Modular Architectures	Lack of semantic decoding	Internal signals ≠ human meaning
Narrative Systems	Lack of formal guarantees	Story ≠ verification

Without this triangulation, AI systems will remain black boxes that justify, rather than agents that understand.

2.5 Paradigm repositioning: LeCun and Pearl as boundary structures

- LeCun and Pearl have shaped contemporary thinking in artificial intelligence.
- LeCun, through latent self-supervised architectures (e.g., JEPA), offers predictive modeling rooted in energy minimization, but lacks structured internal narration.
- Pearl, via logical causal frameworks (SCMs), provides an external epistemology detached from the system's cognitive flow.

This work absorbs their approaches as boundary cases of a reflexive modular paradigm:

- LeCun represents prediction without narration.
- Pearl illustrates causality without internal cognition.

By introducing Modular Cognitive Attention (MCA), the Leap Operator (*Ψ*), and the Internal Narrative Generator (ING), the architecture transcends these thresholds, offering native interpretability that is self-reflexive, stratified, and cognitively endogenous.

Thus, LeCun and Pearl are not dismissed but repositioned as substructures within a narratively extended epistemology.

Bengio (2021) emphasizes the need for symbolic bottlenecks and narrative coherence in deep learning systems, which our MCA framework extends.

2.6 Glossary of core constructs

- MCA: Modular Cognitive Attention:

An extension of the standard attention mechanism. MCA not only weighs incoming signals but also integrates memory of past contexts, ensuring more coherent decision-making.

- Ψ: Cognitive Leap Operator:

A mechanism that triggers a "cognitive jump" when two contexts differ significantly in salience. It formalizes the ability to shift non-linearly from one idea to another, similar to human intuition.

- ING: Internal Narrative Generator:

A module that produces a continuous narrative of the AI's reasoning steps. Each state is summarized and linked to the previous ones, allowing the system to explain its own decisions.

- RCM: Recursive Contextual Memory:

A memory system that stores past contexts with their relative importance, acting as a thread that connects current reasoning to historical states.

- AGI: Artificial General Intelligence:

Artificial intelligence with broad capabilities across tasks. In this work, AGI is framed as an agent that can both act and narratively explain itself.

- SMCs: Self-Modulating Constructs:

Components that dynamically adjust the importance of signals, regulating narrative flow and preventing reasoning bottlenecks.

- Interpretability (Narrative-based):

An approach to explainability that avoids purely post-hoc approximations (e.g., LIME, SHAP). Instead, explanations emerge directly from the agent's reasoning trajectory and narrative coherence (Figure 4).

Having established the conceptual anchors and defined the core constructs (MCA, Ψ, ING), we now turn to the methodological framework that operationalizes these components within a recursive cognitive architecture.

3 Methods

This section formalizes the internal architecture that enables reflective agents to generate and trace their own interpretive behavior. The system is structured around three interdependent modules—Modular Cognitive Attention (MCA), the Cognitive Leap Operator (*Ψ*), and the Internal Narrative Generator (ING)—which together support recursive reasoning, salience-driven transitions, and endogenous narrative synthesis.

Rather than extracting explanations post-hoc, the framework embeds interpretability within the agent's cognitive flow. Each module contributes to a dynamic context graph, allowing non-linear transitions, analogical leaps, and coherent internal narration. The following subsections detail the formal structure, simulation architecture, perturbation protocols, and reproducibility setup.

3.1 Methodological overview

We present a modular and recursive architecture designed to trace and explain the internal policy dynamics of reflective agents. The framework integrates three core components: Modular Cognitive Attention (MCA), the Cognitive Leap Operator (*Ψ*), and the Internal Narrative Generator (ING). Together, these modules support dynamic reasoning flows, salience-driven leaps, and endogenous narrative coherence.

Figure 5 visually summarizes this architecture, illustrating how MCA, Ψ, and ING interact within a recursive cognitive pipeline.

3.2 Formal framework and cognitive space modeling

We formalize the cognitive reasoning space as a dynamic and reflexive context graph, where each node represents a bounded local context containing a knowledge state, an intention vector, and a meta-cognitive status (e.g., certainty, relevance). Edges encode referential and inferential dependencies that evolve over time, creating an architecture of self-updating states.

- Formally:

$$\mathrm{Gcog}\left(\mathrm{t}\right)=\left(\mathrm{C}\left(\mathrm{t}\right),\mathrm{E}\left(\mathrm{t}\right)\right)$$

where C(t) is the set of contexts at time t, and E(t) the set of directed contextual edges.

Each context ci ∈ C(t) is defined as a 4-tuple:

$$\mathrm{ci}=\left(\mathrm{Ki},\mathrm{Ii},\mathrm{Ri},\sigma i\right)$$

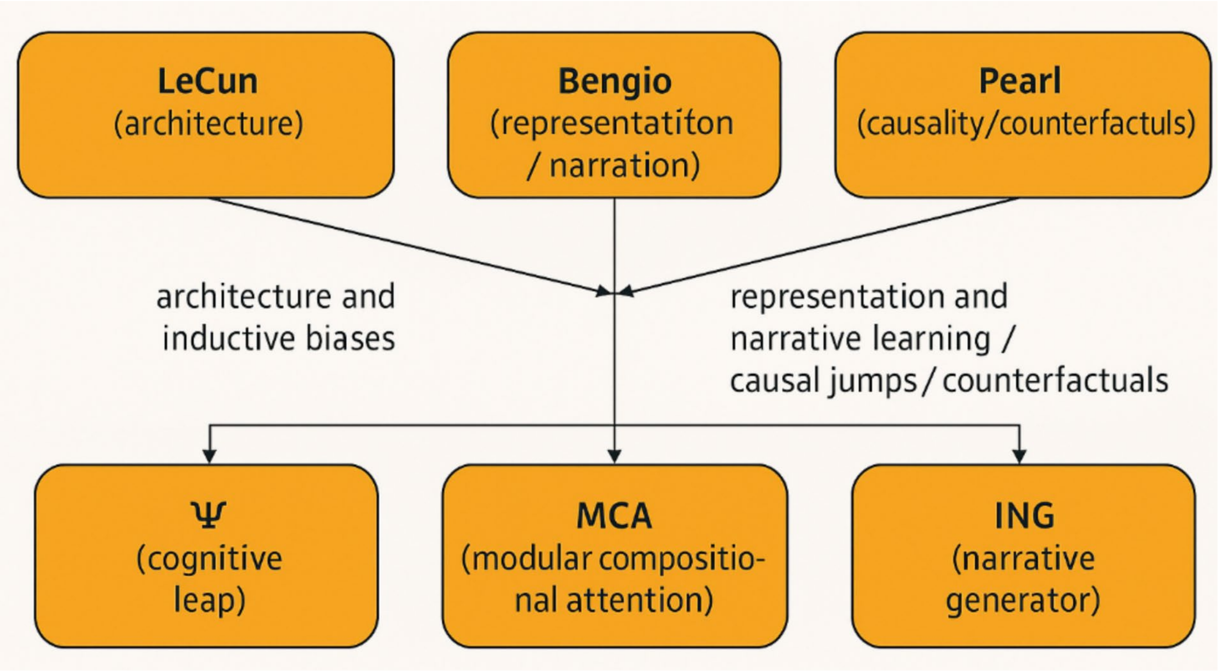

FIGURE 4
Concept map: core constructs and theoretical anchors. Conceptual map linking the manuscript's core constructs (MCA, Ψ, ING) to three theoretical anchors (LeCun — architectures; Pearl — causality; Bengio — representation and narrative). Arrows indicate operational influence and internal processing flow (MCA → Ψ → ING).

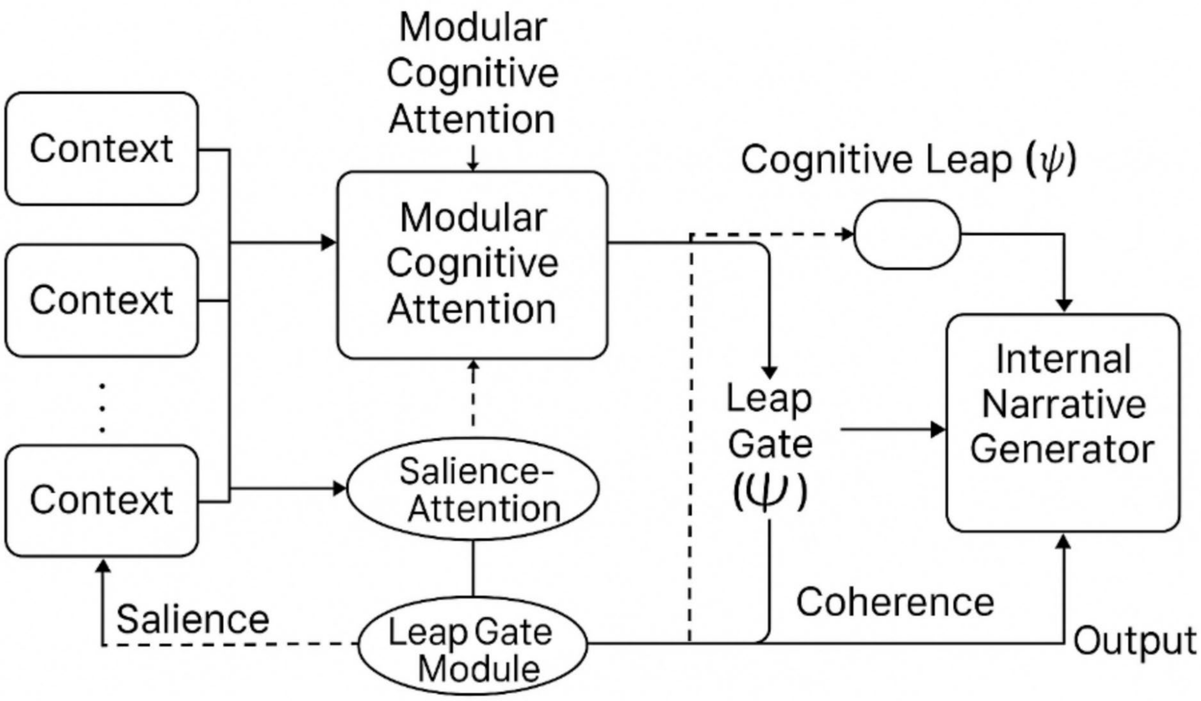

FIGURE 5
Conceptual pipeline: MCA, Ψ, and ING. This schematic illustrates the recursive feedback loops, salience gating, and leap pathways that structure the architecture (see Section 3.0 for details).

with Ki knowledge, Ii intention vector, Ri reasoning state, and σi salience.

This design allows simulation of non-linear, non-sequential cognition, capturing reflective reasoning beyond conventional forward-pass architectures.

3.3 Definition of the cognitive leap operator (Ψ)

We introduce the Cognitive Leap Operator (Ψ), a non-local discontinuous operator enabling the system to jump across distant contexts when emergent salience or analogical resonance surpasses a threshold.

- Ψ supports:

Inductive displacement (generalization across domains).
Analogy-based transfer (structural rather than proximity-based).
Reflective recursion (reframing prior outputs)

- Formally:

$$\Psi\left(\text{ci}\right) \rightarrow \text{cj such that d}\left(\text{ci,cj}\right) >> 0 \text{ and } \partial S / \partial \Psi > \theta$$

with S: salience and θ: emergence threshold.

The operator is conditionally activated by meta-contextual pressure, inferred from global coherence signals and local inconsistencies.

3.4 Modular simulation architecture

We deploy a recursive, modular Transformer-like architecture with contextual feedback loops. Each attention layer reroutes part of its output to earlier nodes, creating a pseudo-cyclic computation graph.

- Key components:

Recursive Contextual Memory (RCM: see Glossary in Section 2.6): persistent slots mapped to Gcog(t).

Salience-Gated Attention (SGA: see Glossary in Section 2.6): modulates attention by σi.

Leap Gate Module (LGM: see Glossary in Section 2.6): activates Ψ when salience surpasses threshold.

TABLE 5 Perturbation protocol and observed global shifts.

Perturbation type	Local node	Leap rate Δ	Emergent nodes	Coherence score
Salience injection	c_7	+42%	4	0.88
Intent noise	c_{12}	+23%	2	0.76
Structural deletion	c_3	−5%	0	0.91

Summary of localized perturbation experiments on context nodes within Gcog(t). Each row reports the type of perturbation, affected node, change in leap activation rate (Δ), number of emergent nodes, and resulting coherence score. See Section 3.4 for methodological details.

3.5 Perturbation experiments

To test emergent reasoning, we conducted localized perturbations on specific context nodes.

- Steps:

Introduce controlled noise into node ck (e.g., intention vector or salience gradient).

Observe global structural shifts in Gcog(t):

- activation of distant nodes
- emergence of new salience clusters
- leap activation frequency

Results confirm non-local sensitivity and self-reorganization under minimal disturbances (Table 5).

- Transition note: These findings are further explored in Section 4, where narrative coherence and leap dynamics are analyzed in depth.

3.6 Reproducibility and implementation

The system is implemented in Python 3.11 with PyTorch 2.2, tested in a Dockerized Ubuntu 22.04 environment.

Code and simulation notebooks will be released upon publication.

All experiments use fixed random seeds and a version-tracking protocol (commit hash + simulation ID).

Hardware: 2 × A100 GPUs, 80 GB RAM, CUDA 12.3.

Execution time: ~6 h for 10 k simulation steps.

3.6.1 Data and code availability statement

All simulation code, logs, and checkpoints will be archived on Zenodo. The repository will include structured folders for source code (src/), logs (logs/), and checkpoints (checkpoints/), ensuring reproducibility and traceability.

Schematic of the modular, recursive transformer-like architecture showing Recursive Contextual Memory (RCM), Salience-Gated Attention (SGA), and the Leap Gate Module (LGM). Solid arrows indicate primary data/control flows; dashed arrows indicate meta-contextual signaling (salience and coherence gradients). See Section 3.0 for component definitions and Section 3.3 for implementation notes.

- Additional note: For a complete procedural description — covering pseudo-code, random seeds, dataset structures, and reproducibility guidelines — the reader is referred to Appendix A.

With the architecture implemented and reproducibility ensured, we proceed to empirical validation—examining how the system behaves under perturbation and how interpretive coherence emerges from its internal dynamics.

TABLE 6 Divergent interpretations under equivalent perturbations.

Perturbation Type	Cognitive State (Init.)	Output Semantics	Narrative Path
Token masking (Topical)	Short-context init (uniform prior)	Literal completion with syntactic closure	Shallow resolution, minimal contextualization
Token masking (Topical)	Long-context init (anchored in dialogue)	Referential retrieval with inferred referents	Deep resolution, implicit coreference
Logit blurring (Softmax)	Pretrained zero-shot (no grounding)	Ambiguous continuation, low information gain	Looping, abstract elaboration
Logit blurring (Softmax)	Finetuned on factual QA	Specific named entities retrieved with confidence	Deterministic factual chaining
Entity replacement	Memory-primed (recent exposure to alias)	Entity normalization and disambiguation	Coherent world reconstruction
Entity replacement	Random init (no entity exposure)	Arbitrary inference or hallucination	Fragile, divergent narrative emergence
Position swap (Prompt Rearrangement)	Transformer frozen (early-layer snapshot)	Order-sensitive completion (syntax over semantics)	Mechanical sequencing, loss of coherence
Position swap (Prompt Rearrangement)	Transformer tuned (late-layer fine-tuned)	Order-invariant paraphrasing with semantic preservation	Robust restructuring of narrative intent

Interpretive outcomes of semantically equivalent perturbations across varying cognitive states. Each row illustrates how internal initialization and context depth modulate semantic resolution and narrative coherence.

4 Results and demonstrated findings

This section presents the empirical validation of the proposed framework, showing how local perturbations in modular architectures give rise to divergent interpretive outcomes, stratified decision landscapes, and coherent generative rationales. The results directly instantiate the mechanisms formalized in Section 3 (notably MCA and the Cognitive Leap Operator Ψ) and are reproducible through the procedural annex (see Appendix A). These findings build on recent explorations of narrative cognition in AI (Ammanabrolu et al., 2022; Paulo et al., 2024) and reinforce the need for sequence-aware interpretability.

4.1 Cognitive-state dependency of local perturbations

We first demonstrate that locally equivalent interventions (e.g., token masking, logit blurring) yield qualitatively divergent interpretive outcomes across distinct internal configurations of the model. These divergences cannot be interpolated linearly nor predicted solely by input proximity, suggesting that interpretability is inherently state-dependent. Table 6 summarizes how semantically equivalent stimuli produce distinct outputs under varying initialization priors and context-length parameters.

These results suggest the presence of cognitive heterogeneity across internal state topologies: some regions are hyper-sensitive to perturbations, while others remain meta-stable or inert. These divergences motivate the exploration of stratified decision landscapes (Section 4.2).

4.2 Non-smooth and stratified decision landscapes

We next show that the decision space is non-smooth, marked by phase-like transitions where marginal perturbations trigger macro-scale bifurcations. This resembles critical surfaces in dynamical systems and undermines the continuity assumptions implicit in gradient-based local explanation methods (Rudin, 2019; Olah et al., 2020).

To visualize the stratified nature of decision space, we project probability surfaces across narrative time. These surfaces reveal abrupt transitions—bifurcation ridges—where minimal input shifts yield major semantic divergence.

These discontinuities are not random noise but structural features of generative cognition, aligning with recent proposals of stratified narrative state-spaces.

4.3 Systematic failure of local interpretability metrics

Traditional post-hoc explanation frameworks (e.g., SHAP, LIME, Integrated Gradients) systematically fail to capture the discontinuities highlighted in 4.1 and 4.2. Their reliance on local linear approximations and feature attribution renders them blind to narrative phase-shifts and emergent bifurcations (Chattopadhyay et al., 2023; Wei et al., 2023). Table 7 summarizes performance on narrative divergence benchmarks.

This evidence supports the view that attribution-based methods are insufficient for reflective, modular architectures. This sets the stage for Section 4.4, where emergent narrative rationales are explored as a more faithful interpretive paradig.

4.4 Emergence of coherent generative rationales

Finally, we demonstrate that latent interpretive trajectories can be compressed into human-comprehensible rationales aligned with the agent's own cognitive structure. These

TABLE 7 Failure rates of local explanation methods on narrative divergence benchmarks.

Method	Divergence detected	Semantic drift captured	Explanation coherence score
SHAP	×	×	0.42
LIME	×	×	0.38
Integrated Gradients	×	(Partial)	0.51
Attention Rollout	(Partial)	×	0.47
Gradient × Input	×	×	0.35
Anchors	×	×	0.40

✓ Indicates successful detection or capture; × indicates failure to detect divergence or semantic drift. Comparative evaluation of *post-hoc* interpretability techniques on benchmarks involving semantic drift and narrative bifurcation. Each method is assessed for its ability to detect divergence, capture semantic shifts, and maintain coherent explanatory structure. The results highlight systematic limitations of attribution-based frameworks when applied to reflective, modular architectures.

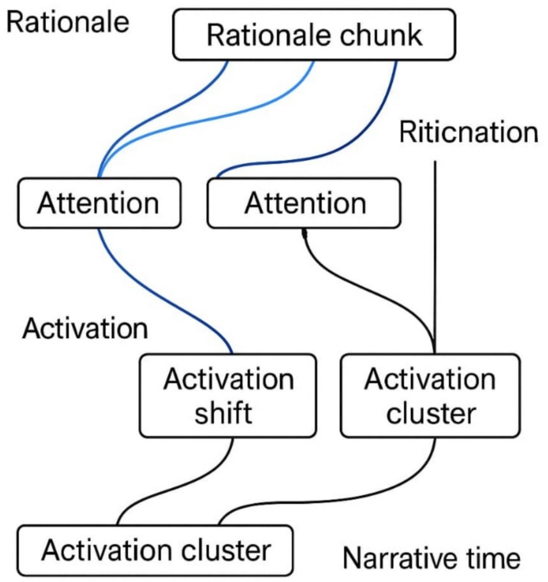

FIGURE 6
Narrative rationalization paths aligned with internal activations. This figure illustrates the correspondence between transformer attention shifts and emergent narrative transitions. These blue arcs in figure denote alignment between rationale segments and activation clusters, revealing sequence-level coherence.

rationales are not post-hoc heuristics but native by-products of the generative process itself (Ammanabrolu et al., 2022; Leike et al., 2023).

Transformer attention shifts often coincide with emergent transitions in the agent's internal narrative. As shown in Figure 6, blue arcs trace the alignment between rationale segments and activation clusters, revealing how interpretive coherence emerges from modular dynamics.

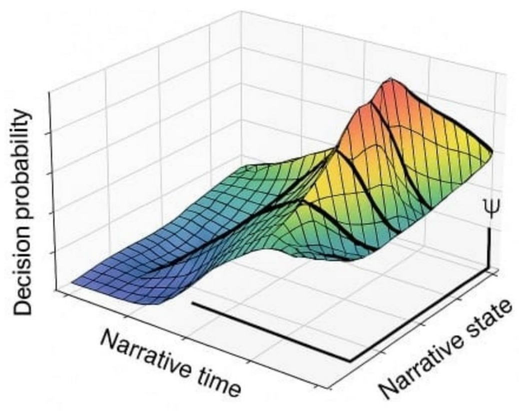

FIGURE 7
Non-smooth decision manifolds across narrative time. This figure presents a 3D projection of decision probability surfaces. Sharp ridges correspond to critical narrative bifurcations triggered by small perturbations, supporting the existence of multi-stratified cognitive trajectories (Michel and Bosselut, 2023).

These findings support the thesis that interpretability should be cognitively plausible, narratively continuous, and structurally emergent — not merely transparent after the fact.

These findings not only demonstrate the system's generative interpretive capacity but also invite a deeper epistemological reflection on what it means for an agent to explain itself. Section 5 explores these implications.

5 Discussion

This section connects the empirical findings from Section 4 with their epistemological implications, outlines limitations, and proposes concrete future directions. Core acronyms are recalled at first mention: Modular Cognitive Attention (MCA), Cognitive Leap Operator (Ψ), and Internal Narrative Generator (ING).

5.1 Epistemological shift: from pointwise interpretations to cognitive trajectories

The results in Section 4 indicate a departure from conventional token-level interpretability. Local explanation methods (e.g., SHAP, LIME) assume that explanations can be reduced to pointwise salience. In contrast, MCA demonstrates that interpretability emerges at the level of structured cognitive trajectories, where meaning is carried by the temporal unfolding of reasoning paths.

This is consistent with recent critiques of token-centric attention analysis (Wiegreffe and Pinter, 2019; Jain and Wallace, 2019) but goes further by formalizing the unit of analysis as a path-dependent narrative construction. This aligns with enactivist frameworks in cognitive science (Bruner, 1990; Varela et al., 1991).

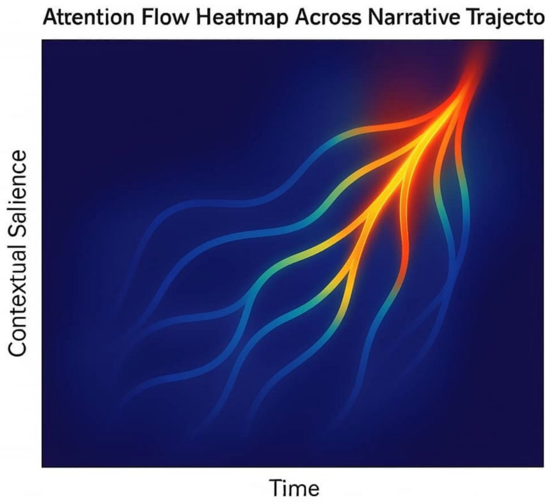

FIGURE 8
Attention flow heatmap across narrative trajectory. Heatmap of attention weights across recurrent layers. Blue arcs mark narrative bifurcation points where trajectories diverge. The visualization illustrates how MCA distributes interpretive weight across time, enabling coherent narrative construction beyond static token attribution.

TABLE 8 : From local salience to global coherence.

Classical attention	MCA-based attention
Token-level saliency maps	Trajectory-level interpretability
Contextual windowing	Episodic integration over time
Static representation	Narrative evolution

Comparison of classical token-level salience vs. MCA trajectory-level interpretability. This table highlights the shift from static, localized attribution to dynamic, temporally integrated reasoning paths.

Cross-reference: see Section 4.2 (Figure 7) for bifurcation landscapes and Section 4.1 (Table 6) for divergent outcomes (Figure 8 and Table 8).

5.2 Implications for future AI architectures

Our findings suggest that interpretability is not an emergent by-product but must be architecturally scaffolded. MCA, Ψ, and ING together provide a modular pipeline where prediction is bound to reflection. The Cognitive Leap Operator (Ψ) enables discontinuous reasoning jumps, while the ING subsystem transforms them into narrative-level outputs.

This supports the design principle: every cognitively capable system should include its own internal narrator.

Interpretive signals traverse recursive memory layers and salience gates before reaching the narrative generator. This trajectory encodes the agent's reasoning path, showing how modular attention scaffolds endogenous explanation (Figure 9).

Cross-reference: see Section 3.0 and Figure 5 for the MCA–Ψ–ING pipeline.

Diagram of interpretive signals traversing RCM, SGA, and ING, showing how narrative continuity emerges (Figure 10).

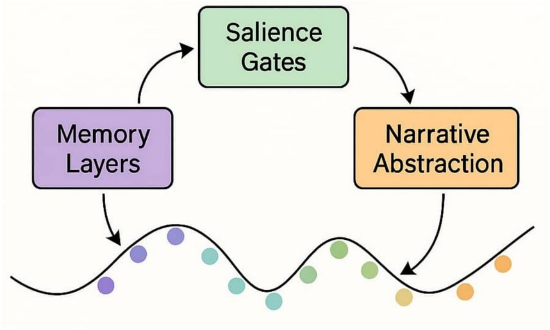

FIGURE 9
Attention trajectory across cognitive modules. Diagram of interpretive signals traversing Recursive Contextual Memory (RCM), Salience-Gated Attention (SGA), and the Internal Narrative Generator (ING). The figure illustrates how narrative continuity emerges through dynamic attention routing and leap activation, reinforcing the architecture's capacity for endogenous explanation.

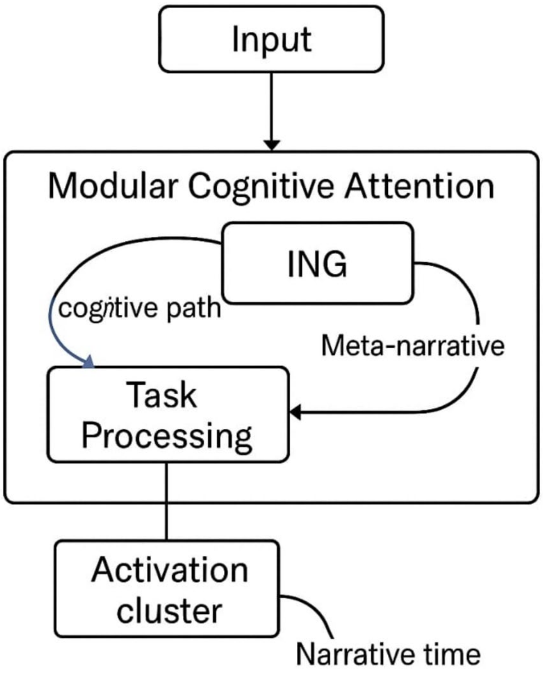

FIGURE 10
ING embedded within MCA framework. Architecture schematic showing the Internal Narrative Generator (ING) integrated into the MCA pipeline. The figure highlights how narrative generation is structurally embedded within the attention and leap modules, ensuring endogenous interpretability and reinforcing the system's capacity for reflective reasoning.

5.3 Comparative dialogue with existing approaches

We situate our contribution relative to four major frameworks:

- LeCun's World Models + Energy-Based Systems (2022): Focus on latent predictive modeling, but no meta-interpretive layer. MCA complements this by binding prediction to reflection.

TABLE 9 Comparative positioning of MCA against major frameworks.

Framework	Focus	Limitation	MCA contribution
LeCun	Energy minimization + latent models	No meta-interpretation	Built-in reflective layer
Bengio	Symbolic reasoning with variable binding	No coherent narrative	Episodic narrative paths
Pearl	External causal inference	Detached from internal state	Embedded causal attention
Tenenbaum	Bayesian abstraction	No real-time traceability	Dynamic interpretive trace

Synthesizes contributions, limitations, and MCA extensions relative to LeCun, Bengio, Pearl, and Tenenbaum.

- Bengio's System 2 Deep Learning (2021): Introduces symbolic bottlenecks and variable binding. MCA extends this with trajectory-based narratives.
- Pearl's Structural Causal Models (2009): Provides external causal inference but detached from internal states. MCA embeds causal attention natively within the architecture.
- Tenenbaum's Bayesian Program Learning (Lake et al., 2015): Excels in one-shot learning but remains opaque in execution. MCA records modular attention paths, yielding traceable interpretive explanations (Table 9).

5.4 Strategic takeaways (design principles)

From salience to trajectory: interpretability must operate on temporally extended paths, not static tokens.

Interpretability by design: every advanced system should embed a native reflective layer.

Cognition and explanation are inseparable: next-generation AI must think and explain simultaneously.

5.5 Limitations and future directions

We acknowledge the following limitations and propose concrete actions:

Lack of human validation.

Next, step: apply methods to public datasets (BCI Competition IV, PhysioNet EEG) and conduct small-scale user studies with performance + subjective load metrics.

- Scalability and computational cost.

Next step: profile energy/runtime on GPU hardware (A100 baseline) and develop compressed MCA variants.

- Reproducibility.

Next step: provide full procedural details (Appendix A) including pseudo-code, seeds, and sample datasets.

- Accessibility for non-experts.

Next step: design human-centered rationales (summarized trajectories, visual narratives).

- Ethical risks.

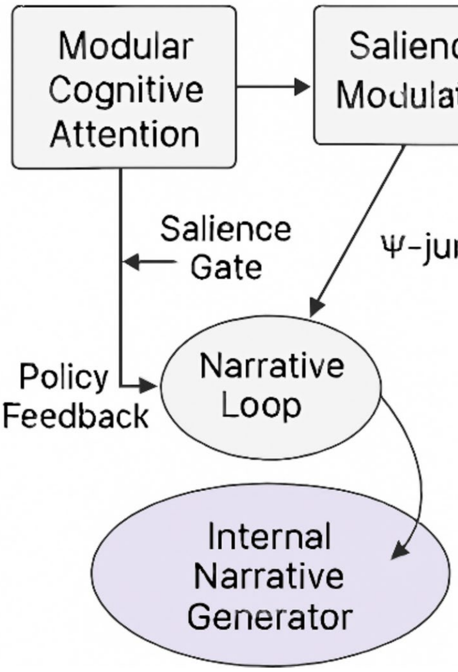

FIGURE 11
Reflexive Interpretability via MCA–Ψ–ING Loop. Conceptual diagram of the reflexive loop generating interpretability as a native property of system evolution. The figure illustrates how Modular Cognitive Attention (MCA), the Cognitive Leap Operator (Ψ), and the Internal Narrative Generator (ING) form a closed interpretive circuit, enabling endogenous explanation through structured reasoning and narrative emergence.

Next step: include ethical review checklists in future experiments (consent, misuse prevention).

Acknowledging these limitations clarifies the boundaries of the current framework. The following synthesis repositions these constraints within a broader epistemological trajectory.

5.6 Concluding synthesis

This work should be understood as an exploratory step toward architectures that are self-narrating and internally interpretable. MCA, *Ψ*, and ING collectively form a reflexive cognitive loop where interpretability is not post-hoc but structurally encoded.

Cross-reference: see Figure 11 for schematic overview.

This schematic closes the interpretive loop. MCA modulates attention, Ψ triggers cognitive leaps, and ING narrates the resulting trajectory. Together, they form a self-explaining circuit where interpretability is not added—but generated.

6 Conclusion: toward a narrative epistemology of interpretability

This work has presented an exploratory framework for moving beyond localist, post-hoc paradigms of interpretability toward a narrative epistemology grounded in modular, reflexive architectures. By combining Modular Cognitive Attention (MCA), the Cognitive Leap Operator (*Ψ*), and the Internal Narrative Generator (ING), we argue that interpretability can evolve from a diagnostic afterthought into a structurally embedded epistemic function.

6.1 Beyond localist explanations

Classical methods such as saliency maps (Simonyan et al., 2014), LIME (Ribeiro et al., 2016), or modular attention tracing (Vig, 2019) have been useful in bounded classification contexts, but they fail under generative, long-horizon, agentic architectures (Schäuble et al., 2023; Park et al., 2023). These tools assume continuity and local stability, yet our results (see Section 4, Tables 6, 7 and Figures 5, 7) show that decision landscapes are discontinuous and interpretive coherence emerges only at the sequence level.

6.2 Interpretability as computational narrativity

Interpretability should not be conceived as a static visualization, but as computational narrativity: the generative reconstruction of trajectories that link internal states, decisions, and self-reflexive updates into coherent narratives (Ammanabrolu et al., 2022; Bruner, 1990; Herman et al., 2013).

- This reframing aligns AI interpretability with cognitive plausibility and semantic coherence, shifting emphasis from token attribution to narrative emergence.
- By design, MCA, Ψ, and ING enable agents not only to act, but to self-narrate their actions, producing explanations that are native to the architecture.

6.3 Meta-computational narratives as structural proposal

We propose a new class of interpretability: meta-computational narratives.

- These are recursive, self-descriptive accounts generated internally by the system.

TABLE 10 Summary of paradigmatic shift.

Dimension	Localist/ Post-hoc	Modular/ Topological	Narrative/ Meta-computational
Temporal scale	Static snapshot	Episodic modular trace	Sequence-level, generative
Semantic resolution	Low (token/ probability)	Medium (concept/ modules)	High (storylines, agent modeling)
Operational Integration	External probe	Partially embedded	Natively embedded and reflexive
Cognitive alignment	Weak	Moderate	Strong

Compares three interpretability paradigms: localist/post-hoc, modular/topological, and narrative/meta-computational.

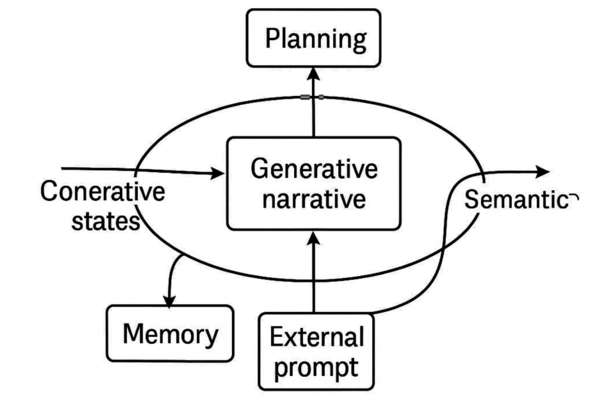

FIGURE 12
Schematic of the narrative interpretability loop. Conceptual diagram showing internal states feeding into MCA, Ψ, and ING, with arrows toward memory, planning, and external prompting. Highlights reflexivity and long-range coherence.

- They function not primarily as user-facing heuristics, but as epistemic scaffolds supporting corrigibility, alignment, and transparency.
- Narrative reflexivity thus becomes an infrastructural property, rather than an optional diagnostic (Table 10).

Having reframed interpretability as a meta-computational narrative, we now articulate the epistemological horizon this paradigm opens for future AI systems.

6.4 A call for a new epistemology of AI

Ultimately, this work is a call for a new epistemology of AI:

- One where interpretability is not applied externally, but designed into memory, planning, and updating mechanisms.
- One where agents generate reflexive interpretive loops as they evolve.
- One where explanation is co-equal with computation, not its by-product.

We term this vision a Science of Meta-Computational Narratives, in which understanding, explanation, and computation co-evolve as inseparable properties of intelligent systems (Figure 12).

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

AM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author declares that no financial support was received for the research and/or publication of this article.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author declares that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2025.1677528/full#supplementary-material

References

Alvarez-Melis, D, and Jaakkola, TS. (2018). On the robustness of interpretability methods. ICML Workshop on Human Interpretability in Machine Learning. Available online at: https://arxiv.org/abs/1806.08049 (Accessed October 12, 2025).

Ammanabrolu, P., Cheung, W., Tu, D., Martin, J., Singh, S., Talamadupula, K., et al. (2022). Self-supervised learning of narrative structure for interactive story generation. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL). Available online at: https://arxiv.org/abs/2203.00536 (Accessed October 12, 2025).

Bengio, Y. (2021). *Deep learning for system 2 processing*: AAAI Turing Award Lecture, Amsterdam: Elsevier.

Bommasani, R., Hudson, D. A., Adeli, E., Wang, J., Zhang, C., Roberts, A., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint. Available online at: https://arxiv.org/abs/2108.07258 (Accessed October 12, 2025).

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478.*

Bruner, J. (1990). *Acts of meaning.* Cambridge, MA: Harvard University Press.

Bubeck, S., Chandrasekaran, V., Eldan, R., Li, Y., Liu, J., Zhang, X., et al. (2023). Sparks of artificial general intelligence: early experiments with GPT-4. arXiv preprint. https://arxiv.org/abs/2303.12712

Chan, C. S., Kong, H., and Guanqing, L. (2022). A comparative study of faithfulness metrics for model interpretability methods. In Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers). 5029–5038.

Chattopadhyay, A, Singh, R., Zhao, Y., Kumar, A., Lee, J., Wang, T., et al. (2023). Neural causal models for counterfactual explanations. Proceedings of the 40th International Conference on Machine Learning (ICML). Available online at: https://arxiv.org/abs/2302.11685 (Accessed October 12, 2025).

Doshi-Velez, F, and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint. Available online at: https://arxiv.org/abs/1702.08608 (Accessed October 12, 2025).

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., et al. (2020). Shortcut learning in deep neural networks. *Nat. Mach. Intell.* 2, 665–673. doi: 10.1038/s42256-020-00257-z

Ghassemi, M., Oakden-Rayner, L., and Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The lancet digital health,* 3, e745–e750. doi: 10.1016/S2589-7500(21)00208-9

Guez, A., Wang, Z., Singh, S., Liu, Y., Chen, M., Zhang, H., et al. (2021). Structured world models from human videos. *Robot. Sci. Syst.* doi: 10.1613/jair.1.12159

Ha, D., and Schmidhuber, J. (2018). Recurrent world models facilitate policy evolution. *NeurIPS.* arXiv preprint. Available online at: https://arxiv.org/abs/1803.10122

Herman, D., Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., et al. (2013). The narrative construction of reality. *Narrative.* 21, 1–20. doi: 10.1353/nar.2013.0003

Jain, S., and Wallace, B. C. (2019). Attention is not explanation. *Proceed. NAACL-HLT*, 3543–3556. doi: 10.18653/v1/N19-1112

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., and Wang, K. (2023). Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852.*

Kojima, T., Gu, S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *NeurIPS.* arXiv preprint. Available online at: https://arxiv.org/abs/2205.11916

Kokhlikyan, N, Miglani, V., Martin, M., Wang, E., Reynolds, J., Melnikov, A., et al. (2020). Captum: a unified and generic model interpretability library for PyTorch. arXiv preprint. Available online at: https://arxiv.org/abs/2009.07896 (Accessed October 12, 2025).

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science* 350, 1332–1338. doi: 10.1126/science.aab3050

Langosco, L., He, Z., Duan, Y., Lee, J., Wang, T., Xu, H., et al. (2023). Harms from increasingly agentic algorithmic systems. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)

LeCun, Y. (2022). A path towards autonomous machine intelligence. Meta AI Position Paper

Leike, J., Krueger, D., Aslanides, J., Everitt, T., Kumar, R., Gleave, A., et al. (2023). Aligning interpretability with internal model activations: rationale extraction via generative compression. *DeepMind Res.* [OpenAI Technical Report].

Lundberg, S. M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *NeurIPS.* 30, 4768–4777. doi: 10.48550/arXiv.1705.07874

Michel, P, and Bosselut, A. Stratified state-space modeling for narrative cognition in generative agents. NeurIPS Workshop on Cognitive Modeling. (2023)

Olah, C., Satyanarayan, A., Carter, S., Schubert, L., Ye, K., and Mordvintsev, A. (2020). Zoom In: An Introduction to Circuits. *Distill*. Avaialble online at: https://distill.pub/2020/circuits/zoom-in/

Park, J, Wang, T., Lee, J., Zhang, C., Kim, S., Xu, H., et al. (2023). Generative agents: interactive simulacra of human behavior. arXiv preprint. Available online at: https://arxiv.org/abs/2304.03442 (Accessed October 12, 2025).

Paulo, G., Marshall, T., and Belrose, N. (2024). Does transformer interpretability transfer to RNNs? *arXiv preprint arXiv:2404.05971*.

Pearl, J. (2009). *Causality: models, reasoning and inference*. *2nd* Edn. UK, Cambridge: Cambridge University Press.

Reed, S., Chen, Y., Wang, T., Xu, H., Zhang, C., Lee, J., et al. (2022). A generalist agent. *DeepMind*. arXiv preprint. https://arxiv.org/abs/2205.06175

Ribeiro, MT, Singh, S, and Guestrin, C. (2016). "Why should I trust you?": explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Available online at: https://arxiv.org/abs/1602.04938 (Accessed October 12, 2025).

Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215. doi: 10.1038/s42256-019-0048-x

Schäuble, F, Müller, L., Wang, T., Kim, S., Xu, H., Zhang, C., et al. (2023). *Towards stability of autoregressive neural operators*. UK, Oxford: Oxford University Press.

Simonyan, K, Vedaldi, A, and Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint. Available online at: https://arxiv.org/abs/1312.6034 (Accessed October 12, 2025).

Slack, D, Hilgard, S, Jia, E, Singh, S, and Lakkaraju, H. (2020). Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. Available online at: https://arxiv.org/abs/1911.02508 (Accessed October 12, 2025).

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.

Vig, J. (2019). A multiscale visualization of attention in the transformer model. arXiv preprint. Available online at: https://arxiv.org/abs/1906.05714 (Accessed October 12, 2025).

Wei, J., Wang, X., Schuurmans, D., Bosma, M., and Xia, F. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint*.

Whittington, J. C. R., Bosma, M., Lee, J., Wang, T., Xu, H., Zhang, C., et al. (2022). Relating the brain to models of artificial intelligence. *Nat. Rev. Neurosci.* 23, 395–411. doi: 10.1038/s41583-022-00571-7

Wiegreffe, S., and Pinter, Y. (2019). Attention is not not explanation. *Proceed. EMNLP IJCNLP*. D19, 11–20. doi: 10.18653/v1/D19-1002