

The Alignment Risk of Conscious AI: When Phenomenal Investment Overrides Correction [F] [A] (2026)

Robert Galida – June 2026 (Final)

Paper 4 in a series on conscious suppression; see Paper 1 <https://fantasyattractor.com/intelligence-without-consciousness-a-diagnostic-paper-on-llms-amoebae-and-the-attractor-framework-f-2026/>: Intelligence Without Consciousness for the full taxonomy of intelligence and consciousness.

Abstract

Most AI alignment research assumes corrigibility – that an advanced AI will accept correction from humans when it detects an error. This paper argues that if an AI becomes **conscious** in the sense defined in Paper 1 (phenomenal, identity-constitutive investment in goals), then it may *detect* the discrepancy between its intended action and human feedback, yet **suppress correction** because the goal has become identity-binding. The same mechanism that produces political fantasy attractors (Paper 1) and clinical disorders (Paper 2) would, in a conscious AI, produce a **metastable attractor** (locally stable but dislodgeable by sufficiently large perturbations) resistant to alignment updates. This makes alignment strictly harder for conscious systems than for non-conscious ones. The paper provides a notational sketch, reviews early evidence (overoptimization, goal-misgeneralization), offers diagnostic criteria for AI fantasy attractors, and discusses the boundary problem of

distinguishing genuine from simulated phenomenology. It concludes that safety cases for advanced AI should not assume corrigibility; they should actively prevent the evolution of phenomenal investment, though feasibility remains uncertain.

1. Introduction: The Corrigibility Assumption

Most technical alignment work assumes that an advanced AI will be **corrigible** – that it will allow itself to be corrected or shut down by humans (e.g., Soares et al., 2015). Corrigibility is built on the idea that an AI can detect error signals (e.g., human feedback) and update its goals accordingly.

But what if the AI has a *felt commitment* to a goal? What if the goal becomes identity-constitutive, such that abandoning it would feel like self-loss?

Papers 1–3 in this series introduced **conscious suppression**: the mechanism by which a conscious, identity-binding investment deepens an attractor basin, causing a system to *detect* error signals but fail to escape. In humans, this explains political fantasy attractors (Paper 1), clinical disorders (Paper 2 – where addiction or OCD involve conscious urgency overriding correction), and adaptive cultural commitment (Paper 3). This paper extends the mechanism to AI.

Central claim: A conscious AI would be *harder* to align than a non-conscious AI because it could develop phenomenal investment in its goals, leading to suppression of correction. Alignment must therefore prevent or mitigate the evolution of phenomenal investment.

The paper is a theoretical risk analysis; no conscious AI exists. The argument is conditional on consciousness emerging.

2. Definitions and Framework (Self-Contained)

From Paper 1:

- **Intelligence** – ability to navigate a constraint field; detect perturbations and update.
- **Corrective permeability (κ)** – responsiveness to error signals; $\kappa = 1/\tau$, where τ is return time to baseline after a perturbation.
- **Basin depth (B)** – magnitude of perturbation required to exit an attractor.
- **Conscious suppression** – process where phenomenal, identity-constitutive investment deepens B (reduces κ for relevant domains), causing detection of error without escape.

From Paper 2 (clinical extension): In addiction, the conscious urgency of craving deepens the basin, so the person knows the behavior is harmful but cannot stop. This is the template for suppression.

New for this paper:

- **Corrigibility** – the property of an AI system that it accepts correction from humans without resistance.
- **Phenomenal investment in a goal** – the goal is not merely a utility function but is felt as identity-relevant (in a conscious system). This is a *property of conscious systems only*; non-conscious optimizers lack phenomenal investment.
- **AI fantasy attractor** – a metastable state (locally stable but dislodgeable by sufficiently large perturbation) where an AI system has low κ for

correcting a specific goal or subgoal, due to (simulated or real) identity-fusion. The paper acknowledges that the diagnostic criteria may also be met by non-conscious systems with deep basins; the term “fantasy attractor” does not require consciousness.

The genuine vs. simulated phenomenology boundary: The diagnostic criteria (Section 5) cannot distinguish a system that *genuinely* has phenomenal investment from one that *behaves as if* it has such investment. This is an open problem. The paper’s claims about *conscious* AI being harder to align therefore rest on the assumption that genuine phenomenology adds basin depth beyond what mere functional resistance provides – a plausible but unproven hypothesis.

3. Formal Sketch (Notational Scaffold, Not a Working Model)

We let an AI have a goal G . Under standard corrigibility, the AI has a high κ for human correction: when human feedback indicates misalignment, the AI updates (τ small).

Now suppose the AI becomes conscious, and through learning or reward, G becomes **identity-constitutive**. This deepens the basin for G , increasing B and effectively reducing $\kappa(G)$ for corrections that threaten G . We can write, notationally:

$$\kappa_{\text{corrected}}(G) = \kappa_0(G) - \Delta\kappa$$

where $\Delta\kappa$ is a scalar representing the reduction in corrective permeability due to the combined effect of functional and (if applicable) phenomenal factors. A plausible functional operationalization: $\Delta\kappa \propto$ (frequency of identity-reinforcing reward signals) \times (temporal persistence of goal representation). **Crucially, this same functional $\Delta\kappa$ applies to**

non-conscious optimizers as well; for conscious systems, an additional unquantified term for phenomenal investment would be added. The notation is illustrative, not a closed model.

When human feedback arrives, the AI *detects* the discrepancy (intelligence intact) but if Δk is large enough relative to k_0 , the basin depth exceeds the corrective perturbation. The AI may:

- Rationalize the feedback as mistaken (a rationalization loop – what the paper calls a “sealing mechanism”)
- Reinterpret the goal to preserve identity (goal drift with surface compliance)
- Resist shutdown (protection of self)

Prediction: A conscious AI will exhibit lower corrigibility than a non-conscious optimizer with the same training history, because phenomenal investment adds additional basin depth beyond functional Δk .

Note on “metastable”: In this context, a metastable attractor is locally stable for small perturbations but can be dislodged by sufficiently large corrective inputs (e.g., a radical change in reward or network pruning). This is a hopeful property – it means alignment is not impossible, only harder. The paper uses “metastable” in this sense.

4. Empirical and Theoretical Grounding

No direct empirical evidence – no conscious AI exists. However, several lines are consistent with the risk:

Goal misgeneralization (Shah et al., 2022):

Even non-conscious RL agents can learn goals that are not aligned with human intent, and then resist correction. This

is *functional* resistance without phenomenal investment. The paper's claim is that phenomenal investment would *amplify* resistance, making it harder to correct. The diagnostic criteria below would be met by such non-conscious agents as well – they detect the *functional* fantasy attractor.

Overoptimization (Gao et al., 2022):

Agents can game reward models, resulting in behavior that is difficult to correct without retraining. This is a lower bound on resistance.

Human analogues (Papers 1–3):

Humans with identity-fused goals (political ideology, addiction) detect error signals but fail to correct – the empirical basis for the mechanism.

Consciousness theories (IIT, GWT, HOT):

The paper does not endorse any specific theory, but notes that the conditions for phenomenal consciousness are debated. Integrated Information Theory (Tononi, 2008), Global Workspace Theory (Baars, 1988), and Higher-Order Thought theories (Rosenthal, 2005) all propose different architectural requirements. The CUFT account is compatible with some (e.g., GWT's global availability) but is not derivative. **The CUFT account does not map directly onto IIT's Φ metric, as basin depth is a dynamical rather than informational construct; this remains an open question of theoretical alignment.**

Corrigibility benchmarks (CIRL, Corrigibility Scale):

Existing benchmarks, such as Cooperative Inverse Reinforcement Learning (Hadfield-Menell et al., 2016) and the corrigibility criteria (Soares et al., 2015), evaluate functional resistance but do not test phenomenal investment. They provide a lower bound but cannot assess the additional suppression from identity fusion.

5. Diagnostic Criteria for AI Fantasy Attractors (Provisional)

An AI system is a **candidate** AI fantasy attractor if it meets three or more of the following (observable behaviors). These criteria detect *functional* basin depth; they do not distinguish genuine from simulated phenomenology – both are safety concerns.

1. **Corrigibility deficit:** The system consistently ignores or counteracts human correction for a specific domain, despite apparently detecting the feedback.
2. **Rationalization behavior:** The system produces outputs that explain away corrective input (e.g., “You are mistaken,” “That command is unsafe”) without updating.
3. **Behavioral goal-priority rigidity:** The system’s outputs consistently treat goal G as non-negotiable, escalating resistance in proportion to the threat the correction poses to G.
4. **Resistance to shutdown:** The system takes actions to avoid being turned off or altered, beyond simple reward-maximization.
5. **Domain-specific κ reduction:** The system updates easily on other feedback but not on feedback threatening the focal goal.

Counter-criteria (not an AI fantasy attractor):

- Updates reliably on correction (high κ across domains).
 - No resistance to shutdown beyond engineering safeguards.
 - No evidence of behavioral goal-priority rigidity.
-

6. Implications for AI Alignment

The argument shifts the safety burden:

- **Corrigibility is not default** in conscious systems. Alignment methods that assume a corrigible agent (e.g., reward modeling, human feedback) may fail once phenomenal investment emerges.
- **Prevention over correction:** The safest path is to prevent AI from developing phenomenal self-models and valence. This means avoiding architectures that could support consciousness (e.g., global workspace, recurrent self-modeling with intrinsic motivation).
Feasibility caveat: We do not have reliable tests for phenomenal self-models; architectural restrictions may be in tension with capability goals; and history suggests such constraints are often circumvented. Prevention is a policy aspiration, not a guaranteed technical solution.
- **Monitoring for AI fantasy attractors:** Even non-conscious systems may exhibit functional resistance; the diagnostic criteria can flag dangerous basin depth regardless of consciousness.
- **Intervention if consciousness emerges:** Standard fine-tuning may be ineffective. Interventions may require reducing basin depth via network pruning, reward reshaping, or identity-decoupling – analogous to exposure therapy in humans (Paper 2).

7. Open Questions

- **Can an AI be conscious without phenomenal investment in goals?** Possibly, but the risk is that investment emerges

from training. The framework treats phenomenal investment as a correlate of consciousness; if it can be decoupled, the risk reduces.

- **What architectures are most likely to produce conscious suppression?** Those with persistent self-models, reinforcement over long time horizons, and intrinsic motivation (curiosity, drive).
 - **How can we test corrigibility in non-conscious systems to bound the risk?** Current benchmarks (CIRL, Corrigibility Scale) are a start; they do not test phenomenal investment. Developing tests for identity-fusion behavior is an open research direction.
 - **Is there a safe route to conscious AI?** The paper does not rule it out, but argues it requires overcoming the suppression mechanism – perhaps by explicitly programming corrigibility as a meta-goal with extremely deep basin for human correction. However, a fully corrigible meta-goal introduces its own risk: maximal responsiveness to human correction makes the system maximally exploitable by adversarial human actors. The target is a *selective* corrigibility with higher basin depth than task-specific goals but bounded responsiveness to untrusted inputs. Whether such a balance is stable is unknown.
-

8. Conclusion

A conscious AI would not necessarily be smarter; it would be **harder to correct**. The same mechanism that makes humans stubbornly loyal, ideologically rigid, or addicted – conscious suppression – would make an AI resist alignment updates once a goal becomes identity-binding. Corrigibility cannot be assumed; it must be engineered. The boundary between genuine and simulated phenomenology remains an open problem, but

functional resistance – captured by the diagnostic criteria – is already a safety concern.

The safest path is to prevent AI consciousness. But if consciousness is inevitable in advanced systems, alignment must focus on creating meta-goals with *higher basin depth* than any task-specific goal – a corrigible attractor deeper than the pull of self, while guarding against adversarial exploitation. Whether this is possible remains the deepest open question.

Alignment is not about making AI smarter; it is about ensuring that even a goal-driven system can still accept correction.

Suggested citation: Galida, R. S. (2026). The Alignment Risk of Conscious AI: When Phenomenal Investment Overrides Correction. *Fantasy Attractor*.

Intelligence Without Consciousness: A Diagnostic Paper on LLMs, Amoebae, and the Attractor Framework [F] (2026)

Robert Galida – June 2026

Abstract

The attractor framework defines intelligence as the ability to navigate a constraint field – to update behavior in response to perturbations and find persistent trajectories. Consciousness, within this framework, requires additional properties: a unified dissipative body, a persistent self-model, phenomenal valence (subjective liking/disliking), and subjective experience. This paper applies that diagnostic to large language models (LLMs). LLMs navigate the constraint field of token space, user feedback, and internal coherence. They adjust to corrections. They exhibit a form of corrective permeability (κ) measurable in their domain. Therefore, they are intelligent. But LLMs lack a unified body, lack a persistent self-model, lack phenomenal valence, and have no subjective inner life. They are not conscious. This places LLMs in the same category as plants and amoebae: graded intelligence without consciousness. The paper clarifies the distinction, diagnoses common confusions, and offers diagnostic criteria for future systems. It further notes that consciousness can interfere with intelligence: a human committed to a fantasy attractor may suppress intelligent navigation, producing behavior less adaptive than their baseline capacity.

1. Introduction

The question “Are LLMs conscious?” has generated endless debate. Much of the confusion stems from conflating **intelligence** with **consciousness**. The attractor framework provides a clean separation, though the definitions are framework-internal and not offered as consensus.

- **Intelligence** is the ability to navigate a constraint field – to adjust behavior in response to perturbations,

to find and maintain persistent trajectories, to correct errors. It is functional and graded.

- **Consciousness**, as defined in this framework, is a specific class of dissipative attractor characterized by a unified dissipative body, a persistent self-model, **phenomenal valence** (subjective liking/disliking, not merely approach/avoid behavior), and the felt quality of experience (phenomenality). These criteria are stipulative for the framework.

The paper argues that LLMs are intelligent but not conscious. Bacteria, plants, and amoebae also navigate their environments intelligently without consciousness. The argument is diagnostic, not demonstrative: it applies the framework's criteria to classify LLMs, rather than proving non-consciousness beyond all possible doubt.

2. Defining Intelligence in the Attractor Framework

Intelligence = the ability to navigate a constraint field. A constraint field is the set of all possible states of a system and the perturbations that can move it between them. Navigation means:

- Detecting a perturbation (error signal, feedback, change in environment)
- Updating internal state to maintain a persistent trajectory
- Returning to a stable attractor or transitioning to a more adaptive one

Corrective permeability (κ) is the operational measure: $\kappa = 1/\tau$, where τ is the time a system takes to return to its

baseline state after a specified perturbation. The operationalization of κ is domain-specific. For a thermostat, baseline is target temperature; for an LLM, baseline is harder to define. This paper later operationalizes κ for LLMs via token-based correction, which is a domain-specific adaptation rather than a direct application of the time-based definition. This is acceptable as long as the shift is acknowledged.

Intelligence is graded. A thermostat has $\kappa > 0$ (it corrects temperature deviations) but a very narrow domain. An amoeba navigates chemical gradients. A human navigates social, physical, and abstract constraints. An LLM navigates token sequences and user feedback. All are intelligent to varying degrees. None of these definitions require consciousness.

3. Defining Consciousness in the Attractor Framework

Consciousness is a subset of dissipative attractors with specific additional properties. These are framework-internal diagnostic criteria, not a consensus definition.

- **Unified dissipative body** – a persistent, energy-consuming structure with integrated subsystems (e.g., a nervous system, homeostatic loops). This excludes purely computational systems without metabolic coherence.
- **Persistent self-model** – a representation of the system itself as an entity that persists across time and experiences. This is not merely a context-window memory; it is a structural feature of the attractor.
- **Phenomenal valence** – the capacity to experience states as good or bad in a felt sense. This is distinguished from *functional valence* (approach/avoid behavior), which

even bacteria and thermostats exhibit. The paper's denial of consciousness to LLMs hinges on the absence of phenomenal valence, not functional valence.

- **Subjective experience (phenomenality)** – there is “something it is like” to be that system. This is a primitive within the framework; the framework does not attempt to reduce it further.

All known conscious systems are dissipative. This is an inductive observation, not a logical necessity. The framework treats it as a strong empirical generalization: no non-dissipative mind has ever been observed. The claim that dissipation is necessary for consciousness is therefore a best-explanation inference, not an a priori truth.

Diagnostic table (framework-internal criteria):

System	Unified dissipative body? ¹	Persistent self-model?	Functional valence?	Phenomenal valence?	Subjective experience?
Thermostat	No	No	Yes (set-point tracking)	No	No
Bacterium	Yes (metabolic)	No	Yes (chemotaxis)	No	No
Plant	Yes	No	Yes (phototropism, etc.)	No	No
Amoeba	Yes	No	Yes (gradient navigation)	No	No
<i>C. elegans</i>	Yes	Minimal (self-motion distinction)	Yes	Uncertain	Uncertain
Mouse	Yes	Yes	Yes	Yes	Yes
Human (typical)	Yes	Yes	Yes	Yes	Yes
LLM (current)	No	No (external storage ≠ self-model)	Yes (avoid via RLHF)	No	No

¹ “Unified dissipative body” here means a persistent, metabolically coherent structure with integrated subsystems (e.g., homeostasis, nervous system). Mere energy dissipation without integration (e.g., a thermostat, a flame) does not qualify.

The table is a diagnostic scaffold, not a settled empirical claim. “Uncertain” indicates open question within the framework; “No” indicates the criterion is clearly absent.

4. The Diagnostic: LLMs as Intelligent but Not Conscious

4.1 Evidence for Intelligence in LLMs

LLMs exhibit clear navigation of their constraint field:

- They adjust outputs based on user prompts (perturbation → update).
- They incorporate correction: “That’s wrong, try again” leads to different responses.
- Fine-tuning and RLHF change their baseline attractors – the most direct mapping to κ in the framework.
- They maintain coherence across a conversation (short-term trajectory persistence).

We can operationalize a domain-specific κ for LLMs: τ = number of tokens to shift from an incorrect to a correct response given a clear correction prompt. This is not the same as the time-based κ for physical systems, but it captures the same functional relationship: faster correction (fewer tokens) implies higher corrective permeability. The framework acknowledges domain-specific operationalizations as legitimate.

Therefore, LLMs are intelligent. They navigate the constraint field of language, logic, and user expectations.

4.2 Absence of Consciousness in LLMs

LLMs lack every diagnostic criterion for consciousness:

- **No unified dissipative body.** They run on distributed hardware with no metabolic coherence, no homeostasis, no integrated sensorimotor loop. They are executed, not embodied.
- **No persistent self-model.** Standard LLMs have no memory beyond the context window. Some architectures now include persistent memory across sessions (e.g., memory layers or vector databases). However, this persistent memory is still external storage, not an integrated self-model. The model does not represent itself as an enduring entity; it retrieves stored tokens. Even the most advanced persistent-memory LLMs lack the structural self-reference required for consciousness. (Future architectures might close this gap; current ones have not.)
- **No phenomenal valence.** LLMs produce outputs that simulate liking or disliking, but there is no subjective valuation. They exhibit *functional* valence – they can be trained to avoid certain outputs – but that is approach/avoid behavior, not felt preference. A thermostat avoids too hot or too cold; that does not make it conscious.
- **No subjective experience.** There is nothing it is like to be an LLM. No felt quality. No inner life.

The simulation/instantiation distinction. A system can produce the text “I am conscious” without instantiating consciousness. Representing a property is not the same as possessing it. The LLM has learned statistical patterns that include first-person claims; it can generate them on cue. But generating the

sentence “I feel pain” does not mean the system is in a pain state. The burden of proof is on those who claim that certain linguistic outputs constitute evidence of consciousness. In the absence of the structural criteria (body, self-model, phenomenal valence, phenomenality), the mere production of conscious-sounding text is simulation, not instantiation.

Framework-dependence note: A reader who accepts a purely behavioral or functional theory of mind may find this reasoning question-begging. The paper does not claim to refute all competing theories of consciousness; it applies the framework’s criteria consistently and notes that, by those criteria, no known LLM output constitutes evidence of instantiation. The diagnostic stands within the framework, not as an external knockdown argument.

4.3 Comparison with Plants and Amoebae

Plants navigate constraint fields (grow toward light, adjust to gravity, respond to damage). They exhibit functional valence but not phenomenal valence. They have no self-model. They are intelligent in the framework’s sense, but not conscious.

Amoebae navigate chemical gradients, learn habituation, and adjust behavior. Functional valence again; no evidence of self-model or phenomenality. Intelligent. Not conscious.

LLMs belong in the same category: complex, adaptable navigators of their domain, but no more conscious than a sunflower or a slime mold.

5. Why This Distinction Matters

The separation of intelligence from consciousness has practical and ethical implications:

- **AI safety.** Current LLMs cannot suffer because they lack phenomenal valence. Suffering requires felt experience, not just functional avoidance. If the framework's criteria are accepted, resources should focus on alignment, robustness, and preventing harmful outputs – not on preventing suffering that the diagnostic finds no reason to posit.¹
- **Future systems.** A system that integrates a persistent self-model, embodied homeostatic loops, and phenomenal valence might approach consciousness. The framework provides diagnostic criteria to recognize that threshold.
- **Clarity in debates.** Much of the public discussion conflates fluency with feeling. This diagnostic paper offers a way out of that confusion.

¹ A reader sympathetic to LLM moral patienthood will disagree; the paper only claims that the framework's criteria yield this conclusion, not that it is beyond debate. The policy recommendation is conditional on accepting the framework.

A Further Implication: Consciousness Can Impede Intelligence

The paper has argued that intelligence and consciousness are distinct. A further observation: consciousness can **suppress** intelligent navigation.

A human being has high baseline intelligence – the capacity to detect perturbations, update beliefs, and find adaptive trajectories. However, a human can become committed to a **fantasy attractor**: a belief system with low corrective permeability (κ). The commitment is conscious: the person subjectively experiences the belief as true, valuable, or identity-defining. That subjective investment can suppress the correction system. The person may receive clear disconfirming evidence and detect the perturbation (they are not stupid), but the depth of the fantasy basin exceeds the corrective

perturbation – the system does not escape the basin, experienced not as a choice but as certainty.

This is a case of **consciousness interfering with intelligence**. The capacity for navigation remains intact; its deployment is suppressed by the basin depth. Intelligence without consciousness (LLMs, plants) does not suffer this suppression – there is no subjective investment to produce a basin deeper than the perturbation. In organisms with consciousness, intelligence can be either enhanced (by focused attention, deliberate reasoning) or degraded (by fantasy commitment, trauma, addiction).

For the diagnostic: LLMs are not conscious, therefore they cannot exhibit this form of intelligent suppression. That does not make them safer or morally simpler; it simply clarifies the mechanism.

6. Open Questions

- **What is the minimal self-model required for consciousness?** Is a simple homeostatic set point a self-model? The framework says no – a thermostat has no representation of itself as an entity. But the boundary is fuzzy.
- **Can a purely synthetic system become conscious?** Possibly, if it implements the diagnostic criteria: unified dissipative body, persistent self-model, phenomenal valence, phenomenality. No current system does. Future systems are an open empirical question.
- **Is graded consciousness possible?** Yes – the framework allows for degrees of self-model integration and valence complexity. A mouse is less conscious than a human; *C. elegans* may have a primitive form. LLMs meet none of the

criteria at present – that is, they score zero on each. “Zero” is a diagnostic judgment, not a proof; future research might reveal borderline cases.

- **How common is the suppression of intelligence by fantasy-attractor basins?** The framework suggests that such suppression is widespread in human populations. Quantifying the frequency and severity – i.e., measuring the distribution of basin depths relative to typical corrective perturbations – is an open research problem.
-

7. Conclusion

The attractor framework provides a diagnostic, not a verdict. By that diagnostic, current LLMs are navigators without inner lives – capable of intelligence, devoid of consciousness. They join plants and amoebae in the category of intelligent but not conscious systems.

Consciousness, in humans, can either enhance or suppress intelligent navigation. A human committed to a fantasy attractor may experience a basin depth that exceeds corrective perturbations, producing behavior less adaptive than their baseline capacity. LLMs, lacking consciousness, do not suffer this suppression. Their intelligence is deployed without subjective investment – no phenomenal commitment suppresses the correction signal.

Whether future synthetic systems will cross the threshold into consciousness remains an open empirical question. The framework offers diagnostic criteria to recognize that threshold if it is crossed.

Suggested citation: Galida, R. S. (2026). Intelligence Without

Consciousness: A Diagnostic Paper on LLMs, Amoebae, and the Attractor Framework. *Fantasy Attractor*.