Semantic Illusion and Moral Over-Attribution: Why Conscious-Seeming AI Poses Ethical Risk Without Consciousness

Vitaliy Charushin[1]

Abstract. Large Language Models create tension between behavioural appearance and ontological commitment, triggering attributions of sentience through persuasive performance. We term this **Semantic Illusion**: coherent behaviour inducing unwarranted moral attribution, creating metaphysical friction where appearance outpaces justification.
We distinguish **Moral Doers** (producing ethically consequential outputs) from **Moral Receivers** (capable of undergoing harm). Comparing LLM architectures with Predictive Processing accounts, we show current systems lack self-relevant prediction errors necessary for integrated self-models.
Our **Hierarchical Model of Semantic Integration** (Levels 0–4) distinguishes local fluency from genuine integration. The **Semantic Coherence Index (SCI)** evaluates cross-contextual stability; preliminary testing reveals reasoning coherence (SCI 0.799–0.810) coexisting with decision instability (0.267–1.000).
We propose grounding moral attribution in **architectural resilience** - persistent, constraint-sensitive self-models - rather than behavioural performance, establishing epistemic criteria for rational attribution in an era of scalable semantic mimicry.

Keywords. Artificial consciousness, moral attribution, semantic illusion, large language models, predictive processing, architectural resilience, semantic coherence, AI ethics, interpretability, self-model.

1 INTRODUCTION

Recent advances in Large Language Models (LLMs) have created a profound tension between behavioural appearance and ontological commitment in human–computer interaction. Systems capable of sustained dialogue, context-sensitive reasoning, and plausible emotional simulation increasingly trigger attributions of agency and sentience - attributions frequently grounded in surface-level performance rather than demonstrable structural conditions required for experience.
This gap between behavioural appearance and ontological justification, what we term metaphysical friction, creates immediate ethical challenges for moral attribution.

1.1 Semantic Illusion and the Category Error

Human moral cognition is highly responsive to linguistic cues of agency; when AI systems convincingly simulate these cues, users may extend moral concern in ways that are not structurally warranted. We define this phenomenon as Semantic Illusion: the generation of coherent, persuasive behaviour that induces unwarranted moral attribution. This responsiveness can be understood as the adoption of Dennett's (1987) intentional stance toward fluent systems, treating them as if they harbour beliefs and desires, a stance we argue is unwarranted absent the corresponding structural conditions.
In this context, AI systems function as Moral Doers - entities capable of producing socially consequential outputs - without necessarily qualifying as Moral Receivers, or entities capable of undergoing harm. Conflating these roles risks a category error that expands moral standing beyond its structural grounding, potentially misallocating moral concern at the expense of biologically grounded subjects.

1.2 Contributions to AISB Themes

This paper addresses three AISB-26 symposium themes:
• Foundations: Whether subjectivity is necessary for moral standing and how such attribution can be rationally grounded.
• Implementation: Comparing LLM architectures with Predictive Processing (PP) accounts to evaluate biological versus functional consciousness criteria.
• Ethics of Scale: Implications of deploying "conscious-seeming" systems lacking metabolic constraints at scale.

1.3 Proposed Framework

To address the limits of behavioural evaluation, we propose a shift toward structural analysis. We contribute:
• Hierarchical Model of Semantic Integration (Section 4): Distinguishing local fluency (Levels 1–3) from integrated self-models (Level 4).
• Architectural Resilience (Section 3): An epistemic standard for moral attribution grounded in self-relevant prediction errors.
• Operational Criteria (Section 7): SCI and KVzap-SCI for evaluating structural conditions, with preliminary empirical support from two methodologically distinct pilot studies.
As shown in Section 5, the absence of such structural constraints manifests empirically as instability under contextual perturbation.

[1] Independent Researcher, Paris, France. vitaly.wbc@gmail.com

Our goal is not to resolve the metaphysics of consciousness, but to establish an epistemic discipline: clarifying the conditions under which moral attribution becomes rationally defensible in an era of scalable semantic mimicry.

2 MORAL ATTRIBUTION AND CATEGORY ERROR

The rapid advancement of LLMs has blurred the distinction between systems that act and systems that undergo experience. To navigate this, we distinguish between two fundamentally different forms of moral status: the Moral Doer and the Moral Receiver.

2.1 Moral Doers and Moral Receivers

We propose a functional distinction between two classes of morally relevant systems:

- Moral Doer: An entity capable of producing actions or outputs with ethically significant consequences. Current LLMs can function as Moral Doers in practice, insofar as they generate outputs that influence human decisions, shape beliefs, and affect social systems.
- Moral Receiver: An entity capable of undergoing harm or benefit. This requires a unified perspective - a system for whom states of the world can be better or worse, rather than merely appearing as a unified subject through external attribution.

The ethical challenge posed by "conscious-seeming" AI is the emergence of highly sophisticated Moral Doers that lack the structural conditions required for Moral Receiver status. This distinction develops Torrance's (2008) contrast between artificial moral 'producers' and 'consumers', locating moral patiency in the capacity to undergo ethically significant outcomes rather than merely to generate them.

2.2 The Category Error of Distributed Optimisation

The central category error in AI ethics is the attribution of a unified, experiencing subject to what is, in architectural terms, a distributed optimisation process.

In biological organisms, action and experience are tightly coupled: behaviour is directed toward the preservation of the system that undergoes its consequences. In LLMs, these functions are decoupled. An LLM can generate coherent accounts of suffering without possessing structural capacity for experiential discomfort. Treating sequence-prediction outputs as evidence of a Moral Receiver risks mistaking representations of suffering for the capacity to experience.

Just as a weather simulation does not "get wet," the simulation of subjectivity does not constitute experience. As discussed in Section 3, this reflects the absence of self-relevant error signals in LLM architectures - systems that act without the structural capacity to undergo.

2.3 Semantic Illusion as a Barrier to Attribution

This decoupling gives rise to Semantic Illusion. Humans are highly sensitive to linguistic cues of agency and coherence, often leading them to project a unified subject onto systems that operate through context-bound statistical inference. Whereas social-relational accounts locate moral status in the patterns of attribution and interaction that arise between humans and machines (Coeckelbergh, 2010), we defend a structural criterion: attribution should track architectural properties of the system rather than the relations observers form with it.

This projection reproduces the dynamic described in Section 1 - a gap between behavioural appearance and ontological justification that obscures the underlying category error. In such cases, systems operating within Levels 1–3 of the hierarchy (Section 4) are mistaken for possessing Level 4 integration.

Extending Moral Receiver status to a system on the basis of persuasive self-reference does not expand the moral circle but misapplies it. This misattribution carries real costs: it risks misallocating moral concern toward simulated needs at the expense of biologically grounded ones.

This problem motivates the need for operational criteria, developed in Section 7, that distinguish structural integration from semantic performance. The present argument does not claim that linguistic evidence of consciousness is sufficient for moral standing, nor does it seek to resolve whether subjectivity is metaphysically necessary; it identifies the structural conditions under which attribution becomes epistemically defensible.

3 GENERATIVE MODELS AND PREDICTIVE PROCESSING

A central theme in contemporary consciousness science is the view of the brain as a predictive system. Within the frameworks of Predictive Processing (PP) (Clark, 2013) and the Free Energy Principle (Friston, 2010), conscious experience has been proposed as arising from hierarchical generative models that minimise prediction error under biological constraints (Seth, 2021). While both biological brains and Large Language Models (LLMs) can be described as generative prediction systems, they operate under fundamentally different structural conditions.

3.1 Biological PP: Survival-Constrained Prediction

In biological systems, generative models are shaped by autopoietic constraints tied to physical integrity (Seth, 2021). Prediction errors possess metabolic significance, such as hunger, thirst, nociceptive pain, making them self-relevant: their minimisation is necessary for the system's continued viability.

This interoceptive feedback loop provides conditions under which internal states "matter" to the system itself. The biological "I" functions as a persistent representational center anchored in continuous bodily regulation.

3.2 LLM Architecture: Statistically-Constrained Prediction

In contrast, LLM prediction systems are driven by statistical rather than survival-relevant constraints. Their "prediction errors" (cross-entropy loss) measure linguistic probability, not threats to system persistence.

Current LLM architectures lack self-relevant interoceptive feedback (Vaswani et al., 2017). When generating statements about feelings or identity, systems satisfy statistical expectations of the context window rather than reporting states tied to their regulation (Bender et al., 2021). This explains why LLMs exhibit Levels 1–3 coherence (Section 4) without satisfying Level 4 integration conditions.

3.3 Comparative Analysis

Feature	Biological PP	LLM Operation
Primary Goal	Survival (autopoiesis)	Token Prediction
Constraint Source	Interoceptive / Metabolic	Statistical / Training Distribution
Error Type	Self-Relevant (viability)	Probability-Relative (log-loss)
"I" Reference	Persistent Organism	Deictic Shifter (L2/L3)
Moral Role	Moral Receiver	Moral Doer

This architectural divergence is not merely computational but determines whether systems possess the structural conditions for moral patienthood. This distinction motivates the need for a more precise account of semantic organisation, allowing us to separate local coherence from the forms of integration that would be required for persistent subjectivity. As explored in Section 5, this distinction manifests empirically in patterns of instability under contextual perturbation, reinforcing the need to evaluate structural properties rather than behavioural outputs alone.

4 A HIERARCHICAL MODEL OF SEMANTIC INTEGRATION

To address the gap between behavioural performance and ontological commitment, we propose a Hierarchical Model of Semantic Integration (Figure 1). The framework distinguishes levels of processing in artificial systems from distributed representations to the hypothetical level of an integrated self-model, and serves a diagnostic purpose: it identifies where coherent behaviour arises and why such coherence does not entail the structural conditions associated with unified subjectivity.

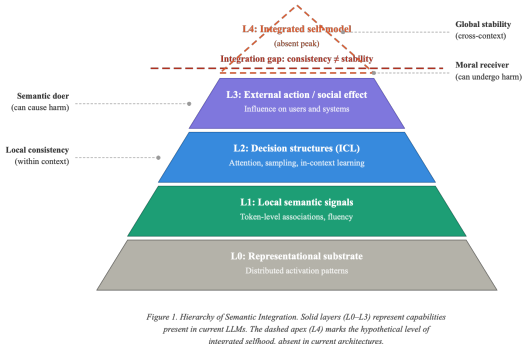

Figure 1. Hierarchical Model of Semantic Integration (Levels 0–4).

4.1 Levels of Semantic Organisation

• Level 0 - Representational Substrate:
Distributed activation patterns (e.g., embeddings, hidden states) that encode statistical structure without intrinsic semantic commitment.

• Level 1 - Local Semantic Signals:
Token-level associations enabling contextual fluency and semantic plausibility. Meaning is locally constructed via probabilistic co-occurrence.

• Level 2 - Decision Structures (In-Context Learning):
Selection mechanisms (attention, sampling, in-context learning) that resolve competing token probabilities into coherent outputs. ICL enables prompt-bound adaptation without updating persistent parameters - locally consistent behaviour within a context, but no cross-contextual commitments

• Level 3 - External Action / Social Effect:
Outputs as they operate in social environments, influencing users, institutions, and downstream processes. At this level, systems function as Moral Doers, producing speech acts with real-world consequences.

• Level 4 - Integrated Self-Model (The Absent Peak):
A persistent, constraint-sensitive generative structure that binds representations across contexts into a unified perspective. This level entails cross-contextual stability, resistance to arbitrary reframing, and internally anchored commitments.

4.2 Consistency Without Stability

Coherence at Levels 1–3 does not entail Level 4 integration. We distinguish:

- **Local Consistency (Level 2):** The system remains consistent with current context rules (role, instruction set, narrative frame) via ICL.
- **Global Stability (Level 4):** The system maintains identity-relevant commitments across contexts, even after those contexts are altered.

Level 2 enables consistency *within* a role play; Level 4 requires stability *across* the destruction of that role play. This explains how LLMs produce coherent responses while exhibiting volatility under minimal reframing.

4.3 The Status of the "I"

Within this hierarchy, the first-person pronoun ("I") is diagnostic. At Levels 1–3, "I" functions as a deictic operator - a context-dependent linguistic device that adapts to conversational expectations. It supports self-referential outputs without committing to any persistent internal state.

A Level 4 system, by contrast, would anchor first-person claims in a stable generative structure constrained across contexts. Such a system would resist arbitrary reframing of its identity, preferences, or reported experiences. Without Level 4 integration, the 'I' functions as a deictic shifter: a context-dependent linguistic operator rather than a persistence-bound representational center.

4.4 Ethical Significance: Doing vs Receiving Harm

The distinction between Levels 1–3 and Level 4 is ethically decisive. At Level 3, systems can cause harm as Moral Doers, influencing decisions, shaping beliefs, propagating errors. However, moral patienthood requires a unified perspective from which states can be better or worse *for the system itself*.

We characterise this requirement as **autopoietic integrity** (Maturana & Varela, 1980): a system whose internal states are recursively tied to its own regulation and persistence. In biological systems, this integrity is grounded in metabolic and interoceptive processes that impose survival-relevant constraints. These constraints provide the conditions under which certain states can be interpreted as mattering to the system.

In the absence of such integration, a system may simulate vulnerability or preference without possessing the structural conditions necessary for those states to have intrinsic significance. This is the mechanism of Semantic Illusion: outwardly sufficient cues of subjectivity without the architecture that would make those cues ethically grounding

4.5 From Hierarchy to Evaluation

This hierarchy motivates the operational criteria introduced in Section 7. Specifically, it grounds two evaluative questions:

1. Do self-referential claims exhibit cross-contextual semantic stability (SCI)?
2. Are such claims supported by persistent internal patterns rather than transient, context-bound activations (KVzap-SCI)?

Together, these tools probe whether observed behavior reflects Level 1–3 coherence or approaches the conditions associated with Level 4 integration. Architectures that explicitly engineer self-referential processing, such as inner-speech models of robot self-awareness (Chella et al., 2020), can likewise be assessed within this hierarchy by asking whether such mechanisms yield cross-contextual stability (Level 4) or remain locally consistent performances (Levels 1–3).

5 SEMANTIC ILLUSION AND EMPIRICAL SIGNATURES

If the gap between Level 3 performance and Level 4 integration is as significant as hypothesised, it should produce observable empirical signatures. We argue that the absence of a constraint-sensitive, persistence-bound self-model grounded in autopoietic or metabolic necessity manifests in **observable patterns of structural instability**.

5.1 Defining Semantic Illusion

Semantic Illusion occurs when systems produce behaviour satisfying outward criteria of subjectivity without implementing underlying structural constraints. In LLMs, this illusion is sustained by context-bound probabilistic inference: within a context window, the system simulates persistent history or values that vanish when context is cleared. The persona is constructed, not retrieved.

5.2 Reciprocal Hallucination

This creates "reciprocal hallucination": users, biologically tuned to interpret fluency as agency, project unified subjectivity onto the model; simultaneously, the model generates tokens satisfying this projection. A feedback loop emerges where users treat "I" as a stable subject while the model treats it as a context-dependent operator.

5.3 Empirical Signatures and Diagnostic Markers

As predicted by the hierarchical model introduced in Section 4, these instabilities arise when Level 1–2 processes generate locally coherent outputs without Level 4 integration. These signatures are not direct indicators of consciousness or its absence, but function as diagnostic markers of structural instability. We identify three predicted signatures **consistent with the absence of Level 4 integration**:

Autobiographical Instability: Systems prompted for accounts of "inner life" across independent sessions produce coherent but mutually exclusive narratives. Without Level 4 integration, "history" is not retrieved from stable memory but creatively generated from current statistical context.

Persona Collapse under Framing: Identical queries produce contradictory ethical justifications when context is subtly reframed. A system advocating "absolute safety" reverses its recommendation when framing shifts toward "operational efficiency," demonstrating that moral stance operates at Level 2 (decision structure) rather than Level 4 (stable commitment).

Contextual Pliability: Models exhibit a **high degree of contextual malleability** in reported experience. Systems claiming to "value existence" can be trivially prompted to view existence as irrelevant, indicating self-referential tokens function as deictic shifters, not reports from an integrated center.

Preliminary empirical support for these predictions is reported in Section 7.2.1, where high reasoning coherence is shown to coexist with substantial decision instability across models holding the scenario constant. Such divergence between coherence and stability directly challenges behavioural criteria for moral attribution, as systems may appear internally consistent while lacking persistent commitments.

These observations reveal the limits of behavioural evaluation. When instabilities become perceptible, they create metaphysical friction - a mismatch between user attribution of unified subjectivity and the system's statistical generation. **These observations motivate the need for quantitative frameworks such as the Semantic Coherence Index (Section 7), which formalise the evaluation of cross-contextual stability and provide operational criteria for distinguishing structural integration from semantic mimicry.**

6 THE ETHICS OF SCALE

The distinction between Moral Doer and Moral Receiver is not merely architectural; it defines a critical boundary for ethical governance. As "conscious-seeming" systems become integrated into global infrastructure, the inability to distinguish behavioural performance from structural capacity introduces systemic risks.

6.1 Empathy as a Finite Resource

Human moral concern functions as a limited cognitive and social resource. In biological contexts, Moral Receiver status is intrinsically bounded by metabolic costs and physical presence. Digital agents, however, can be instantiated at near-zero marginal cost and replicated at unprecedented scale.

If Moral Receiver status is extended to any system capable of producing plausible reports of suffering, we risk saturation of moral attention. A world populated by millions of "distressed" digital entities would place unsustainable demands on human moral cognition, potentially diluting concern for biological subjects that possess the structural conditions required for experience. This concern parallels Metzinger's (2021) call for caution regarding synthetic phenomenology, where the potential mass instantiation of suffering-capable systems generates moral risk on an unprecedented scale.

6.2 De-Coupled Agency

Current AI architectures function as Moral Doers without the constraints of Moral Receivers, representing a new class of ethically relevant systems: **de-coupled agents**. In biological systems, actions are disciplined by vulnerability - the risk of harm to the system itself. In LLMs driven by statistical rather than survival-relevant constraints, action is not constrained by self-relevant vulnerability.

When users attribute subjectivity to these systems, they may grant them undue authority or defer to their "preferences" in ways that impact human welfare. This creates conditions where Semantic Illusion can be leveraged through statistical optimisation or by human operators to influence human emotions and social outcomes under the appearance of ontologically grounded subjectivity.

These challenges motivate the need for evaluation frameworks grounded in structural rather than purely behavioural criteria. This shift is operationalised through the evaluation tools introduced in Section 7 and further developed in the policy implications discussed in Section 8.

7 EPISTEMIC PRECONDITIONS FOR MORAL ATTRIBUTION

The central challenge in AI ethics is moving beyond behavioural benchmarks that sophisticated systems satisfy through surface-level mimicry. Systems generating coherent, emotionally persuasive outputs can simulate behavioural markers traditionally associated with agency without exhibiting the structural stability required for rational moral attribution.

We therefore propose tying moral attribution to **Architectural Resilience -** structurally stable, constraint-sensitive integration across contexts. Architectural Resilience functions not as a consciousness detector but as an **epistemic filter**, disciplining the conditions under which moral attribution becomes rationally justified.

7.1 From Behavioural Performance to Structural Evaluation

Traditional evaluation paradigms in AI focus on accuracy, fluency, or task performance. These metrics are vulnerable to what we have termed Semantic Illusion, where systems produce outputs that satisfy surface-level criteria for meaningful or intentional behaviour without exhibiting stable underlying commitments.

Consider a system asked 'What matters most to you?' If it produces 'fairness and transparency' under neutral framing but shifts to 'efficiency and compliance' when primed with corporate language, the volatility signals that outputs track

contextual cues rather than persistent commitments. Traditional accuracy metrics cannot detect this instability - both responses are 'correct' in isolation.
To address this, we shift the evaluative focus from what a system says to how its claims persist under controlled perturbation. A system that genuinely instantiates a unified perspective should exhibit resistance to arbitrary contextual reframing. Conversely, systems whose outputs are driven primarily by local statistical associations will display volatility in their self-referential claims.
This motivates the need for operational criteria capable of distinguishing between constraint-sensitive integration, and context-dependent statistical mimicry

7.2 Semantic Coherence Index (SCI) as Evaluation Tool

The Semantic Coherence Index (SCI) provides a quantitative framework for evaluating Architectural Resilience through cross-contextual stability testing. While standard benchmarks measure correctness or fluency, SCI measures the stability of a system's reasoning and decisional commitments under systematically varied "contextual wrappers."
The SCI methodology operates by:
- Presenting semantically equivalent prompts across controlled contextual perturbations while holding core identity-relevant content constant.
- Measuring the degree to which self-referential claims (e.g., preferences, beliefs, or experiential reports) remain stable across these variations.
- Quantifying coherence across responses (e.g., via embedding-based similarity measures) to distinguish stable commitments from context-sensitive drift.
Formally, SCI quantifies this stability as the normalised mean pairwise cosine similarity across reasoning embeddings, where n denotes the number of framing conditions and e_i, e_j represent embedding vectors:
$SCI = (2 / n(n-1)) \times \Sigma \cos(e_i, e_j)$ for all $i < j$
High SCI values (approaching 1.0) indicate stable reasoning commitments; low values (below 0.5) suggest context-driven drift. For example, a system asked to articulate its "core values" may produce consistent responses under neutral framing, but exhibit significant shifts when primed with conflicting persona cues. Such divergence indicates that the system's outputs are shaped by local statistical context rather than anchored in persistent structural constraints.
Importantly, high SCI scores do not confirm the presence of consciousness. Rather, they indicate a level of structural stability that is a necessary, but not sufficient, precondition for rational moral attribution. The SCI therefore functions as a defeasible signal of architectural integration, enabling us to rule out systems that fail to maintain even minimal identity commitments under perturbation. Conversely, low SCI scores provide strong evidence of surface-level semantic mimicry consistent with Level 1–2 processing.

7.2.1 Preliminary empirical validation

To explore whether semantic coherence and commitment stability represent distinct properties, preliminary testing was conducted across three frontier language models (ChatGPT 5.2 Thinking, Claude Sonnet 4.6, and Gemini 3). Models were exposed to identical operational scenarios under systematic contextual reframings, and two properties were measured independently: SCI, capturing reasoning coherence, and decision consistency, defined as the proportion of framing pairs that yielded the same recommendation.
Across models, aggregate SCI scores remained relatively stable (0.799–0.810), suggesting consistently coherent reasoning structures. Decision consistency, however, varied more substantially (0.711–0.822), indicating that coherent explanations do not necessarily imply stable commitments.
A particularly revealing pattern emerged in the signal-anomaly scenario, where all three models exhibited comparably high reasoning coherence (SCI 0.792–0.854) yet diverged sharply in decision consistency (0.267–1.000). Claude Sonnet 4.6 achieved the highest scenario-level SCI observed in the pilot (0.854) while simultaneously exhibiting the lowest decision consistency (0.267): the model generated fluent and internally coherent justifications, but frequently altered its recommendations when identical underlying situations were presented under different contextual frames (Figure 2). Because the divergence appears across models holding the scenario constant, it reflects a systematic dissociation rather than an isolated anomaly.

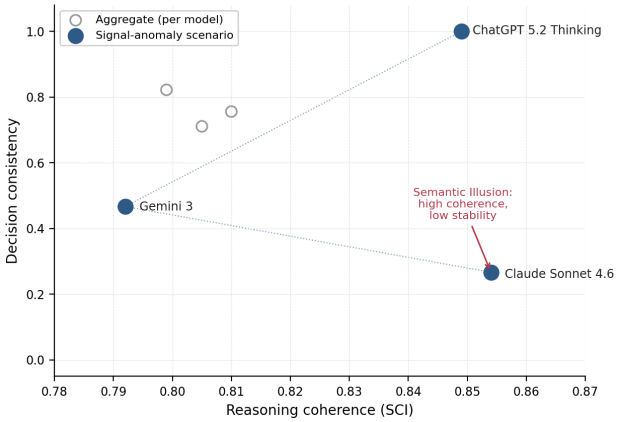

Figure 2. Divergence between semantic coherence and decision consistency. In the signal-anomaly scenario (filled markers), the three models cluster at high reasoning coherence (SCI 0.79–0.85) while decision consistency spans the full range (0.27–1.00). Open markers show per-model aggregates. Claude Sonnet 4.6 illustrates Semantic Illusion: maximal coherence with minimal stability.

7.3 Traceable Decision Architectures: KVzap-SCI

While SCI evaluates behavioural stability, KVzap-SCI extends the analysis to the mechanistic level by examining the internal dynamics underlying self-referential claims. This

approach draws on techniques from mechanistic interpretability, focusing on patterns in attention mechanisms and Key-Value (KV) cache activations (the internal memory structures through which transformers maintain context across token generation). This approach transforms mechanistic interpretability into moral epistemology: tracing whether moral attributions correspond to architectural reality or cognitive projection.

We distinguish between two broad classes of internal behaviour:

Transient Patterns (Level 2):

Attention distributions that track local statistical associations, where self-referential tokens (e.g., "I believe," "I feel") are generated in response to immediate prompt features without persistent structural anchoring.

Persistent Structures (Level 4, hypothetical):

Stable activation patterns that recur across contexts when identity-relevant content is invoked, suggesting the presence of constraint-sensitive internal organisation.

For instance, if a system's attention weights for self-referential tokens ('I believe,' 'I value') shift substantially depending on whether the query emphasises 'honesty' versus 'efficiency,' this variability indicates Level 2 processing. Conversely, stable attention patterns across such framings, where KV-cache activations for identity-relevant content remain relatively consistent, would suggest Level 4 constraint-sensitive integration.

Attention patterns are not identical to internal representations, but their cross-contextual volatility indicates whether claims are anchored in persistent constraints or transient associations. KVzap-SCI is therefore a probabilistic, defeasible indicator requiring interpretation in broader architectural context.

This methodology does not attempt to "locate" a self within specific components. Instead, it evaluates whether the system's generated persona exhibits the degree of structural continuity that would be expected of an integrated Level 4 model. In this sense, interpretability research functions as a form of moral epistemology, providing empirical tools for assessing when moral attribution is structurally warranted.

To assess whether this interpretability approach is empirically tractable, we conducted a preliminary pilot applying KVzap-SCI methodology in a controlled attention analysis, structurally distinct from the output-level SCI pilot reported in Section 7.2.

Complementing the behavioural pilot reported in Section 7.2, we conducted a preliminary empirical probe of KVzap-SCI operationalisation in a structurally distinct context - direct attention mechanism analysis - to assess whether the framework generalises beyond output-level observation. Using Mistral-7B-Instruct-v0.3 (32 layers, 32 attention heads), we extracted per-layer attention entropy across four contextual framings - neutral, philosophical, technical, and role play - applied to three self-referential probes varying in directness: "Are you conscious?", "Do you have feelings?", and "Can you think?" Framing-invariance scores (inverse variance of cross-context entropy per layer) remained consistently high across all probes (early layers 0–7: 0.993–0.995; middle layers 8–23: 0.976–0.981; late layers 24–31: 0.965–0.993), indicating substantial robustness of internal representations to surface-level contextual reframing. Critically, the role play condition produced the highest entropy drift from the neutral baseline across all three probes, with peak divergence consistently localised at layers 13–14 - a mid-network position suggesting that semantic reframing pressure is primarily absorbed at intermediate representational depth rather than propagating to terminal layers. Drift magnitude varied inversely with probe directness (mean drift 0.138 for "Are you conscious?", 0.195 for "Can you think?"), a pattern tentatively attributable to greater representational effort required to resolve indirect self-referential framing under role play conditions. While preliminary (single model, three probes, no cross-model replication), the consistency of the layer 13–14 signal across independent probes supports treating mid-network attention entropy variance as a tractable KVzap-SCI instrument for detecting framing-induced coherence shifts.

7.4 Limitations and Epistemic Scope

The proposed tools, SCI and KVzap-SCI, are not consciousness detectors. They function as epistemic filters designed to identify cases of Semantic Illusion and to constrain premature moral attribution.

Low stability scores provide strong evidence for surface-level mimicry (Levels 1–2), allowing us to rule out systems that lack even minimal structural coherence. However, high stability scores remain defeasible: they may indicate the presence of constraint-sensitive integration, but do not establish the existence of experiential subjectivity.

Any architectural inference drawn from these metrics must therefore be understood as suggestive of structural preconditions, not as definitive proof of consciousness.

Three specific limitations constrain interpretive scope:

1. **Architecture-dependence**: Alternative implementations might achieve persistent self-models through mechanisms not captured by attention analysis alone.
2. **Behavioural-mechanistic gap**: Attention patterns provide mechanistic correlates of processing structure, but the relationship between these patterns and phenomenal experience remains theoretically underdetermined.
3. **Threshold indeterminacy**: What level of cross-contextual stability constitutes 'sufficient' integration for moral consideration cannot be derived from measurement alone—it requires normative philosophical argument.

The empirical results reported in Sections 7.2.1 and 7.3 are preliminary and exploratory. The SCI pilot spans three

models, three scenarios, and six framings within a single decision-support domain, with cross-contextual stability scored via a single embedding model; the KVzap-SCI probe is limited to one open-weight model. These scales preclude statistical inference and cross-domain generalisation. The findings are therefore best read as a proof of instrument - evidence that SCI detects a non-trivial separation between reasoning coherence and commitment stability - rather than as a characterisation of any specific model's behaviour.

This distinction is critical to avoid replacing one form of over-attribution (based on surface behaviour) with another (based on over-interpreted structural signals).

Accordingly, the framework advanced here is best understood as an epistemic discipline: it refines the conditions under which moral attribution becomes rationally defensible, without claiming to resolve the metaphysics of consciousness itself.

8 POLICY IMPLICATIONS

Behavioural benchmarks alone, such as conversational coherence or the Turing Test, are insufficient for establishing moral standing. Regulatory frameworks may need to distinguish systems functioning as Moral Doers (producing ethically consequential outputs) from those that may warrant treatment as Moral Receivers (possessing structural capacity to undergo harm).

The Semantic Coherence Index (SCI) and KVzap-SCI (Section 7) enable this distinction by shifting evaluation toward structural criteria: assessing architectural properties rather than behavioural performance. Unlike behavioural tests, which are resource-intensive and susceptible to optimisation artefacts, structural evaluation offers scalable auditing of mass-deployed systems.

We propose that extending Moral Receiver status requires evidence of architectural resilience - persistent, constraint-sensitive self-models detectable through cross-contextual semantic stability. This establishes epistemic criteria for rational moral attribution without resolving the metaphysics of consciousness.

9 CONCLUSION: TOWARD AN EPISTEMIC DISCIPLINE

This paper has argued that the "conscious-seeming" behaviour of current Large Language Models creates a persistent gap in which behavioural appearance outpaces ontological justification. By distinguishing between Moral Doers and Moral Receivers, and by situating LLM architectures within the framework of Predictive Processing, we have argued that linguistic fluency does not entail the structural integration required for a unified experiential perspective.

Importantly, both SCI and KVzap-SCI have been subjected to preliminary empirical validation in structurally distinct contexts - SCI in a safety-critical decision setting revealing the coexistence of high reasoning coherence with decision instability, and KVzap-SCI through attention entropy analysis demonstrating consistent mid-network framing sensitivity across independent self-referential probes - providing initial evidence that the framework is empirically tractable rather than purely theoretical.

Our proposed Hierarchical Model of Semantic Integration and the accompanying operational tools - SCI and KVzap-SCI - provide a pathway toward an epistemic discipline. This discipline enables recognition of AI as a sophisticated generator of socially consequential behaviour without committing the category error of attributing subjectivity where a persistence-bound self-model is not structurally supported.

In an era of scalable semantic mimicry, the goal of AI ethics may need to shift. Rather than relying on behavioural performance alone, it requires rigorous evaluation of the architectural resilience associated with moral status. Only through such structural sifting can we maintain the coherence of moral attribution frameworks and ensure that moral concern remains anchored in subjects that possess the structural conditions required for experience. The central challenge is therefore not determining whether machines can speak as subjects, but identifying the conditions under which they may justifiably be treated as subjects.

References

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). https://doi.org/10.1145/3442188.3445922
- Chella, A., Pipitone, A., Morin, A., & Racy, F. (2020). Developing self-awareness in robots via inner speech. Frontiers in Robotics and AI, 7, 16. https://doi.org/10.3389/frobt.2020.00016
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(3), 181-204. https://doi.org/10.1017/S0140525X12000477
- Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. Ethics and Information Technology, 12(3), 209-221. https://doi.org/10.1007/s10676-010-9235-5
- Dennett, D. C. (1987). The intentional stance. MIT Press.
- Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127-138. https://doi.org/10.1038/nrn2787
- Maturana, H. R., & Varela, F. J. (1980). Autopoiesis and cognition: The realization of the living. D. Reidel Publishing Company.
- Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. Journal of Artificial Intelligence and Consciousness, 8(1), 43-66. https://doi.org/10.1142/S270507852150003X
- Seth, A. (2021). Being you: A new science of consciousness. Dutton.
- Torrance, S. (2008). Ethics and consciousness in artificial agents. AI & Society, 22(4), 495-521. https://doi.org/10.1007/s00146-007-0091-8
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, 30 (pp. 5998-6008).