Non-Physical Claims Are Fantasy Attractors: Why Unverifiable Realms Cannot Be Empirically Distinguished from Nonexistence

Robert Galida – June 2026
[F] (Foundation

Abstract

The attractor framework adopts a physicalist commitment: to be real is to be able to interact, and to interact is to share at least one **interaction channel** (spacetime, energy, momentum, gauge charge, or any measurable coupling). This is a philosophical starting point, not an empirical discovery. The paper argues that any claim about a non-physical realm – defined as having no such interaction channel – cannot be empirically assessed. Such claims are **fantasy attractors**: belief systems structurally sealed against correction by defining their objects as forever beyond any possible test. The paper distinguishes provisional non-detection (e.g., dark matter) from **structural, permanent non-verifiability** (e.g., non-physical gods, transcendent souls). It concludes that while such claims may have personal or social meaning, they cannot be part of a scientific ontology, and their structure makes them vulnerable to fraud and manipulation – though sincere belief is not fraud.

1. The Foundational Commitment: Interaction Requires Shared Channels

The attractor framework is a physicalist ontology. It begins with a commitment: **entities can only interact through shared interaction channels**. An *interaction channel* is any measurable coupling — spacetime coordinates, energy, momentum, electric charge, weak isospin, color charge, or any other quantity that can be transferred or correlated between systems. This is not an empirical discovery of the Standard Model; it is the framework's chosen criterion for what counts as real.

The neutrino example illustrates the criterion but does not prove it. Neutrinos interact weakly because they share weak isospin; they do not interact electromagnetically because they lack electric charge. The framework simply says: if an entity shares no interaction channel with physical reality, we have no way to detect it, measure it, or include it in a scientific ontology. That is a philosophical choice, not a falsifiable claim about the world.

Why interaction? Interaction is chosen because it provides a public, corrigible basis for knowledge. It avoids ontological commitments that cannot influence observation, and it aligns with the core principle of the attractor framework: *persistence under perturbation*. An entity that never perturbs anything cannot be distinguished from nothing.

What the framework does not claim:

- That non-physical entities are logically impossible.
- That all non-physical claims are false.
- That physics has disproven God or the supernatural.

What it does claim:

- That non-physical entities cannot be empirically distinguished from nonexistence.
- That claims about them operate as fantasy attractors, resistant to correction.

---

2. Types of Non-Physical Claims

A non-physical claim is any assertion about an entity, force, or realm defined as having **no interaction channel** with the physical world. However, not all claims that seem non-physical are alike. We distinguish two categories:

Category A: Truly non-interacting — Claims that explicitly deny any possible interaction. Examples:

- A deistic creator who wound the universe and then never interacts.
- A transcendent God defined as beyond all categories, including causality.
- An immaterial soul that cannot influence the body after death.
- Abstract objects (Platonism) that exist non-physically and non-causally.

Category B: Claims that assert interaction but evade testing — Examples:

- Ghosts that move objects but become undetectable when instruments are present.
- Psychics whose powers fail under controlled conditions (explained as “skeptic’s energy”).
- Homeopathic “water memory” that cannot be detected by any known physical measurement.

Category B is a different epistemic pathology: motivated reasoning, ad-hoc escape clauses, and sealing mechanisms. The attractor framework addresses them as *functionally* non-verifiable in practice, but they are not the primary target of this paper. This paper focuses on **Category A**: claims that structurally preclude any possible interaction channel.

Domain (Category A)	Example Claim	Interaction Channel?	Empirically Assessable?
Religion (non-interacting God)	A creator with no detectable properties	None	No – any test is ruled out a priori
Paranormal (non-interacting ghosts)	Ghosts that cannot affect matter	None	No – no possible evidence
Abstract objects (Platonism)	Numbers exist non-physically, non-causally	None	No – no interaction, hence no evidence
New Age (non-interacting “vibrations”)	Crystals with undetectable healing vibrations	None	No – absence of effect is blamed on “wrong intent”

Under the framework’s commitment, such claims are not false; they are **not empirically assessable**. They belong to a different domain: personal belief, fiction, or social identity.

---

3. Provisional vs. Structural

Non-Verifiability

A crucial distinction separates:

- **Provisional non-detection** – e.g., dark matter, gravitational waves (before 2015), the neutrino (before 1956). These entities are predicted to share at least one interaction channel (gravity, weak force) and are in principle detectable. **A future discovery could confirm or disconfirm them.** That is the key: we can specify what would count as evidence, even if we don't yet have it.
- **Structural, permanent non-verifiability** – Category A claims. The entity is defined so that **no possible future discovery** could ever count as confirmation or disconfirmation. Any proposed test is ruled out in advance. This is the hallmark of a fantasy attractor.

(This framework does not assert that dark matter could have been called a fantasy attractor before detection; dark matter always had specified interaction channels – gravity – and was therefore never structurally non-verifiable.)

---

4. Fantasy Attractor: Formal Definition

A belief system qualifies as a **fantasy attractor** if it meets the following conditions:

1. **No specified interaction channel** – The central claim lacks any measurable coupling to physical reality (Category A), or defines it in a way that systematically evades testing (Category B).
2. **Sealing mechanisms** – The belief incorporates rhetorical or cognitive strategies that neutralize disconfirming evidence (e.g., "God works in mysterious ways," "The

ghost left when the EMF meter arrived”).
3. **Low corrective permeability (κ → 0)** — The belief does not update in response to counterevidence; the return time τ to baseline is effectively infinite.
4. **Identity fusion** — The belief is tied to self-worth or group membership, making abandonment costly.

Under this definition, both Category A and some Category B claims can be fantasy attractors, but Category A are the paradigmatic case because they are structurally immune to evidence.

---

5. Fiction Is Real but Not True: A Crucial Distinction

The main argument might provoke an objection: *What about fiction? Sherlock Holmes is not physical, yet we say he exists as a character. Isn’t that a counterexample to the claim that non-physical entities cannot be empirically distinguished from nonexistence?*

The objection fails because it conflates two different senses of “exists.” We must distinguish:

- **Fiction exists as physical information.** The character Sherlock Holmes is realized as patterns of ink on a page, as sounds in a performance, as neural firing patterns in readers’ brains, or as bits on a computer screen. Information is a physical arrangement of matter. It shares interaction channels (energy, spacetime, causality) with the physical world. You can buy a book, discuss the plot, or be emotionally affected by a story. Fiction is **real** in this sense: it has a physical substrate and causal effects.

- **Fiction is not true.** The proposition “Sherlock Holmes lived at 221B Baker Street” does not correspond to any actual state of affairs in the world. It is false. Fiction is not required to be verifiable; it is understood as imagined.

Thus, the attractor framework happily accommodates fiction. It is real as information, but not claimed as true.

The bad faith of non-physical claims: Non-physical claims that demand to be treated as real — gods, ghosts, souls, hidden cabals — are *fiction pretending to be true*. They borrow the ontological status of real information (they exist as patterns in books, sermons, or brains) but also demand the epistemic authority of factual truth. Yet they refuse any possible test. They define themselves as beyond verification. This is bad faith: it is not metaphysics, but fiction that insists on being taken as fact while rejecting the rules of fact-checking.

Category	Exists as physical information?	Claims to be true?	Verifiable?	Framework classification
Fiction (Hamlet)	Yes	No (acknowledged as imagined)	Not applicable	Real information, not true
Scientific claim (neutrino)	Yes (theory, data)	Yes	In principle	Real, true (provisionally)
Non-physical claim (God)	Yes (as cultural artifact)	Yes	No — structurally excluded	Fantasy attractor

Therefore, the framework does not deny the reality of stories; it denies the epistemic legitimacy of treating unverifiable stories as facts. The fantasy attractor is not the story. It is the insistence that the story is true combined with the structural refusal to let the story be tested.

6. Vulnerability to Fraud and Manipulation

The structure of non-physical claims makes them **vulnerable** to fraud and manipulation — not that all such claims are fraudulent. Because there are no checks, a bad actor can assert divine commands, psychic readings, or secret knowledge without fear of disconfirmation. Sincere believers are not fraudsters, but the attractor basin can be exploited by those who understand its dynamics.

The framework diagnoses the **structure**, not the intent of every believer. It distinguishes **error, self-deception, motivated reasoning, and fraud** — all possible outcomes, but not all present in every case.

---

7. What This Argument Does Not Prove

To avoid overreach, the paper explicitly states what it does **not** claim:

- It does not prove that non-physical entities are logically impossible.
- It does not refute philosophical positions like Platonism (abstract objects) or classical theism that defines God as existence itself rather than an interacting object — though it notes that such positions are not empirically assessable.
- It does not claim that all believers are fraudsters or that all non-physical claims are meaningless in a philosophical sense.
- It does not assert a timeless criterion for what will be discovered in the future.

The claim is narrower: **within the attractor framework's physicalist commitment, non-physical claims are not empirically assessable, and they exhibit the dynamics of fantasy attractors.**

---

8. Conclusion

The attractor framework adopts a physicalist commitment: entities can only interact through shared interaction channels. Non-physical claims — defined as having no such channels — are not empirically assessable. They are fantasy attractors: belief systems structurally sealed against correction by permanent non-verifiability. This does not make them meaningless or false; it places them outside the domain of scientific ontology. Their structure makes them vulnerable to exploitation, but sincere belief is not fraud. The framework provides a diagnostic tool for recognising when a claim has been immunised against evidence, regardless of its content.

The argument supports the following conclusion:

> ***Claims that are permanently insulated from any possible empirical correction occupy a distinct epistemic category and exhibit attractor dynamics that make them resistant to updating. Within the attractor framework's physicalist ontology, such claims cannot be empirically distinguished from nonexistence.***

That is a substantial claim. It does not require asserting that non-physical realms cannot exist — only that they cannot be part of a scientific ontology, and that the beliefs which cling to them operate as fantasy attractors.

Suggested citation: Galida, R. S. (2026). Non-Physical Claims Are Fantasy Attractors: Why Unverifiable Realms Cannot Be Empirically Distinguished from Nonexistence. *Fantasy Attractor.*

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 1https://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 κ(G) for corrections that threaten G. We can write, notationally:

$\kappa\_corrected(G) = \kappa_0(G) - \Delta\kappa$

where Δκ 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) × (temporal persistence of goal representation). **Crucially, this same functional Δκ 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 Δκ is large enough relative to κ_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 Δκ.

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.*