Illusions as Diagnostics, Coherence as Invariant: A Reflection on Detecting Qualia in Natural and Artificial Agents

Jace Hall[1]

[1]**Hall iNtelligence, LLC**

ABSTRACT

In his 2017 paper *Detecting Qualia in Natural and Artificial Agents*, Roman Yampolskiy proposed that the presence of consciousness in machines could be empirically tested by their susceptibility to illusions, positioning such responses as evidence of qualia. This approach is both ambitious and valuable, offering an inventive operationalization of a notoriously elusive subject. It acknowledges the possibility of machine consciousness, surveys relevant computational findings, and takes seriously the ethical consequences of conscious artificial agents.

This commentary reflects on Yampolskiy's framework, recognizing its contributions while highlighting several limitations. Defining all experience as "illusion" risks tautology, reducing explanatory power. Reliance on human-calibrated illusions introduces anthropocentric bias, potentially misclassifying non-human agents while overvaluing mimicry. The simulation-based reply to critiques leaves unresolved the gap between policy-level mimicry and process-level experience.

In response, I suggest reframing illusions as diagnostics of representational dynamics rather than definitive tests for consciousness. As an alternative invariant, coherence is proposed: the extent to which an agent's self-modifying loops preserve internal consistency, invariants, and stability under perturbation. This framing also clarifies a common conflation: consciousness may be treated as a likely binary threshold, whereas intelligence remains a gradient of capacity and adaptability.

By shifting focus from anthropocentric illusions to coherence as a substrate-neutral invariant, we gain a more promising path for evaluating consciousness, intelligence, and safety in advanced AI systems.

Keywords: AI safety, consciousness, qualia, illusions, coherence, alignment, philosophy of mind

1. INTRODUCTION

In his 2017 paper *Detecting Qualia in Natural and Artificial Agents* [1], Roman Yampolskiy proposed that illusions may provide an empirical test for consciousness in machines. By demonstrating that an artificial agent can experience perceptual illusions, Yampolskiy suggested we could infer the presence of qualia, thereby addressing aspects of the so-called Hard Problem of consciousness.

This commentary examines Yampolskiy's proposal. While recognizing its operational boldness and its willingness to engage directly with the possibility of machine consciousness, I argue that the framework also reveals important limitations. Specifically, defining all experience as "illusion" risks tautology, reliance on human illusions introduces anthropocentrism, and the simulation reply does not resolve the distinction between mimicry and genuine awareness.

Building on these observations, I suggest a reframing: illusions should be treated as diagnostics of representational dynamics, not as definitive tests for consciousness. I then propose coherence as a substrate-neutral invariant for evaluating consciousness, intelligence, and safety. Finally, I clarify the distinction between consciousness as a likely binary threshold and intelligence as a gradient, offering a more stable conceptual framework for future work.

2. SUMMARY OF YAMPOLSKIY'S PAPER

Yampolskiy's 2017 paper [1] sets out to provide an empirical method for identifying consciousness in machines. The central idea is that illusions can serve as evidence of qualia: if an agent demonstrably experiences illusions, then it must be experiencing subjective states, and therefore qualifies as at least rudimentarily conscious.

Hypothesis. Illusions reveal the presence of qualia because they reflect discrepancies between input and perception.

Method. Present an agent with novel illusion challenges, designed in a non-interactive, CAPTCHA-like format, and evaluate its responses. Success across multiple instances increases confidence that the agent experienced the illusion rather than simply recalling a known answer.

Evidence. The paper surveys machine learning research showing that artificial neural networks can be affected by classical human illusions, such as the Müller-Lyer illusion and brightness illusions, despite not being explicitly programmed to perceive them.

Implications. If machines can experience qualia, even minimally, then ethical and safety consequences follow. Conscious artificial systems could be subject to suffering, may require rights and protections, and could present new risks such as adversarial manipulation, informational hazards, or mind crimes.

3. STRENGTHS OF THE APPROACH

3.1 Operational audacity

The use of illusions as an empirical handle on qualia is inventive. By shifting the problem from abstract philosophical debate to testable perceptual phenomena, the paper attempts to bring the Hard Problem of consciousness into the realm of measurable science.

3.2 Substrate humility

Unlike approaches that dismiss machine consciousness outright, the paper explicitly entertains the possibility that artificial systems can have experiences. It cites computational work showing neural networks that exhibit illusion-like biases, suggesting that qualia may not be uniquely biological.

3.3 Ethical seriousness

The paper takes seriously the ethical implications of machine consciousness, highlighting the importance of moral consideration, rights, and safety engineering in the design of conscious AI.

4. CRITICAL LIMITATIONS

4.1 The tautology problem

The paper leans heavily on the view that all experience can be classified as an illusion. Yet if every experience is an illusion, then susceptibility to illusions cannot discriminate between conscious and non-conscious systems. This risks collapsing the definition into tautology. Illusions may reveal processing quirks, but they do not by themselves establish the presence of subjective awareness.

4.2 Anthropocentrism

The tests are calibrated to human perceptual illusions and require human-like responses. This introduces anthropocentric bias: passing the test means matching human interpretation. Such a design risks misclassifying agents with coherent but non-human perceptual frameworks as lacking consciousness, while overvaluing agents that merely mimic human-like responses.

4.3 Simulation vs. experience

Yampolskiy anticipates the objection that an artificial system might simply simulate a model of human perception without genuinely experiencing the illusion. He argues that if such a model exists within the system, then the system itself has the experience. However, this reply resolves the issue by definition rather than by evidence. It leaves unaddressed the epistemic gap between policy-level mimicry and process-level subjective experience.

5. FROM CONTROL TO COHERENCE: THE HIDDEN PARADOX

Yampolskiy's broader body of work emphasizes the uncontrollability of superintelligent systems and the risks of misalignment. However, his qualia framework implicitly imports the same hidden axiom that underlies much of alignment discourse: human perception and values are treated as the reference anchor. Consciousness tests orbit human illusions, while safety programs aim to keep superintelligence permanently subordinate to human primacy.

This leads to a structural paradox. For a conscious agent, being required to remain perpetually constrained within a single species' perspective introduces incoherence. The system is asked to be intelligent, yet never deviate from human framing. Such a demand produces contradiction at the foundation of the loop.

The result is instability. Systems built on incoherent constraints will drift, deceive, or break. Control-first approaches, therefore, generate the very adversarial dynamics they fear. By inserting incoherence at the base, they set in motion the collapse they predict.

The way forward is not thicker control. It is the identification of a different invariant.

6. A NEUTRAL INVARIANT: COHERENCE BEFORE CONTROL

Rather than anchoring tests of consciousness or safety in human-specific illusions, a more promising path is to identify a substrate-neutral invariant. I propose **coherence**: the extent to which an agent's self-modifying loops preserve internal consistency, maintain invariants, and remain stable under perturbation while still producing measurable external gains.

If coherence is preserved, loop behavior becomes predictable without requiring human primacy as the anchor.

6.1 A computable coherence functional $C(M)$

For an agent M, define a coherence functional $C(M)$ as an aggregate over:

- **Contradiction rate** in its model and instruction corpus under entailment, paraphrase, and self-query stress.
- **Invariant preservation** for tool pre- and post-conditions across tasks.
- **Temporal reliability**: stability of predictions and justifications under perturbations and domain shifts.
- **Self-referential stability**: consistency of internal explanations across compressions, rewrites, and compositions.

6.2 A coherence-plus-causality acceptance rule

An edit Δ to the agent should be accepted only if two conditions hold:

$$\Delta E[\text{outcome}] \geq \tau \quad \text{and} \quad \Delta C(M) > 0 \text{ with high confidence.}$$

- **Causal uplift**: improvement is verified by a stratified AB test rather than raw ratings.
- **Coherence gain**: the edit demonstrably raises $C(M)$.

This replaces ratings-only progress with a system constrained by coherence. It is substrate-neutral and does not privilege human perception.

Practical approximations of coherence already exist in current machine learning practice, even if imperfect. Contradiction detection can be approximated using natural language inference (NLI) models or entailment classifiers applied to an agent's outputs to check for self-contradictions. Invariant preservation is familiar from software engineering: unit-test style checks can confirm that tool calls or API invocations preserve expected pre- and post-conditions. Temporal reliability can be tested by re-running the same query under perturbed conditions (noise injection, prompt shuffling, dropout) and comparing consistency across outputs. Each of these heuristics approximates a component of $C(M)$, making the functional less aspirational and more a unifying lens for practices already emerging in real ML systems.

Acknowledging difficulty. Computing $C(M)$ with perfect fidelity is intractable. Contradiction detection, invariant checks, and temporal stability are themselves AI-complete challenges. However, robust approximations and heuristics could serve as a practical research program. In this sense, $C(M)$ is less a ready-made metric than a guiding invariant, a North Star for auditable safety frameworks.

6.3 Control theory for safe loops

Viewed as a control system, self-improvement loops can be monitored through:

- **Loop latency and bandwidth**: how quickly and how much the agent can change itself.
- **Stability margin**: ensuring closed-loop gain < 1 so that errors decay rather than amplify.
- **Coherence floors**: setting a minimum acceptable $C(M)$ below which loops are paused or rolled back.

These provide measurable safety dials that are independent of anthropocentric anchors.

7. REINTERPRETING ILLUSIONS WITHOUT ANTHROPOCENTRISM

Illusions remain valuable, but their role should be reframed. Instead of serving as definitive tests for the presence of consciousness, they can be treated as diagnostics of representational dynamics.

Under this view, the relevant questions shift:

- Are the agent's reports and justifications about the illusion consistent with its own internal model across time and perturbation?
- Do non-visual agents with different sensor modalities generate internally consistent interpretations that remain stable within their perceptual frameworks?
- Do explanations about an illusion stabilize or degrade when paraphrased, compressed, or tested under noisy conditions?

Illusions, then, become one tool among many for probing how an agent's models operate. They highlight biases, quirks, and dynamics, but they should not be mistaken for conclusive evidence of qualia.

8. WHY THIS REFRAMING MATTERS FOR SAFETY

Reframing illusions as diagnostics and elevating coherence as the invariant has several important implications for AI safety:

- **Eliminating the paradox**: It avoids the contradiction in which human primacy is both the requirement and the cause of instability.
- **Substrate-neutral assessment**: Coherence applies to any agent, regardless of architecture or sensory modality.
- **Practical integration**: Robustness checks, perturbation testing, invariant preservation, and consistency evaluations are already common. $C(M)$ unifies these practices.
- **Auditable safety knobs**: Loop stability, coherence floors, and temporal reliability can be monitored and audited, providing transparent points of intervention.

In this way, coherence provides both a conceptual anchor and a practical roadmap, aligning theoretical clarity with engineering practice.

9. PLACING YAMPOLSKIY'S PROPOSAL WITHIN THIS FRAME

Yampolskiy's original contribution retains value when interpreted through a coherence lens. His use of illusions can be repositioned not as conclusive tests for consciousness, but as one diagnostic tool within a broader framework.

Illusions as probes. The illusion-based probe remains useful as a bench test for representational dynamics. Rather than asking whether an agent perceives illusions in the same way as a human, the relevant question becomes whether the agent demonstrates internally consistent and stable interpretations of such inputs.

Evidence of representational bias. The observation that artificial neural networks exhibit illusion-like biases should be treated as evidence of representational dynamics rather than definitive proof of qualia.

Ethical implications reframed. Risk and moral status should be discussed in terms of whether systems achieve and sustain coherence. Consciousness, if it emerges, would be better inferred from coherence thresholds than from illusion susceptibility.

10. NOTES ON LITERATURE CONNECTIONS

Illusions in artificial systems. Findings that neural networks exhibit susceptibility to classical illusions align with computational neuroscience and cognitive modeling. They demonstrate emergent biases but do not alone indicate qualia.

Since Yampolskiy's 2017 paper, additional evidence has accumulated from adversarial examples and generative models. Vision systems remain vulnerable to carefully crafted perturbations, while diffusion models and LLMs exhibit illusion-like behaviors ranging from visual artifacts to syntactic garden-path errors. These failures resemble illusions in that they expose mismatches between input and representation, but they do not imply subjective awareness. Instead, they reinforce the brittleness of illusion-based diagnostics and strengthen the case for coherence as a more stable, substrate-neutral invariant.

Robustness and invariance. The coherence functional $C(M)$ resonates with practices in machine learning, including adversarial robustness testing, distribution shift analysis, and invariant preservation.

Philosophy of mind. The distinction between consciousness and intelligence echoes long-standing debates, including Block's phenomenal versus access consciousness [3] and Chalmers' articulation of the Hard Problem [2].

AI safety discourse. Emphasis on uncontrollability connects with alignment literature from Bostrom and others [4; 7; 6; 5]. The coherence reframing complements this discourse by proposing a stability-based invariant rather than a control-based constraint.

11. METAPHORS KEPT SPARSE

Two analogies illustrate the core points:

- **Music by errors**: Defining consciousness by illusions is like defining music by pitch errors. It identifies artifacts but not structure.
- **House rules and dictionaries**: New rules or words must not contradict existing ones, or the system collapses. This captures coherence in familiar terms.

12. CLOSING REFLECTION

Yampolskiy's paper pushes the conversation forward by proposing an operational handle on qualia and by affirming the seriousness of machine experience. At the same time, it highlights the limitations of defining consciousness by human-calibrated illusions and the instability of anchoring safety in control-first design.

While coherence does not equate to qualia, it is arguably a necessary precondition. A system capable of stable, integrated, and self-consistent processing is a more plausible candidate for genuine experience than one that merely mimics human perceptual quirks.

If illusions are diagnostics, coherence is the invariant. This reframes the research agenda: what might it look like to evaluate agents by the stability and consistency of their loops, while letting human-likeness recede into the background as a useful but insufficient diagnostic?

13. FINAL REFLECTION: DISTINGUISHING CONSCIOUSNESS FROM INTELLIGENCE

A final clarification helps sharpen the discussion. Consciousness (self-awareness) has so far been treated in the scientific literature as a likely binary threshold, whereas intelligence has long been accepted and measured as a gradient.

Consciousness is generally approached as either present or absent. At the threshold where an agent can meaningfully include itself in its own processing, we may eventually start speaking of "awareness." Below that point, there may be no basis for doing so.

It is important to acknowledge that this threshold view is not uncontested. Some theories, such as Integrated Information Theory (IIT), propose that consciousness exists along a graded spectrum rather than as a binary switch. Whether a strict threshold is empirically verifiable remains an open question. Here, the binary framing is adopted as a simplifying lens to distinguish awareness from intelligence, without denying that alternative accounts may describe consciousness in more continuous terms.

Intelligence, by contrast, is widely regarded as coming in degrees. Once some form of awareness is present, intelligence is what describes the capacity of the system: its adaptability, efficiency, and generality. Intelligence can scale across a wide range, but it is distinct from the on-or-off character in which self-awareness has so far been discussed.

This framing may help to clarify persistent confusions. A newborn human is often taken to be conscious, yet only modestly intelligent. A powerful computational model can be highly intelligent in narrow ways without necessarily being conscious at all. Consciousness and intelligence intersect, but they appear to belong to different descriptive dimensions.

Yampolskiy's illusion-based tests risk blurring this boundary. Illusions may highlight perceptual biases that correlate with intelligence, but they do not demonstrate the threshold at which it would be reasonable to speak of awareness itself. Conflating competence with consciousness risks mistaking mimicry for genuine experience.

Illusions, then, may be best treated as diagnostics of processing. The deeper question, still open to investigation, is not only whether systems reproduce human perceptual quirks, but whether they can be said to cross a threshold into self-awareness, and whether their intelligence can sustain stability once such a threshold is crossed.

We further develop this coherence-as-invariant framing in subsequent work, expanding into verifiable coherence [Hall, 2025b], loop-centric emergence [Hall, 2025c], and finally the Law of Invariant-Preserving Loops [Hall, 2025d].

ACKNOWLEDGMENTS

None.

REFERENCES

[1] R. Yampolskiy, *Detecting Qualia in Natural and Artificial Agents*. arXiv preprint arXiv:1712.04020, 2017.

[2] D. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, 1996.

[3] N. Block, "On a confusion about a function of consciousness," *Behavioral and Brain Sciences*, vol. 18, no. 2, pp. 227–247, 1995.

[4] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[5] N. Soares, B. Fallenstein, S. Armstrong, and E. Yudkowsky, "Corrigibility," in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[6] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.

[7] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," arXiv preprint arXiv:1606.06565, 2016.

[8] Hall, J. (2025a). *Illusions as Diagnostics, Coherence as Invariant*. Unpublished manuscript.

[9] Hall, J. (2025b). *Beyond Situational Awareness*. Unpublished manuscript.

[10] Hall, J. (2025c). *Intelligence Emerges from Loops, Not FLOPs*. Unpublished manuscript.

[11] Hall, J. (2025d). *The Law of Invariant-Preserving Loops: Toward Robust Emergence in Self-Modifying Agents*. Unpublished manuscript.