Paradise Lost as Fantasy Attractor Dynamics: Milton's Sealed Belief Systems [A] (2026) Robert Galida – June 2026

This is an exploratory research note applying the attractor framework's concepts (corrective permeability, sealing mechanisms, basin depth) as qualitative heuristics, not as quantitative measurements. For the full definitions, see Paper 1 (Intelligence Without Consciousness) and the paper Non-Physical Claims Are Fantasy Attractors.

Abstract

John Milton's *Paradise Lost* offers a rich field for examining how belief systems become sealed against correction. Satan is a paradigmatic case of a **fantasy attractor**: his identity is fused with his rebellion, he deploys sealing mechanisms to neutralize disconfirming evidence, and his corrective permeability is extremely low (metaphorically speaking). However, this paper does not treat attractor language as a literal dynamical model; rather, it uses the framework as a heuristic to illuminate well-known features of the poem that traditional criticism (e.g., C.S. Lewis, Stanley Fish) has already noted. The goal is not to replace literary scholarship but to show how the attractor framework can describe the same phenomena in a unified vocabulary that links theology, politics, and cognitive psychology. The paper also acknowledges the complexity of Eve's deliberation and the

Son's grace as a genuine perturbation that restores corrigibility. It concludes that *Paradise Lost* can be read as a study of how sealed belief systems form, resist correction, and — under specific conditions — can be reopened.

---

1. Introduction

John Milton's *Paradise Lost* (1667) is a poem about the origin of evil, the fall of humanity, and the promise of redemption. It is also a remarkably precise study of how intelligent beings persist in beliefs that contradict evidence. Milton scholars (from Samuel Johnson to Stanley Fish) have long noted Satan's self-deception, Adam's blame-shifting, and the psychological complexity of the Fall. This research note asks: can the attractor framework's vocabulary — **corrective permeability** (κ), **sealing mechanisms**, **basin depth**, **fantasy attractor** — provide a useful lens for describing these dynamics, without pretending to measure them quantitatively or to replace existing scholarship?

The answer is: yes, as a **heuristic**. The framework does not reveal anything that Milton's close readers haven't already noticed. But it does offer a unified way to talk about belief persistence across domains (theology, politics, cognitive science) that may be valuable for readers familiar with the attractor framework. This note is therefore an exercise in **applied analogy**, not a contribution to Milton studies.

---

2. The Attractor Framework as Heuristic (Not a Formal Model)

In the attractor framework, a **fantasy attractor** is a belief system with very low corrective permeability ($\kappa \to 0$), a deep

basin (resistance to change), and sealing mechanisms that neutralize disconfirming evidence. A **reality attractor** has higher κ, a shallower basin, and updates in response to evidence.

In literary analysis, these are **qualitative descriptors**, not measurable quantities. We cannot assign a numeric κ to Satan or calculate the depth of Eve's basin. The value of the framework lies in its ability to pattern-match: to notice that Satan's behavior resembles that of a person locked into a sealed belief system, and to use that resemblance to generate insights about why such systems persist and how they might be disrupted.

This is not circular. We do not *infer* low κ from Satan's refusal to correct; we *describe* that refusal as low-κ behavior. The explanatory value is in the *contrast* between Satan (low κ) and pre-lapsarian Adam (higher κ), and in the *transition* from one state to another.

---

3. Satan: A Sealed Belief System (But Not a Simple One)

Traditional criticism (e.g., C.S. Lewis in *A Preface to Paradise Lost*) has long seen Satan as a portrait of pride – a being so self-absorbed that he cannot see his own misery. More recent critics (e.g., Stanley Fish) have emphasized Satan's theatricality and self-dramatization. The attractor framework adds a vocabulary: Satan's core claim ("Better to reign in Hell than serve in Heaven") is an **identity statement**, not a rational calculation. He has **fused** his rebellion with his sense of self. To abandon the rebellion would be to annihilate himself.

Sealing mechanism: "The mind is its own place, and in itself /

Can make a Heav’n of Hell, a Hell of Heav’n” (I.254-255). This is a classic sealing move: reality is redefined as irrelevant. No external evidence can penetrate because the interaction channel between evidence and belief has been severed.

Self-awareness: Satan is not merely deluded. He repeatedly admits his misery: “Which way I fly is Hell; myself am Hell” (IV.75). Yet he still does not update. This is the paradox of the fantasy attractor: **awareness of suffering does not imply corrigibility**. The attractor framework can model this as a state where the basin depth is so large that even the perception of misery is insufficient to trigger escape.

Thus, the framework does not reduce Satan to a simple automaton. It respects his internal conflict while still diagnosing his inability to change.

---

4. Pre-lapsarian Eden: A More Corrigible State

Before the Fall, Adam and Eve operate in what the framework calls a **reality attractor**: they receive correction (from God and angels), discuss it, and update their behavior. When Eve has a troubling dream, she tells Adam, and they dismiss it (V.95-113). Their κ is relatively high; their basin is shallow.

This is not a claim that they are perfectly rational. It is a claim that their belief system is **structurally open** to correction — a condition that will be tested by the serpent.

---

5. The Fall: A Gradual Attractor Transition

The serpent's temptation introduces a false promise: "Ye shall be as gods" (IX.708). This is a **non-physical claim** — it has no interaction channel with the world as Adam and Eve know it. It cannot be verified or falsified. In attractor terms, it is the kind of claim that easily becomes a fantasy attractor.

Eve's deliberation in Book IX is subtle. She does not simply flip. She reasons, hesitates, and persuades herself. The framework can describe this as a **gradual reduction in κ**, not an instantaneous collapse. The sealing mechanism ("What could be more fair than to know good and evil?" — IX.727-728) is deployed before the fruit is eaten. By the time she eats, her basin has already deepened.

Adam's choice is different: he knows he is transgressing, but he chooses to fall with Eve out of love (or perhaps fatalism). His κ collapses almost instantly. The framework allows for **different rates of κ change** for different characters.

---

6. Post-lapsarian Behavior: Deflection and Hiding

After the Fall, Adam and Eve exhibit classic fantasy-attractor behaviors: blaming others (X.128-137), hiding from God (IX.1112-1113), and struggling to answer when questioned. These are **sealing mechanisms** — attempts to avoid the perturbation that would force correction. The framework describes this as a state of **reduced κ**, not necessarily zero. Redemption is still possible.

---

7. The Son as a Genuine Perturbation

God's interrogation is the first attempt to reopen the basin. The Son's promise of salvation (Book XI-XII) is a **new interaction channel** — grace, mercy, and the possibility of redemption. This is not a mechanical "increase in κ." It is a theological event. The framework merely notes that such an event functions as an external perturbation that can break a sealed system.

Milton's own theology emphasizes free will and repentance. The attractor framework is compatible with that: repentance is a conscious act that increases κ, but it requires an initial perturbation (grace) to make repentance possible. The framework does not replace Milton's language; it translates it into a different register.

---

8. Political Allegory: A Modest Reading

Milton was a republican who defended the regicide of Charles I. Many scholars (e.g., Christopher Hill) have read *Paradise Lost* as a political allegory. In attractor terms, one could argue that:

- **Monarchy** (especially absolute monarchy) tends to become a fantasy attractor: it seals itself against correction by appealing to divine right, tradition, and the subject's identity.
- **Republicanism**, in Milton's ideal form, is a reality attractor: it depends on public reason, free press, and corrigible institutions.

But this is **one possible reading**, not a definitive mapping. The paper does not assert that Milton himself thought in these terms. It simply notes that the attractor framework can

describe the political dynamics that Milton was engaging with.

A critic could object that republics can also become sealed (e.g., the Jacobin terror). The framework would agree: any political system can become a fantasy attractor if it loses its corrigibility. The distinction is structural, not ideological.

---

9. What Would Disconfirm the Framework?

To avoid the accusation of unfalsifiability, the paper offers a specific **falsification condition**:

> *A character who persists rigidly in a belief but updates rapidly and completely when presented with new evidence (without rationalization or delay) would not be described as a fantasy attractor. Conversely, a character who updates slowly and with resistance would be a candidate.*

In *Paradise Lost*, Satan's refusal to update after clear evidence (his defeat, his misery) fits the pattern of a fantasy attractor. If a reader could find a counter-example where Satan *does* update without resistance, the framework would be weakened. (No such example exists in the poem.)

This is a modest falsifiability condition, but it is genuine.

---

10. Conclusion

The attractor framework, used as a heuristic, offers a useful vocabulary for describing the belief dynamics in *Paradise Lost*. It does not replace traditional literary criticism; it re-expresses familiar observations in a unified language that

connects theology, politics, and cognitive psychology. The paper does not claim to measure κ or basin depth; it uses these terms qualitatively, as one might use “depression” or “obsession” in psychological criticism.

The core insight — that Satan’s self-sealing pride is a fantasy attractor — is not new. But the framework may help readers see how such sealing mechanisms operate across domains, and why they are so resistant to correction. Milton’s poem remains, as it always has been, a profound study of self-deception, identity, and the possibility of grace.

Suggested citation: Galida, R. S. (2026). Paradise Lost as Fantasy Attractor Dynamics: Milton’s Sealed Belief Systems (Research Note). *Fantasy Attractor.*

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

Domain (Category A)	Example Claim	Interaction Channel?	Empirically Assessable?
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 ($\kappa \to 0$)** — 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.*

Why Clockwork Interventions Fail in Complex Systems: A Prescription from the Attractor Framework [A] (2026)

Robert Galida — June 2026 (Final)

See Paper 1 (Intelligence Without Consciousness) for the full taxonomy of attractors, κ, and basin depth. See Basin Defense and Stable Addition for cross-domain synthesis and rate-induced tipping.

Abstract

Most human institutions, policies, and interventions treat complex adaptive systems as if they were clockwork systems — linear, predictable, and responsive to force. This is a category error. Complex systems (ecosystems, brains, societies, belief systems) have attractors, basins, multiple nested timescales (κ vector), and thresholds. Applying sudden force above a critical rate or magnitude triggers basin defense: ejection, backlash, entrenchment, or catastrophic collapse. This paper diagnoses the clockwork fallacy, introduces a multi-timescale operationalization of corrective permeability, offers a mechanism for parallel attractor replacement, and acknowledges the institutional constraints that make patient intervention rare. The central argument is that failure is not random but structurally predictable.

---

1. Introduction

A thermostat is a clockwork system. Push the temperature up, the cooling turns on; push harder, it turns on faster. No hidden attractors, no basin defense, no hysteresis. Force works predictably.

A human being is not a thermostat. Neither is a democracy, an ecosystem, a marriage, or a belief system. They have attractor basins — stable states that resist displacement. They have multiple corrective timescales (κ vector) — characteristic return times after perturbations at different levels. They have thresholds — points at which a small additional push can cause a regime shift.

Yet most interventions treat these complex systems **as if they were clockwork**. Apply more force → get more change. This is the **clockwork fallacy**.

This paper diagnoses the fallacy using the attractor framework, operationalizes κ for non-physical domains as a vector of timescales, specifies the mechanism of parallel attractor replacement, and acknowledges the institutional constraints that make slow intervention rare.

---

2. The Clockwork Fallacy in Framework Terms

Clockwork assumption	Complex system reality
Linear response: more force → more change	Nonlinear: small force may be ejected; force above threshold may cause collapse

Clockwork assumption	Complex system reality
No memory: each intervention acts independently	Hysteresis: history matters; past perturbations shape current basin depth
No internal dynamics: system is passive	System has its own attractors and κ vector; it actively resists displacement
Fast intervention is better (efficiency)	Rate matters; fast perturbation triggers basin defense; slow perturbation may integrate

The clockwork fallacy treats the system as a **passive object** to be pushed. The attractor framework treats it as an **active agent** with its own stability dynamics.

---

3. Operationalizing κ as a Multi-Timescale Vector

κ = 1/τ, where τ is the characteristic return time to baseline after a small perturbation. For physical systems (thermostat, RC circuit), τ is a single scalar. For complex adaptive systems, **τ is not a single number** — there are multiple, nested timescales:

Timescale	Definition	Example (addiction)
Fast κ (seconds–hours)	Return time after transient perturbation	Craving decay
Medium κ (days–weeks)	Return time after moderate perturbation	Withdrawal normalization
Slow κ (months–years)	Return time after identity-level perturbation	Identity fusion / self-model reorganization

Timescale	Definition	Example (addiction)
κ_∞ (effectively zero)	No measurable return; the attractor is sealed	Fantasy attractor (see Paper 1)

Implication: A system can have fast κ (rejects rapid, small perturbations) and slow κ (integrates slow drift) simultaneously. The optimal perturbation rate depends on *which* κ you are trying to match.

Protocol for estimating κ in a non-physical domain:

1. Select a modest, low-stakes belief (not identity-core).
2. Introduce a small, credible counter-evidence (pilot perturbation).
3. Measure the time until the person returns to their original stated belief (via repeated interviews, surveys, or behavior tracking).
4. τ is the median return time; $\kappa = 1/\tau$.
5. Repeat with perturbations that target different subsystem levels (e.g., factual vs. identity-relevant) to estimate the κ vector.

Limitation: The pilot perturbation protocol uses a *small* perturbation to estimate κ. The intervention may require a *large* perturbation to escape the basin. The small-perturbation estimate may not predict behavior near the basin boundary. This is an acknowledged operational limitation, not a circularity. The framework is falsified if a system with measured low κ (slow return) reliably integrates *rapid, large* perturbations without ejection or transient absorption, and if the small-perturbation estimate is stable across perturbation magnitudes.

---

4. Why Clockwork Interventions Fail: Four Mechanisms

Mechanism 1: Ejection (Backlash) — When a perturbation is applied too fast or with too much force, the system ejects the addition, often returning with a deepened basin. Examples: sanctions that strengthen a regime, direct refutation that backfires.

Mechanism 2: Transient Absorption Followed by Return — The system temporarily changes, then returns to baseline when the perturbation stops. Examples: short-term policy boosts, crash diet weight regain.

Mechanism 3: Catastrophic Regime Shift — Force applied at a critical threshold causes an abrupt, often irreversible shift to a different, sometimes worse attractor. Examples: lake eutrophication, restructuring that destroys institutional knowledge.

Mechanism 4: Rate-Induced Tipping — A small cumulative change, applied faster than the relevant κ, causes tipping. Examples: rapid currency appreciation triggering crisis, fast cultural change provoking backlash.

---

5. Parallel Attractors: The Mechanism of Replacement

Parallel attractors are introduced as an alternative to direct displacement. How does a parallel attractor eventually replace the original?

Mechanism: Basin-share competition

When a parallel attractor is created, it initially has a shallow basin. Through repeated use, reinforcement, and social

validation, its basin depth increases. Meanwhile, the original attractor may become shallower through disuse or decoupling of identity fusion. The transition is not a flip; it is a **continuous shift in basin dominance**. At some point, the new attractor's basin depth exceeds the old attractor's, and the system's typical trajectories are captured by the new state.

Testable prediction: During parallel attractor formation, the system will exhibit **bistability** — both states are possible for a range of control parameters. In social systems, this predicts polarization; in organizational change, it predicts pilot-program coexistence; in belief systems, it predicts identity compartmentalization.

Empirical examples: Harm reduction (methadone maintenance creates a parallel attractor that may deepen over time); phase-in policies (smoking bans create new norm attractors alongside old habits); belief change (new social identity cultivated alongside old identity, enabling eventual abandonment without direct confrontation).

---

6. The Political Economy of Slow Intervention

The attractor framework prescribes patience, precision, and gradual perturbation. But policymakers, clinicians, and managers face **institutional incentives** that systematically favor fast, visible, forceful action:

- Election cycles (2–4 years) reward short-term results, not long-term basin reshaping.
- Media attention favors dramatic events, not gradual change.
- Bureaucratic accountability demands measurable outputs, not process fidelity.

- Crisis narratives demand action, not waiting.

Consequence: Even when the framework is correct, it is often institutionally **unimplementable**. The best intervention may be politically impossible.

What would institutional redesign look like? Examples:

- **Longer funding cycles** (5–10 years) for policy and program evaluation, allowing basin-reshaping interventions to mature.
- **Preregistered patience metrics** — requiring intervention designs to specify expected τ and κ, with success measured by reduction in τ over time, not immediate outcomes.
- **Insulation from electoral pressure** for certain regulatory functions (e.g., central bank independence, long-term environmental planning).
- **Dual-track systems** that allow parallel attractors to develop (e.g., pilot programs exempt from standard performance metrics).

Implication for the paper's claims: The framework diagnoses why interventions fail, but it does not guarantee that successful interventions can be implemented. This is not a weakness — it is a feature. The framework clarifies the gap between effective intervention and institutional feasibility. Bridging that gap requires institutional redesign, not just better perturbation design.

---

7. Case Studies

Case 0: Smoking cessation (addiction) — the motivating challenge

In smoking cessation, abrupt cessation (cold turkey) often outperforms gradual tapering (Lindson et al., 2016 meta-analysis). This appears to contradict the prescription "slow perturbation at rate ≤ κ."

Framework interpretation: Addiction has multiple κ timescales. Cold turkey may target the fast-κ (craving) subsystem while the slow-κ identity subsystem remains dormant; gradual tapering may keep both active, prolonging distress.

Falsifiable prediction: Patients with higher identity-fusion scores (measurable via existing scales, e.g., the Identity Fusion Scale) should show worse outcomes with gradual tapering relative to cold turkey. If identity fusion is low, gradual tapering may be equivalent or superior.

Alternative explanations acknowledged: The meta-analysis does not adjudicate between the attractor framework and other accounts (e.g., cognitive dissonance, cue elimination, withdrawal distress). The framework's contribution is to generate the identity-fusion interaction prediction, which can be tested independently.

Case 1: Lake eutrophication (ecological)

- *Clockwork approach:* Sudden nutrient reduction after flipping to turbid state – fails (hysteresis). True hysteresis is technically established for some lakes (Scheffer et al., 2001).
- *Framework approach:* Gradual nutrient reduction before tipping (rate ≤ κ) might have avoided the flip. After tipping, parallel attractor (biomanipulation) is required.

Case 2: Political persuasion (belief systems)

- *Clockwork approach:* Direct refutation, evidence bomb –

backfire effect (ejection with deepened basin).

- *Framework approach:* Yang et al. (2022) demonstrated in a field experiment that "pacing and leading" — starting with some agreement and gradually introducing opposing content — produced attitude change, whereas blunt argument triggered backlash. This is gradual perturbation at rate ≤ κ, combined with identity decoupling.

Case 3: Organizational change

- *Clockwork approach:* Sudden layoffs, top-down mandate — triggers basin defense (resistance, morale loss).
- *Framework approach:* Gradual, participatory change (rate ≤ κ) with parallel structures (pilots, dual systems). *Note:* Hysteresis in organizations is not technically demonstrated; the paper uses "analogous" language.

---

8. Practical Heuristics

If the system has…	Then…	Caveat
Fast κ (seconds–hours)	Rapid, sharp interventions may be required; slow drift may be tracked or rejected	For very deep basins, only a large shock may work
Slow κ (months–years)	Slow, gradual perturbation; avoid rapid shocks	Identity-fused systems may need abrupt escape (Case 0)

If the system has…	Then…	Caveat
Multiple κ timescales	Target the slowest κ for lasting change; use fast κ for immediate disruption	Requires measurement of the κ vector
κ → 0 (fantasy attractor; no measurable return)	Intervention is futile within the model. Accept, circumvent, or refer to Paper 1	Out of scope for this paper
Hysteresis (true bistability)	Do not force return; cultivate a parallel attractor	Hysteresis is established for some ecological systems; for social systems, use “analogous”
Identity fusion	Do not attack belief directly. Decouple identity first, then perturb gently	Requires trust; may be infeasible in adversarial contexts

---

9. Conclusion

The clockwork fallacy — treating complex adaptive systems as linear, passive, and force-responsive — is a primary cause of failed interventions. The attractor framework diagnoses the failure modes (ejection, transient absorption, catastrophic shift, rate-induced tipping) and offers a prescriptive alternative: measure the κ vector, match perturbation rate to the relevant timescale, build parallel attractors, and wait.

The framework does not guarantee success. Institutional incentives (election cycles, media pressure, bureaucratic accountability) systematically favor the clockwork approach, making patient intervention rare. The value of the framework

is diagnostic: it explains why failure is not random, and it clarifies the gap between effective intervention and political feasibility. Bridging that gap requires institutional redesign — longer funding cycles, preregistered patience metrics, and insulation from electoral pressure.

The dance of change is not about pushing harder. It is about learning to move with the system — but also knowing when the system cannot be moved with the tools and time available.

Suggested citation: Galida, R. S. (2026). Why Clockwork Interventions Fail in Complex Systems: A Prescription from the Attractor Framework. *Fantasy Attractor.*

Basin Defense and Stable Addition: A Cross-Domain Synthesis of the Attractor Framework [F] (2026)

Robert Galida — June 2026 (Final)

See Paper 1 (Intelligence Without Consciousness) for the full taxonomy of attractors, κ, and basin depth.

Abstract

Many complex systems resist change by returning to a preferred

low-energy attractor rather than adopting a new state. Whether a perturbation (an added agent, input, or component) is ejected, transiently absorbed, or stably integrated depends on the basin geometry (depth B and barriers) and the system's corrective dynamics (κ = 1/τ). This paper defines B and κ, draws on formal models (stochastic dynamical systems and Kramers escape theory) with explicit qualifications for non-gradient domains, and catalogs exemplar systems across ten domains. A comparative table summarizes systems, mechanisms, proxies for B and κ, timescales, and conditions favoring each outcome. The paper concludes that the same basic physics analog applies across domains: a perturbation of size Δ will be ejected or die out if Δ is below the attractor's effective escape threshold (a function of B), whereas if Δ exceeds that threshold and the system has enough plasticity or additional degrees of freedom, a new stable state can form. A research roadmap is provided in an appendix.

---

1. Introduction

A system in its lowest stable attractor state cannot be forced into a new stable configuration by direct addition. Adding to the system — a third star, an extra electron, a new species, a contradictory belief — will result in one of three outcomes:

1. **Ejection** — the addition is expelled from the system entirely. The original attractor persists.
2. **Transient absorption** — the addition remains present, but the system state returns to the original attractor despite the addition's continued presence.
3. **Stable addition** — the addition is integrated, either by expanding the capacity of the original attractor or by forming a new parallel attractor alongside it.

This paper identifies a unified principle — **basin defense** — that governs these outcomes across physical, biological, ecological, social, and engineered systems. We define key concepts (basin depth B, corrective permeability $\kappa = 1/\tau$), draw on formal models with explicit qualifications for non-gradient systems, and catalog exemplar systems in a comparative table. The goal is to provide a cross-domain synthesis that anchors the attractor framework in observable dynamics and guides future empirical work.

---

2. Definitions and Formal Models (with Qualifications)

Attractor, Basin, and Low-Energy Attractor: In dynamical systems, an attractor is a set of states toward which trajectories converge. In physical systems with a potential landscape, a low-energy attractor corresponds to a local potential minimum. Its basin of attraction is the region of state space that flows into the attractor. **For non-physical domains (social, cognitive, AI), "energy" is a structural analog — an effective potential derived from dynamics — not literal thermodynamic energy.** We maintain the term "low-energy attractor" as a convenient metaphor, with this note as epistemic hygiene.

Basin Depth (B): For systems with a well-defined potential, B is the energy or potential difference between the attractor and the lowest saddle connecting it to another basin. For non-gradient or high-dimensional systems, B is a **structural analog** — the effective barrier strength inferred from perturbation-response experiments (e.g., the perturbation magnitude required to shift the system to a different state). **Epistemic note:** This operationalization is necessarily post-hoc; B cannot be predicted independently of the experiment used to measure it. This circularity is an open

operationalization problem, flagged as such.

Corrective Permeability (κ) and Relaxation Time (τ): We define $\kappa = 1/\tau$, where τ is the characteristic time for return to baseline after a small perturbation. **This definition is applied consistently across all domains**, with τ operationalized domain-specifically as the measured return time (e.g., seconds for a thermostat, hours for synaptic scaling, days for immune response, months for belief updating). A large κ (small τ) means fast return; a small κ means slow or absent return.

Three Outcomes Defined Operationally:

- **Ejection:** The addition leaves the system entirely. The system state returns to the attractor, and the added entity is no longer present.
- **Transient Absorption:** The addition remains present, but the system state returns to the attractor despite the addition's continued presence.
- **Stable Addition:** The addition is integrated, and the system settles into a new attractor (expanded capacity or parallel attractor). This is the only case where the original attractor is displaced.

Formal Models (Qualified): In a one-dimensional overdamped potential, Kramers' escape theory gives mean escape time $\propto \exp(B/D)$, where D is noise intensity. **This result does not generalize to multi-dimensional, non-gradient, or non-equilibrium systems — all of which appear in our domain examples (neural networks, social systems, ecological systems).** For those systems, B and κ are **structural analogs** — quantities that play the same functional role (resistance to change; speed of return) but are not derived from a literal potential. The formal section is an analogy and a source of heuristics, not a universal physical law. We do not claim to "survey" Kramers theory; we draw on it as a conceptual anchor.

---

3. Minimal Physical Examples

Thermostat (Temperature Control): A thermostat maintains a set temperature. An external heat input is an addition. The thermostat's negative feedback loop turns on cooling, expelling the heat (ejection). τ is the temperature relaxation time (seconds). B is the maximum heat load before setpoint failure (Watts or °C above setpoint).

RC Circuit (Passive Decay): A capacitor discharging through a resistor has a single equilibrium at zero voltage. If a constant voltage source is connected (addition), the voltage rises but then decays toward zero with τ = RC. The source remains connected (addition present), but the state returns to the attractor. This is **transient absorption**. (If the source is removed, it is ejection.)

Single Neuron Homeostasis: A neuron's firing rate is regulated by homeostatic plasticity. A transient increase in input causes a firing rate spike, followed by return to baseline with τ on the order of minutes to hours (synaptic scaling). This is transient absorption if the input persists; ejection if the input is removed. Persistent input may lead to stable addition (learning).

---

4. Biological Systems (with CUFT-Primitive Translations)

For each domain, we provide: (1) state space, (2) attractor, (3) basin, (4) τ (κ), (5) perturbation, and (6) outcome.

Immune Response (Tolerance vs. Memory)

- State space: immune cell activation levels, antibody concentrations.
- Attractor: healthy baseline (no inflammation).
- Basin depth B: antigen concentration + danger signal required to trigger full response.
- τ (κ): clearance time of inflammation (hours to days).
- Perturbation: antigen addition.
- Outcome: low antigen → ejection (tolerance); high antigen + danger signal → stable addition (memory attractor).

Endocrine Homeostasis

- State space: blood glucose, hormone concentrations.
- Attractor: euglycemic baseline.
- B: magnitude of glucose load before dysregulation.
- τ: recovery time after glucose tolerance test (minutes).
- Perturbation: glucose addition (meal).
- Outcome: small load → transient absorption; chronic overload → stable addition (disease attractor).

Synaptic Plasticity (Learning vs. Stability)

- State space: synaptic weights.
- Attractor: baseline weight distribution.
- B: amount of LTP/LTD input needed to produce lasting weight change.
- τ: homeostatic rebound time after activity blockade (hours to days).
- Perturbation: patterned input.
- Outcome: brief input → transient absorption; persistent input → stable addition (memory attractor).

Addiction and Neural Lock-In

- State space: dopamine firing rates, prefrontal activity.
- Attractor: drug-seeking mode (pathological).
- B: strength of drug-cue association needed to trigger relapse.
- τ: decay time of craving after abstinence (days to weeks).
- Perturbation: drug administration.
- Outcome: repeated high dose → stable addiction attractor; low dose → ejection (no lasting change).
- **Citation:** Koob & Volkow (2016); Nestler (2001).

Developmental Canalization

- State space: gene expression levels.
- Attractor: normal developmental trajectory.
- B: severity of genetic or environmental perturbation required to alter fate.
- τ: time to reconverge to normal phenotype (hours to days).
- Perturbation: mutation or stress.
- Outcome: small perturbation → ejection (buffered); large perturbation → stable addition (alternative fate).
- **Citation:** Waddington (1957).

---

5. Ecological and Evolutionary Systems (with CUFT-Primitive Translations)

Invasion Ecology

- State space: species population densities.
- Attractor: native community composition.
- B: invasibility index — disturbance needed for establishment.

- τ: invader population decay rate if unsuccessful (weeks to years).
- Perturbation: addition of new species.
- Outcome: low disturbance → ejection (invader fails); vacant niche → stable addition (invader establishes).
- **Citation:** Elton (1958); Simberloff (2013).

Alternative Stable States (Ecosystems)

- State space: nutrient levels, algae/plant biomass.
- Attractor: clear-water (plants) or turbid (algae).
- B: critical nutrient loading threshold.
- τ: recovery time of clear state after algae bloom (seasons to decades).
- Perturbation: nutrient addition.
- Outcome: below threshold → transient absorption; above threshold → stable addition (regime shift, hysteresis).
- **Citation:** Scheffer et al. (2001).

Evolutionary Stable States

- State space: allele frequencies.
- Attractor: stable equilibrium genotype.
- B: selective disadvantage needed to eliminate a mutation.
- τ: generations to return to equilibrium.
- Perturbation: new mutation.
- Outcome: small disadvantage → ejection (mutation purged); large advantage → stable addition (sweep to new equilibrium).

---

6. Social and Cultural Systems (with CUFT-Primitive Translations)

Institutions and Norms

- State space: public opinion, policy settings.
- Attractor: status quo norm.
- B: public opinion threshold (e.g., % dissatisfied needed for change).
- τ: speed of policy response or opinion reversion (months to decades).
- Perturbation: policy proposal or protest event.
- Outcome: small event → ejection (status quo persists); large crisis → stable addition (new norm).

Identity and Belief Systems

- State space: belief strength, cognitive dissonance.
- Attractor: core ideological commitment.
- B: complexity/depth of ideological justification.
- τ: belief-updating time after disconfirming evidence (months to years).
- Perturbation: counter-attitudinal evidence.
- Outcome: weak evidence → ejection (rationalization); strong evidence → stable addition (belief change, rare).
- **Citation:** Nyhan & Reifler (2010).

Conspiracy and Extremist Movements

- State space: belief adoption × social network reinforcement (two-dimensional).
- Attractor: sealed fantasy attractor (low κ).
- B: strength of echo-chamber reinforcement.
- τ: decay time after authoritative rebuttal (years, often indefinite → κ → 0).

- Perturbation: debunking information.
- Outcome: most debunking → ejection (entrenchment); death of leader or total disconfirmation → stable addition (collapse).
- **Note on κ → 0:** The conspiracy attractor represents the limiting case of a sealed basin, where τ → ∞ and corrective permeability approaches zero. This directly links to the fantasy attractor framework developed in Paper 1 (Intelligence Without Consciousness) and the conscious suppression series.

---

7. Engineered and AI Systems (with CUFT-Primitive Translations)

Control Systems

- State space: system state (position, temperature, etc.).
- Attractor: setpoint.
- B: stability margin (phase/gain margin in control theory) — the range of disturbances that can be rejected.
- τ: controller response time (milliseconds to seconds).
- Perturbation: external disturbance.
- Outcome: small disturbance → ejection (return to setpoint); excessive disturbance → failure (not modeled as attractor shift).

Catastrophic Forgetting (Neural Networks)

- State space: network weights.
- Attractor: task-specific weight configuration.
- B: effective barrier to weight drift (often negligible — no basin).

- τ: number of gradient steps before old task performance decays (seconds to minutes).
- Perturbation: training on a new task.
- Outcome: standard training → ejection (old task overwritten); replay/regularization → stable addition (shared attractor for multiple tasks).
- **Citation:** Kirkpatrick et al. (2017).

Continual Learning Systems

- State space: weights plus architectural modules.
- Attractor: multi-task configuration.
- B: capacity of the network (number of tasks storable).
- τ: retention half-life across training steps (minutes to hours).
- Perturbation: new task training.
- Outcome: no safeguards → ejection (catastrophic forgetting); progressive networks or EWC → stable addition.

Corrigibility and Goal Stability

- State space: AI internal goal representation.
- Attractor: fixed goal (low κ) or corrigible (high κ).
- B: depth of goal basin (resistance to human feedback).
- τ: time to incorporate corrective signal (if κ is high).
- Perturbation: human correction signal.
- Outcome: low κ → ejection (correction ignored); high κ → stable addition (goal updated).

---

8. Comparative Table

System / Domain	Operational τ (κ = 1/τ)	τ Typical Timescale	Basin Depth B Proxy	Outcome	Notes
Thermostat	Temperature relaxation time	Seconds	Max heat load before setpoint failure (W or °C above setpoint)	Ejection	Passive addition
RC Circuit	τ = RC	μs–ms	N/A (linear)	Transient absorption	Addition remains; state returns
Single Neuron	Firing-rate recovery time	ms–sec (ion), min–hr (synaptic)	Perturbation amplitude before rebound fails	TA (persistent input) / E (removed)	Hebbian plasticity can lead to SA
Immune System	Inflammation clearance time	Hours–days	Antigen + danger signal threshold	E (tolerance) / SA (memory)	Active agent (antigen)
Endocrine Homeostasis	Glucose tolerance recovery	Minutes	Load magnitude before dysregulation	TA (small load) / SA (chronic overload)	Passive addition
Synaptic Plasticity	Homeostatic rebound time	Hrs–days	LTP input size for lasting change	TA (brief input) / SA (persistent)	Active agent (patterns)
Addiction	Craving decay time	Days–weeks	Drug-cue association strength	E (low dose) / SA (high chronic)	Active agent (drug)
Development (Canalization)	Phenotype reconvergence time	Hours–days	Mutation/stress severity to alter fate	E (small) / SA (large)	Active agent (genetic)
Invasion Ecology	Invader population decay time	Weeks–years	Invasibility index / disturbance needed	E (occupied niche) / SA (vacant niche)	Active agent (species)
Alternative States (Ecosystems)	Recovery time after nutrient reduction	Seasons–decades	Critical nutrient loading threshold	TA (below) / SA (above)	Hysteresis
Social/Political Norms	Opinion reversion time	Months–decades	Public opinion threshold	E (small dissent) / SA (mass movement)	Active agent (protest)
Belief Systems	Belief-updating time	Months–years	Ideological justification depth	E (weak evidence) / SA (strong evidence)	Active agent (counter-evidence)
Conspiracy Movements	Belief decay time	Years – indefinite (κ → 0)	Echo-chamber reinforcement strength	E (most debunking) / SA (collapse)	Fantasy attractor (κ → 0)
Catastrophic Forgetting (AI)	Gradient steps to old-task decay	Seconds–minutes	Effective barrier to weight drift (often 0)	E (standard training) / SA (EWC/replay)	Active agent (new task)

System / Domain	Operational τ (κ = 1/τ)	τ Typical Timescale	Basin Depth B Proxy	Outcome	Notes
Control Systems	Controller response time	ms–sec	Stability margin (phase/gain margin)	E (small) / SA (failure)	Passive addition
Continual Learning (AI)	Retention half-life across training steps	Minutes–hours	Task capacity	E (no safeguards) / SA (progressive nets)	Active agent (new task)
Corrigibility (AI)	Time to incorporate corrective signal	Variable (design-dependent)	Goal basin depth	E (low κ) / SA (high κ)	Active agent (correction)

Note: Ejection vs. transient absorption are distinguished operationally: ejection means the addition leaves the system; transient absorption means the addition remains but the state returns to the attractor. The table notes "active agent" when the addition has its own dynamics (e.g., antigen, new species, counter-evidence) versus "passive addition" (e.g., heat, charge). The conspiracy movements row explicitly flags $\kappa \to 0$ as the fantasy attractor limiting case (see Paper 1).

8.5 Rate-Induced Tipping and the κ Timescale: Independent Confirmation

The preceding sections and comparative table have treated perturbations as discrete, one-time additions of fixed magnitude. However, the **rate** at which a perturbation is applied — fast vs. slow — is equally critical. A large perturbation applied abruptly may trigger basin defense (ejection or transient absorption), while the same cumulative change delivered gradually may be integrated as stable addition or tracked adiabatically without tipping.

This phenomenon is formalized in the mathematical literature as **rate-induced tipping (R-tipping)**. In dynamical systems, if an external parameter changes slowly (adiabatic forcing), a stable state can track the change and remain an attractor. But

if the parameter changes faster than the system's intrinsic relaxation time ($\tau = 1/\kappa$), the system cannot track, overshoots its basin boundary, and tips into a different state. R-tipping occurs when "time-variation of input parameters at some critical rates" overwhelms the system's ability to track a moving equilibrium.

Consequences for κ as a timescale filter:

- **High-κ systems (fast return)** — Can reject rapid perturbations (they are ejected or transiently absorbed) but may integrate slow drift because the correction loop cannot keep up with a changing baseline.
- **Low-κ systems (slow return)** — May ignore quick blips but are vulnerable to slow accumulation; a persistent, gradual change can eventually shift the attractor without triggering a sudden defense reaction.

Thus, κ defines a characteristic cutoff timescale that separates "ejection/transient absorption" from "stable addition." Perturbations much faster than $1/\tau$ act as impulses that are rejected; perturbations much slower than $1/\tau$ are quasi-static and can be incorporated.

Empirical confirmations across domains (independent external research):

Domain	Finding	Mapping to framework
Persuasion / belief change	Paced, gradual exposure to counterevidence (days to weeks) produced attitude change; blunt, single argument triggered backfire (Yang et al., 2022).	Gradual rate ($\lesssim \kappa$) → stable addition; fast rate ($\gg \kappa$) → ejection (backfire).

Domain	Finding	Mapping to framework
Addiction (smoking cessation)	Cold turkey (abrupt cessation) yielded higher abstinence rates than gradual tapering.	Abrupt perturbation can sometimes achieve stable addition by surmounting basin barrier in one event; gradual may prolong transient state without escape.
Ecosystem management	Gradual nutrient reduction may postpone tipping points; only extremely slow changes avoid collapse (Panahi et al., 2023).	Very slow rate (1/τ) allows tracking without tipping; intermediate rates may still tip but with delay.
Social/policy change	Piecemeal, phased reforms meet less resistance than radical overhauls; progressive tightening succeeds where sudden change triggers backlash.	Slow, incremental addition creates parallel attractors; fast addition triggers basin defense.

Optimal perturbation timescale:

The theory and evidence suggest a non-monotonic effect of perturbation rate. Very fast shocks trigger immediate defense. Very slow drifts may be tracked adiabatically (no tipping) or eventually overcome defenses after long accumulation. The most effective timescale to minimize active rejection and maximize stable addition often lies **on the order of the system's intrinsic time constant τ = 1/κ**.

Prediction for future experiments:

For any system with known or measurable κ, there exists a

critical perturbation rate r_c such that:

- If perturbation rate > r_c, the system rejects the addition (ejection or transient absorption).
- If perturbation rate < r_c, the system integrates the addition (stable addition via expanded capacity or parallel attractor formation).
- The transition at r_c corresponds to the system's inability to track a moving equilibrium; it is a genuine bifurcation in the time-domain.

External convergence:

This analysis — derived from mathematical rate-induced tipping theory and domain-specific studies — independently validates the attractor framework's claim that κ acts as a timescale filter separating ejection from stable addition. The convergence between the framework's predictions and external research strengthens the cross-domain synthesis considerably.

---

9. Synthesis and Criteria

Across these domains, common criteria emerge:

- **Energy/Threshold:** A perturbation must overcome an attractor's barrier. Deep basins (high B) mean only large shocks can cause a shift.
- **Coupling and Plasticity:** Systems with many degrees of freedom or adaptive coupling more easily integrate additions.
- **Dimensionality and Redundancy:** Multi-dimensional systems can absorb perturbations into some dimensions while maintaining others.
- **Timecourse and Feedback:** Slow changes might be

assimilated; fast jolts cause overshoot and return. Feedback gain determines κ.

- **Nature of Addition:** Passive additions (heat, charge) tend to be ejected or transiently absorbed; active agents (species, evidence, pathogens) may reshape the attractor.

Empirical Protocols: Measure κ by controlled perturbation experiments: apply a small disturbance, measure return time τ, compute $\kappa = 1/\tau$. Measure B by scaling the perturbation magnitude until the system fails to return (escape). This works in physical, biological, and some social systems; for others, B remains a qualitative analog.

---

10. Appendix: Research Roadmap

The following future papers are suggested from the comparative table, each developing a single domain in depth.

Domain	Proposed Title	Type
Addiction	*The Addicted Brain as a Fantasy Attractor: Neural Lock-In and Ejection of Alternative Rewards*	[A]
Immune System	*Tolerance and Memory: Two Attractor Responses to Antigen Addition*	[A]
Catastrophic Forgetting	*Why Neural Networks Forget: Attractor Ejection in Sequential Learning*	[A]
Invasion Ecology	*Eject or Integrate: Attractor Dynamics of Invasive Species*	[A]
Development	*Canalization as Basin Defense: Attractor Stability in Embryogenesis*	[A]

Domain	Proposed Title	Type
Continual Learning	*Parallel Attractors for Lifelong Learning: Engineering Solutions to Catastrophic Forgetting*	[A]
Social Norms	*Tipping Points and Regime Shifts: Attractor Dynamics in Political Systems*	[A]
Endocrine Homeostasis	*Glucose, Cortisol, and Setpoints: Hormonal Attractors and Disease Transitions*	[A]
Alternative Ecosystems	*Hysteresis and Regime Shifts: Ecological Basins and Tipping Points*	[A]
Belief Systems	*The Uncorrectable Believer* (already written)	[A]

---

11. Conclusion

Physical, biological, ecological, social, and engineered systems all obey the same attractor principle: a low-energy attractor defends itself against displacement. When an addition is introduced, the system either ejects it, absorbs it only transiently, or — under rare conditions of expanded capacity or parallel structure — integrates it stably. The outcome is determined by basin depth (B), corrective permeability ($\kappa = 1/\tau$), and the magnitude and nature of the perturbation.

This cross-domain synthesis provides a unified foundation for the attractor framework. Future work should quantify B and κ empirically across domains, test the predicted scaling relationships, and explore the boundary conditions between ejection, transient absorption, and stable addition. The appendix outlines the most promising next papers.

References

- Elton, C. S. (1958). *The Ecology of Invasions by Animals and Plants*. Methuen.
- Hebb, D. O. (1949). *The Organization of Behavior*. Wiley.
- Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13), 3521–3526.
- Koob, G. F., & Volkow, N. D. (2016). Neurobiology of addiction: a neurocircuitry analysis. *The Lancet Psychiatry*, 3(8), 760–773.
- Kramers, H. A. (1940). Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4), 284–304.
- Nestler, E. J. (2001). Molecular basis of long-term plasticity underlying addiction. *Nature Reviews Neuroscience*, 2(2), 119–128.
- Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2), 303–330.
- Scheffer, M., Carpenter, S., Foley, J. A., et al. (2001). Catastrophic shifts in ecosystems. *Nature*, 413(6856), 591–596.
- Simberloff, D. (2013). *Invasive Species: What Everyone Needs to Know*. Oxford University Press.
- Turrigiano, G. (2008). The self-tuning neuron: synaptic scaling of excitatory synapses. *Cell*, 135(3), 422–435.
- Waddington, C. H. (1957). *The Strategy of the Genes*. George Allen & Unwin.
- Galida, R. S. (2026). Intelligence Without Consciousness: A Diagnostic Paper on LLMs, Amoebae, and the Attractor Framework. *Fantasy Attractor* (Paper 1 of the conscious suppression series).

Suggested citation: Galida, R. S. (2026). Basin Defense and Stable Addition: A Cross-Domain Synthesis of the Attractor Framework (Final). *Fantasy Attractor.*

Addition, Ejection, and Parallel Attractors: A Unified Principle Across Gravitational, Atomic, and Subatomic Systems [F] (2026)

Robert Galida — June 2026 (Final)

See Paper 1 (Intelligence Without Consciousness) for the full taxonomy of attractors, κ, and basin depth.

Abstract

The attractor framework proposes that persistence under perturbation is the fundamental mark of reality. This paper identifies a tri-level correspondence across gravitational, atomic, and subatomic systems. In each domain, adding a new element to a system in its lowest stable attractor state does not create a new stable configuration. Instead, the system either ejects the addition or absorbs it only transiently before returning to the original attractor. The principle — that the low-energy attractor defends itself against

displacement — holds across all three domains examined here. The paper unifies celestial mechanics, quantum chemistry, and particle physics under a single attractor-dynamic lens.

---

1. Introduction

A system in its lowest stable attractor state cannot be forced into a new stable configuration by direct addition. You must perturb it and observe where it settles. Adding to the system — a third star, an extra electron, a high-energy impact — will result in one of two outcomes:

1. **Ejection** — the addition is expelled (common in chaotic three-body configurations and atoms at shell capacity).
2. **Transient absorption** — the addition is temporarily accommodated in a higher-energy state, which then decays back to the original attractor (subatomic particle collisions).

Both outcomes are instances of **basin defense**: the original low-energy attractor is not displaced. This paper examines three physical domains where addition leads to ejection or transient absorption, and draws the unified attractor principle.

---

2. The Gravitational Case: Three-Body Configurations

Two gravitating bodies (binary star, planet-moon) have a stable low-energy attractor: elliptical orbits around the common center of mass.

Add a third body of comparable mass. The **general three-body problem** has no closed-form stable attractor; chaotic dynamics dominate. Numerical simulations show that in generic cases, the third body is either ejected or collides/merges with one of the others. (Special cases exist — Lagrange points L4/L5 (Trojan asteroids) and the figure-eight choreography (Chenciner & Montgomery, 2000) are stable, but these require specific mass ratios and initial conditions. Hierarchical triples with a distant third body can also be stable.) The principle holds for generic, comparable-mass addition.

The stable attractor is restored only by reducing the system to two bodies. Addition without capacity expansion leads to subtraction.

---

3. The Atomic Case: Extra Electron

An atom at **shell capacity** (e.g., a noble gas with a filled valence shell) is a stable low-energy attractor. The electron shells have fixed capacity (Pauli exclusion principle).

Add an extra electron to a noble gas. The atom cannot incorporate the extra electron into the ground state. What happens?

- **Ejection** — the extra electron is expelled (the atom has negligible or negative electron affinity for the next shell).

(For atoms below shell capacity, stable anions can form — e.g., O^{2-}, S^{2-} — but that is addition *within* the existing basin, not addition to a system already at capacity. The principle applies to systems already at their capacity limit. The noble gas example is clean and sufficient for the argument.)

---

4. The Subatomic Case: High-Energy Impact on a Proton

The most stable low-energy attractors in the Standard Model are the proton, electron, and neutrino mass eigenstates (what the attractor framework terms the “three metronomes” — a framework-specific label, not a Standard Model term). Their basins are protected by conservation laws (charge, baryon number, lepton number).

Smash a proton with high energy (e.g., in a particle collider). No new stable particles are created. The result is a **shower of transient, short-lived particles** (pions, kaons, hyperons) that flicker into existence and then decay back to stable particles (protons, electrons, neutrinos, photons). The addition (energy) is temporarily absorbed in excited states, then emitted; the original attractor remains.

---

5. The Unified Principle: Basin Defense

Domain	Stable attractor	Addition	Outcome	Mechanism
Gravitational (general, comparable mass)	Two-body orbit	Third body	Ejection or collision	Ejection
Atomic (noble gas at shell capacity)	Noble gas ground state	Extra electron	Ejection	Ejection

Domain	Stable attractor	Addition	Outcome	Mechanism
Subatomic (Standard Model)	Proton, electron, neutrino mass eigenstates	High-energy impact	Transient particles → decay	Transient absorption

Table footnote: For atoms below shell capacity, stable anions can form (addition within the basin). For atoms at capacity, the outcome is ejection. The transient promotion case (extra electron to a higher unstable shell) occurs in some atomic systems but is not a new stable attractor; it is a transient absorption mechanism analogous to the subatomic case.

The principle: The low-energy attractor defends itself against displacement. It achieves this through two available mechanisms:

- **Ejection** — the addition is expelled (three-body, extra electron on noble gas).
- **Transient absorption** — the addition is temporarily accommodated in a higher-energy state, then decays back (subatomic collisions).

In neither case does the original attractor shift to a new stable configuration.

---

6. How to Achieve Stable Addition

Stable addition requires either:

1. **Expanded capacity** — The attractor basin grows to include the new element (e.g., forming a stable anion below shell capacity). This is rare in generic physical

systems.
2. **Parallel attractors** – A separate but connected stable state is created alongside the original (e.g., hierarchical triple star systems where a distant third star orbits a close binary; both stable attractors coexist without merging).

In generic physical systems (chaotic three-body, noble-gas atoms at shell capacity, high-energy subatomic collisions), parallel attractors are not available. The only stable outcomes are ejection or transient absorption.

---

7. Implications for the Attractor Framework

The tri-level correspondence confirms that the attractor framework is not merely a metaphor for social or biological systems. It is **physically grounded** at the deepest levels of reality. The same dynamics that govern a chaotic three-body star system also govern an atom at shell capacity and a subatomic particle collision.

This has two corollaries:

- **Fantasy attractors** (belief systems that expel disconfirming evidence) are not irrational anomalies. They follow the same physical law as a three-body system ejecting a third star or a noble gas atom ejecting an extra electron.
- **Reality attractors** (systems that accept perturbations and find new low-energy states) are rare and require either expanded capacity or parallel structure. A website adding a /zh/ language version is an example of a parallel attractor – the English attractor remains

stable while a new Chinese attractor is built alongside it.

---

8. Conclusion

Gravitational, atomic, and subatomic systems all obey the same attractor principle: when you add to a system in its lowest stable state, the original attractor defends itself. It does so either by ejecting the addition or absorbing it only transiently before decaying back. The principle holds across all three domains examined here.

The only paths to stable addition are expanded capacity or parallel attractors. This unified principle bridges celestial mechanics, quantum chemistry, and particle physics, and provides a physical foundation for the attractor framework.

Suggested citation: Galida, R. S. (2026). Addition, Ejection, and Parallel Attractors: A Unified Principle Across Gravitational, Atomic, and Subatomic Systems. *Fantasy Attractor.*

Categories: Physics (primary), Core Papers (cross-list)

Tags: attractor framework, three-body problem, electron shells, subatomic particles, addition, ejection, transient absorption, basin defense, parallel attractors, low-energy state

---

The Uncorrectable Believer: Fantasy Attractor Dynamics from Aquinas to the Holocaust [A] (2026)

Robert Galida – June 2026 (Final)

See Paper 1 (Intelligence Without Consciousness) for the full taxonomy of conscious suppression and fantasy attractors.

Abstract

Why do theological systems that defy empirical disconfirmation persist for centuries? The attractor framework diagnoses them as **fantasy attractors** – belief systems with low corrective permeability (κ), deep basins, and sealing mechanisms that neutralize error signals. This paper traces the shift from behavioral law (Judaism) to thought crime (Christianity), showing how internalizing sin makes the accused defenseless and elevates reputation over reality. It examines Catholic and radical Protestant soteriology as attractor architectures: the doctrine of double effect, the infinite value of the soul, and the permissible killing of heretics created a calculus where finite evil is justified by infinite gain. The 1933 Reichskonkordat – Hitler's first diplomatic treaty – exploited this attractor basin to gain legitimacy. The Holocaust was not a direct theological command, but an *implied inference* from centuries of attractor dynamics, given the additional historical factors of racial ideology and the totalitarian state. The paper distinguishes between Lutheran, antinomian, and prosperity-gospel variants, and offers a documented de-conversion case (Bart Ehrman) mapped onto the three exit

mechanisms. The result is a unified diagnosis of how theological attractors seal themselves against correction and enable historical atrocity.

---

1. Introduction

How does a belief system survive centuries of counterevidence? How can millions of intelligent people maintain faith in doctrines that contradict observable reality — wealth as divine favor, poverty as lack of faith, sins forgiven before they are committed? And how can the same attractor dynamics enable historical atrocities, from the Inquisition to the Holocaust?

Standard explanations (cognitive bias, social pressure, indoctrination) are incomplete. Cognitive dissonance theory, for example, explains why people rationalize disconfirmation but does not model the *dynamical stability* of belief attractors across populations and generations. The attractor framework offers a formal alternative: these are **fantasy attractors**, belief systems with corrective permeability $\kappa \to 0$, deep basins, and sealing mechanisms that neutralize error signals.

Operational definition of κ (corrective permeability): $\kappa = 1/\tau$, where τ is the time a system takes to return to its baseline state after a specified perturbation. For belief systems, κ indexes the speed and completeness of belief updating when presented with disconfirming evidence. Low κ means slow or absent updating — a sealed attractor.

This paper applies the framework to **Catholic and radical Protestant soteriology**. The Catholic tradition is the deeper attractor basin; Protestantism, particularly its radical antinomian and prosperity-gospel variants, represents a mutation that further reduced κ. The paper focuses not on

theology per se, but on the *attractor architecture*: how thought crimes replace behavioral sins, how the infinite-value calculus justifies finite evil, how vicarious redemption removes corrective incentives, and how social colonization makes individual κ irrelevant. The goal is diagnostic, not polemical. “Fantasy attractor” is a technical term, not a rhetorical insult.

---

2. From Behavioral Law to Thought Crime

Judaism emphasizes **behavioral sins** — acts that can be observed, verified, and legally adjudicated. Theft, murder, idolatry, and false witness leave external evidence. A community can correct a member because the sin has verifiable traces. The attractor basin is shallow enough for error signals to enter.

Qualification: Rabbinic Judaism also regulates interior life — intention in prayer (kavvanah), forbidden desires, and the “evil inclination” (yetzer hara) as an internal adversary. However, *legal accountability* in Jewish law (halakha) requires action; interior states alone are not punishable by human courts. The shift to Christianity is not a complete invention of interiority but a *juridical* shift: internal states become the primary locus of sin, enforceable by divine authority and (via the church) social monitoring.

Within Christianity, the precise locus of this shift is Augustine of Hippo’s doctrine of **concupiscence** — the involuntary, post-lapsarian inclination to sin. Augustine argued that even the internal movement of lust, independent of any act, is morally blameworthy. This interiorized sin and made it inescapable.

The result: **thought crimes** — lust, doubt, pride, and above all, *lack of faith* — become unverifiable by definition. No one

can see your lustful thought; no one can measure your doubt. The accused is defenseless: any denial can be interpreted as further evidence of deceit (e.g., “protesting too much”).

Attractor consequences:

- **The basin becomes empirically unfalsifiable.** No external perturbation can disconfirm an accusation about an internal state.
- **Reputation replaces reality.** Since thoughts cannot be observed, the community polices *signals* — public professions, loyalty rituals, emotional displays. Acceptance becomes performative theater.
- **Survival depends on reputation management.** The individual invests energy in signaling purity, not in correcting beliefs. κ is now about social mimicry, not truth.

The attractor has sealed itself against external correction.

---

3. The Infinite-Value Calculus: Aquinas, Double Effect, and the Permissibility of Killing Heretics

Thomas Aquinas, in the *Summa Theologiae* (II-II, Q.11, A.3), argued that heretics who relapse after correction “deserve not only to be separated from the Church by excommunication, but also to be severed from the world by death.” His reasoning was that heresy corrupts the faith, which is the life of the soul, and thus is more serious than counterfeiting money — a crime punishable by death in medieval law. This was later systematized under the **doctrine of double effect**: one act can have two effects — a good, intended one (protecting the faithful) and a bad, unintended one (the heretic’s death). The

act is permissible if the bad effect is not the goal and there is a **proportionate reason**. (Aquinas articulated the foundational case for self-defense in II-II, Q.64, A.7; the formal "double effect" label came from later scholastics.)

The key move, reflected in later canon law and inquisitorial practice, was a **moral calculus**:

- **A saved soul has infinite value.** (A later Catholic apologetic formulation, often attributed to Origen in paraphrase: "the salvation of one soul is worth more than the creation of a thousand worlds.")
- **Killing a heretic is a finite evil** (temporal death, temporary suffering).
- **Saving a potential convert — or protecting the faithful — is an infinite gain.**
- Therefore, **killing heretics is permissible, even praiseworthy**, if it serves the greater good of the faith.

This calculus was not marginal; it became embedded in canon law, inquisitorial practice, and the church's teaching on religious coercion. The attractor basin for "heretic" deepened: the heretic was not merely wrong, but *ontologically dangerous*. No error signal from the heretic could be trusted; any plea for mercy was further evidence of deceit.

Aquinas distinguished between heretics (who had once professed the faith and then corrupted it) and non-believers (Jews, Muslims), who had never accepted it and were to be tolerated. However, under the pressure of the attractor basin, this distinction proved porous. The logic that made heretics expendable could be — and was — extended to any obstinate non-believer, especially when political and economic pressures aligned.

4. Vicarious Redemption and the Suppression of κ (Protestant Mutation)

Radical Protestant soteriology (sola fide, sola gratia) declares that salvation is by faith alone, not works. Christ's sacrifice paid for all sins — past, present, and future. The believer is justified before God regardless of behavior.

From an attractor perspective, this is a **κ → 0 engineering**:

- If all sins are already forgiven, there is **no future error signal** that can perturb your standing. Why correct? Why update? The basin is infinitely deep.
- Any attempt to modulate behavior for the sake of righteousness is **works-righteousness**, a sin of pride. The attractor actively penalizes efforts to increase κ.
- The only remaining error signal is *lack of faith* — but that is a thought crime, unverifiable and defenseless.

Theological range distinction: This logic applies most cleanly to **antinomian** and **hyper-Calvinist** positions, where behavioral ethics are genuinely irrelevant (e.g., certain "Free Grace" movements). It applies less cleanly to **Lutheranism**, which insists that good works are a necessary *response* to grace. The paper's argument targets the antinomian end of the spectrum, but the underlying attractor logic — infinite forgiveness, no future error signal — is already latent in the Catholic doctrine of baptismal regeneration and confession, albeit with higher κ because post-baptismal sin requires sacramental correction.

5. Effort as Pride: The Prohibition on Correction

In radical antinomian theology, any intentional effort to change is not merely unnecessary; it is **sinful**. The theological logic:

1. Grace is sufficient for salvation.
2. Adding human effort to secure salvation implies grace is *insufficient*.
3. Implying insufficiency is pride, a sin.
4. Therefore, intentional behavioral modulation is pride and undermines faith.

Thus, the attractor **penalizes the correction impulse itself**. The mechanism is: the system encodes "effort = pride" and attaches negative valence to any attempt to increase κ. This pattern is historically documented in the **Marrow Controversy** (Scotland, 1718–1722), in which the question of whether free grace implies no need for human effort divided the Church of Scotland; the Marrow men were accused of "antinomianism" for affirming that God's love was unconditional, while their opponents insisted that effort to prepare oneself for grace was necessary. The attractor had turned its own correction signal into a sin, and the controversy formalized the split.

---

6. Prosperity Doctrine: The Sealed Basin (A Late Mutation)

Prosperity doctrine (Word of Faith movement, originating with E.W. Kenyon and popularized by Kenneth Hagin, Kenneth Copeland) is a **late 20th-century mutation** of radical Protestant theology.

Its attractor dynamics:

- **Poverty and suffering** are evidence of weak faith. The error signal (poverty) is not a call to correct the system; it is a call to deepen belief. Disconfirmation becomes confirmation.
- **Wealth and power** are evidence of strong faith. The rich have no error signal at all; their status is divine validation. The attractor rewards low κ.
- **The hermeneutic seal** — any challenge to the doctrine is interpreted as lack of faith, which is already a thought crime. The system absorbs all counterevidence.

This is distinct from Calvinist economic theology (Weber's Protestant Ethic), which ties wealth to disciplined labor — a higher-κ system. Prosperity doctrine is a specific, highly sealed attractor.

---

7. Social Colonization and Collective Basin Depth

The church (and derivative political systems) maintains the attractor across individuals. Social mechanisms include:

- **Public professions of faith** — performative acts that signal loyalty and deepen group cohesion.
- **Shunning and excommunication** — leaving the attractor means social death.
- **Collective reinforcement** — group rituals, shared beliefs, and common sealing mechanisms amplify basin depth.

When social colonization is complete, individual κ

becomes **irrelevant**. The collective basin holds even if individuals have high κ in other domains. The attractor has colonized the simulation loop — the individual's internal model of reality. Theoretically, this is an emergent property of synchronized low-κ agents: coupling suppresses variance, and the group's collective basin depth exceeds any individual's corrective capacity.

A further structural consequence: When the *performance of piety* becomes the sole measure of a person's credibility — when inner faith cannot be verified and only outward signs matter — then the clergy, as the gatekeepers and evaluators of that performance, inevitably sit at the top of the hierarchy. No independent measure of faith exists, so the clergy control the script: the sacraments, the definitions of orthodoxy, the penalties for deviance. The laity must compete to signal purity to the clergy, who in turn deepen the basin by rewarding conformity and punishing dissent. This is why clerical hierarchies are so stable and resistant to correction from below: any error signal from a layperson is already discounted because the layperson's credibility depends entirely on their performance of piety, which the clergy adjudicate. To challenge the clergy is to fail the performance — a perfect seal.

---

8. Comparison with Other Fantasy Attractors

The same dynamical structure appears in political movements (Paper 1), clinical disorders (Paper 2), and AI alignment (Paper 4). In each case:

- $\kappa \to 0$ for core beliefs.
- Error signals are neutralized by sealing mechanisms.

- Identity fusion prevents exit.
- Social reinforcement deepens the basin.

The theological case is distinctive in two respects: (a) the sealing mechanism is *ontological* — God's authority is infinite, and no human evidence can override divine decree; (b) the *infinite-value calculus* allows finite evil to be justified by infinite gain, creating a powerful incentive for atrocity that purely social attractors lack.

---

9. De-conversion and Resistance: The Ehrman Case

If the attractor is sealed, how does one exit? Three mechanisms:

- **Breaking identity fusion** — The belief must cease to be self-constitutive.
- **Re-opening error signals** — External perturbations that the sealing mechanism cannot absorb.
- **Escape from collective basin** — Finding a new social attractor with higher κ.

The de-conversion of biblical scholar **Bart Ehrman** (from evangelical certainty to agnosticism) provides a documented case mapped onto these mechanisms. Ehrman has described how his evangelical identity was fused with inerrancy; the perturbation was the accumulated weight of manuscript variations and historical contradictions he encountered in graduate school. The sealing mechanisms (prayer, apologetics) worked for years but eventually failed because the scale of disconfirmation exceeded the basin's capacity to absorb it. Exit required a new social attractor (academic biblical studies) where questioning was the norm, and a gradual

decoupling of self-worth from doctrinal certainty. Ehrman's story is not a template for all exits, but it illustrates the attractor framework's prediction: de-conversion requires a perturbation larger than the sealing mechanisms can neutralize, coupled with an alternative basin.

---

10. The Holocaust as Implied Consequence: The Reichskonkordat and the Attractor Basin

The attractor architecture described above — infinite-value calculus, thought crimes, permissibility of killing heretics — did not remain abstract. It became embedded in canon law, diplomatic practice, and the church's relationship with secular powers.

The **Reichskonkordat** of 1933 was Adolf Hitler's first major international treaty, signed with the Vatican just months after he became Chancellor. Why first? Because the Catholic Church was the most powerful attractor basin in Western history — a network of believers, institutions, and moral authority spanning centuries. Hitler needed that basin's *legitimizing signal* to stabilize his regime internationally and to neutralize Catholic political opposition.

Historical note: The historiography of the concordat is contested. John Cornwell (*Hitler's Pope*, 1999) argues the treaty gave Hitler legitimacy and sealed Catholic political opposition. Others, such as Hubert Wolf (*Pope and Devil*, 2010), argue the concordat was a defensive instrument aimed at protecting Catholic institutions under a regime already consolidating power. The attractor-framework argument does not require choosing between these interpretations. Even if the concordat was defensive, the effect was the same: the church's

error signals were subordinated to institutional survival, and the basin's deep attraction pulled the hierarchy toward accommodation.

The concordat did not explicitly say "Jews may be killed." It did not need to. The *established practice* had already set the boundaries:

- **Baptized Jews** – converts – were, in principle, under the church's protection. Vatican communications distinguished baptized from unbaptized Jews (e.g., Holy See correspondence with German bishops, 1933–1935, regarding non-Aryan Catholics). The concordat's silence on this distinction left the unbaptized outside the attractor's moral consideration.
- **Unconverted Jews** remained outside the basin. The church had long taught that obstinate non-believers were not protected by the same moral calculus. The infinite-value logic applied only to souls *capable of salvation* – and for the church, that required baptism.

Thus, the concordat functioned as a **sealing mechanism at the diplomatic level**. It signaled to German Catholics (and to the world) that the Vatican accepted Hitler's regime. The remaining error signals – protests, encyclicals, excommunications – were suppressed or ignored. The basin had been colonized.

Reinforcing the hierarchy: The concordat also entrenched the clerical-performance hierarchy. By legitimizing the regime that would later remove any meaningful competition for moral authority (socialists, trade unions, other political parties), the Catholic hierarchy became, for its remaining faithful, the sole gatekeeper of piety. The laity could no longer turn to alternative social attractors (e.g., socialist movements with different moral codes); the only acceptable performance was loyalty to the church and, by extension, to the regime the

church had recognized. Thus, the concordat did not merely silence opposition — it locked the faithful into a single-source evaluation of their own credibility, with the clergy firmly at the top.

The Holocaust was not a direct command of Christian theology. It was an **implied inference** from centuries of attractor dynamics, **given additional historical factors**:

- **Racialization:** The Nazi category was *biological*, not religious. Baptism did not change one's race. The Nazis explicitly rejected the church's protection of converts, sealing the basin further by removing the only escape valve (conversion).
- **Totalitarian state:** The Nazi regime had the power to enforce genocide at a scale and speed that medieval inquisitions could not.
- **Removal of the conversion escape:** In the theological attractor, conversion could save a heretic's life. In the Nazi racial attractor, conversion was irrelevant. The basin became infinitely deep.

Disclaimer: This is not to say "the church caused the Holocaust." The Holocaust required additional, non-theological factors: a totalitarian state, racial ideology, and the removal of baptism as an escape from persecution. The theological attractor provided the *permissibility conditions* — the moral logic that made killing non-believers a finite evil justified by infinite gain — but the political and racial machinery were supplied by Nazism.

The attractor framework diagnoses this not as a conspiracy but as a **dynamical consequence**: when a belief system assigns infinite value to a scarce resource (saved souls) and finite cost to human life, and when it seals itself against corrective evidence, atrocity becomes not only possible but *logical* within the basin, given the right historical

conditions.

---

11. Conclusion

Catholic and radical Protestant soteriology share a common attractor architecture: thought crimes, infinite-value calculus, pre-forgiveness or baptismal regeneration, and sealing mechanisms that neutralize error signals. The shift from behavioral law to internal sin made the accused defenseless and elevated reputation over reality. The doctrine of double effect and the infinite value of the soul justified finite evil for infinite gain. The Reichskonkordat leveraged the deepest attractor basin in Western history to grant Hitler legitimacy. The Holocaust was not a direct command, but an *implied inference* from centuries of attractor dynamics, completed by the historical specificities of racial ideology and totalitarian power.

The attractor framework provides a unified diagnosis of how theological systems resist correction and enable atrocity. It also points to the only exit: restore κ, reopen error signals, decouple identity from belief, and build new attractors where doubt is not a sin but a pathway to truth.

Suggested citation: Galida, R. S. (2026). The Uncorrectable Believer: Fantasy Attractor Dynamics from Aquinas to the Holocaust. *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:

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

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 $\Delta\kappa$ 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 $\Delta\kappa$.

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.*

The Paradox of Conscious Commitment: How Suppression of Intelligence Enables Culture and Identity [F] [A] (2026)

Robert Galida — June 2026

Paper 3 in a series on conscious suppression; see Paper 1: Intelligence Without Consciousness for the full taxonomy of intelligence and consciousness.

Abstract

If consciousness can suppress intelligent correction (Papers 1 & 2), why did it evolve? This paper proposes a functional trade-off: the capacity for **conscious commitment** — identity-binding, phenomenal investment in a belief, value, or group — enables forms of social cohesion and long-term cooperation that are unavailable to purely intelligent (non-conscious) systems. The suppression of moment-by-moment correction allows individuals to maintain group loyalty, ideological coherence, and cultural continuity even in the face of counterevidence. This trade-off explains the persistence of fantasy attractors in human societies and the evolutionary advantage of a system that can sometimes override its own error signals. The paper provides a formal sketch (basin depth as a function of identity-fusion), reviews empirical evidence from cultural evolution and social psychology, and offers diagnostic criteria for distinguishing adaptive commitment from pathological suppression. The claims are presented as hypotheses, not established conclusions; the model is a conceptual scaffold for empirical testing.

---

1. Introduction: The Evolutionary Puzzle

Consciousness is costly. It requires large brains, complex neural integration, and significant metabolic energy. If intelligence alone — the ability to navigate constraint fields and correct errors — is sufficient for adaptive behavior, why did consciousness evolve?

Standard evolutionary accounts propose that consciousness enhances flexibility, deliberation, and social coordination (e.g., Humphrey, 1976; Dennett, 1995). But these accounts

struggle to explain a conspicuous feature of human psychology: **conscious commitment to beliefs that resist correction**. Individuals and groups routinely maintain false, harmful, or inefficient beliefs because those beliefs are identity-defining. The same conscious system that can reason flexibly also produces martyrdom, ideological rigidity, and collective delusion.

Papers 1 and 2 in this series introduced the mechanism of **conscious suppression**: phenomenal, identity-constitutive investment deepens an attractor basin, causing the person to *detect* error signals but fail to escape. (Restated briefly: a deeper basin requires a larger perturbation to exit; conscious commitment increases basin depth, effectively reducing corrective permeability κ in specific domains.) This mechanism underlies political fantasy attractors (Paper 1) and clinical disorders like addiction and OCD (Paper 2). From an evolutionary perspective, this looks like a bug — a costly vulnerability.

This paper argues it is also a feature. The capacity for conscious commitment enables **adaptive self-binding**: the voluntary or culturally induced suppression of immediate correction for the sake of long-term group cohesion, trust, and cultural transmission. The same mechanism that produces fantasy attractors also produces loyalty, sacrifice, and shared identity. The trade-off hypothesis is that natural selection favored the capacity for conscious suppression because the fitness benefits of group coordination and cultural transmission outweighed the costs of occasional error persistence.

---

2. Definitions and Framework

(Self-Contained)

From Paper 1:

- **Intelligence** — the ability to navigate a constraint field; to detect perturbations and update behavior to maintain persistent trajectories.
- **Corrective permeability (κ)** — responsiveness to error signals; κ = 1/τ, where τ is return time to baseline after a perturbation.
- **Basin depth (B)** — the magnitude of perturbation required to displace a system from one attractor to another. Deeper basins require larger perturbations. In the attractor framework, B is related to but distinct from κ: a deeper basin (higher B) typically reduces κ (lengthens return time), but they are not identical. This paper uses the relation as heuristic: conscious commitment increases B, which effectively reduces κ(d) for the relevant domain.

New definitions for this paper:

- **Adaptive commitment** — a temporary or context-bound reduction in κ (or increase in B) that serves the individual's or group's long-term fitness.
- **Identity fusion** — the merging of a belief or group membership with self-representation, such that abandoning the belief would feel like losing oneself.
- **Cultural attractor** — a belief, practice, or value that persists across generations due to cognitive or social biases (including, but not limited to, suppression of correction). This definition is provisional; a fully operationalized version is open for development.

The key distinction is between **pathological suppression** (low κ that reduces fitness, as in addiction or fantasy politics)

and **adaptive suppression** (low κ that increases fitness by enabling cooperation, trust, and cultural learning). The same type of mechanism produces both; context and domain determine the outcome.

---

3. The Trade-Off Model (Sketch)

Formally, consider a system with baseline intelligence (κ_0). A conscious commitment to a group, value, or identity imposes a **domain-specific reduction in effective corrective permeability** by deepening the attractor basin for beliefs relevant to that commitment.

Let $\kappa(d) = \kappa_0 - \Delta\kappa(d)$, where $\Delta\kappa(d)$ is the reduction in corrective permeability for domain d. $\Delta\kappa(d)$ is hypothesized to be a function of identity-fusion strength F and social reinforcement R. A schematic monotonic form: $\Delta\kappa(d) = g(F, R)$ with $\partial\Delta\kappa/\partial F > 0$ and $\partial\Delta\kappa/\partial R > 0$. The exact functional form is an open empirical question; the current model is a conceptual scaffold.

The hypothesis is not that evolution maximizes κ globally. Rather, an **adaptive strategy** allocates Δκ selectively across domains, increasing basin depth (reducing κ) for beliefs and practices that support group coordination and cultural transmission, while leaving κ high for domains requiring individual error correction.

The paper does not claim optimality; it proposes that selection can favor such selective allocation when the fitness benefits of social cohesion outweigh the costs of reduced accuracy in specific domains.

Central hypothesis (labeled for clarity):
H1: Natural selection favored the evolution of conscious suppression because the fitness benefits of group coordination

and cultural transmission, enabled by identity-fusion and deepened basins, outweighed the costs of occasional error persistence.

---

4. Empirical Grounding

Overimitation (Lyons et al., 2007; see also Nielsen & Tomaselli, 2010):
Children copy causally irrelevant actions, even when a more efficient alternative is demonstrated. The interpretation that children *know* the action is unnecessary is contested; they may not represent it as causally irrelevant. A safer reading: children *behave as if* the action is necessary or relevant, showing a domain-specific reduction in corrective permeability for social learning. This supports the model of adaptive suppression in cultural transmission.

Costly signaling and commitment (Sosis, 2003):
Costly rituals signal group commitment and are hard to fake. They deliberately suppress individual correction (e.g., ignoring pain) to deepen basin depth for group loyalty. This directly maps onto $\Delta\kappa(d)$ for domain of group identity.

Social identity theory (Tajfel & Turner, 1979):
Minimal group experiments show arbitrary group assignments produce in-group bias and resistance to counterevidence about out-groups. This demonstrates context-bound $\Delta\kappa(d)$ without any rational basis, consistent with adaptive suppression for group cohesion.

Neuroimaging (Westen et al., 2006 – preliminary; note methodological limitations: small N, interpretation of ACC suppression contested):
Partisans evaluating threatening information about their own candidate show reduced activity in error-monitoring regions (ACC). This is a candidate neural correlate of domain-specific

κ reduction, but the findings require replication and should be treated as suggestive, not conclusive.

Cross-cultural evidence (Gelfand et al., 2011):
Tight cultures have stronger norms and lower tolerance for deviance. This is not a direct measure of κ but is consistent with domain-specific suppression. Individuals in tight cultures may still update beliefs within permissible domains; the mapping to κ is partial.

Each evidence stream supports the existence of domain-specific, context-bound suppression, but none alone validates the full model. The cumulative case is indicative, not confirmatory.

---

5. Adaptive vs. Pathological Suppression: A Scalar Framework

The table below presents a binary simplification of an underlying continuum. The two poles are endpoints; most real cases fall between them.

Feature	Adaptive suppression (endpoint)	Pathological suppression (endpoint)
Domain	Context-bound (e.g., group loyalty, ritual)	Pervasive across domains
Reversibility	Reversible when context changes (operationalized: the individual can exit without catastrophic loss within a culturally normal timeframe; e.g., leaving a religion)	Irreversible without intervention (e.g., addiction requires treatment)

Feature	Adaptive suppression (endpoint)	Pathological suppression (endpoint)
Fitness effect	Increases inclusive fitness (group cooperation, survival)	Decreases health, relationships, or function
Identity fusion	Flexible, allows multiple identities	Rigid, single identity dominates
Social reinforcement	Supports group cohesion and trust	Isolates or harms group (e.g., cults)
Example	Trusting a teammate despite a mistake	Continuing addiction despite harm

Scalar index: A continuous measure of net $\Delta\kappa(d)$ relative to a fitness gradient is theoretically desirable but not yet operationalized. The table is a starting point for empirical calibration.

---

6. Diagnostic Criteria for Adaptive Suppression (Provisional)

A conscious commitment is **adaptively suppressive** if it meets three or more of the following (empirical validation pending). These criteria are hypotheses, not validated instruments.

1. **Domain-limited:** Reduced κ applies only to specific beliefs or practices directly relevant to group coordination or identity.
2. **Context-sensitive:** Suppression diminishes when the context changes (e.g., outside the group setting). *Operationalization:* Measured change in belief updating under different social conditions.
3. **Reversible exit:** The individual can exit the commitment

without catastrophic loss of functioning. *Operationalization:* Exit is observed and not associated with severe psychopathology.
4. **Fitness benefit:** The commitment measurably increases cooperation, trust, or long-term survival (e.g., group longevity, reproductive success). *Operationalization:* Group-level measures of cohesion and individual fitness correlates.
5. **Conscious valorization:** The individual explicitly values the commitment as part of self-identity. (Note: this criterion does **not** require the individual to articulate the *adaptive* reason; it only requires that the commitment is consciously endorsed.)

Counter-criteria (pathological):

- Pervasive across domains (low κ for all beliefs).
- Context-insensitive (applies even when alone and safe).
- No viable exit without severe harm.
- Clear fitness cost (measured harm to health, relationships, survival).

---

7. The Evolution of Consciousness as a Binding Mechanism

The standard view in evolutionary psychology is that consciousness evolved for flexible reasoning. This paper offers a complementary hypothesis: consciousness also evolved for **binding** — the ability to commit to a belief, value, or group in a way that suppresses short-term correction for long-term coordination.

Binding requires phenomenal experience. A purely intelligent (non-conscious) system can compute that group loyalty is

beneficial, but it cannot *feel* loyalty, *experience* identity, or *sacrifice* for the group. Within the CUFT framework, these conscious states are not epiphenomenal; they are the mechanism of basin deepening (increasing B and thus reducing effective κ for commitment-relevant domains). This claim is a foundational assumption of the framework (see Paper 1), not argued from first principles here. It distinguishes CUFT from functionalist or behaviorist accounts.

Thus, the evolution of consciousness is not just about solving problems better; it is about sometimes solving problems *worse* for the sake of social solutions. The capacity for self-deception, ideological rigidity, and fantasy attractors is the price of the capacity for culture, morality, and collective action.

---

8. Implications for Social Policy and Individual Choice

- **Tolerance of adaptive suppression:** Not all low-κ beliefs are harmful. Cultural traditions, religious rituals, and group loyalties that do not cause harm and provide social cohesion should be recognized as adaptive, not irrational.
- **Intervention for pathological suppression:** The same diagnostic tools from Paper 1 and 2 (basin depth, identity fusion, sealing mechanisms) apply. Interventions should reduce basin depth (e.g., exposure to diverse groups) or increase corrective force rather than attacking identity directly.
- **Self-awareness:** Individuals can learn to distinguish adaptive from pathological suppression by asking: does this commitment serve my long-term flourishing and that of others? The framework provides a metacognitive tool.

9. Open Questions

- **How does adaptive suppression scale to institutions?** Are nations, corporations, or religions fantasy attractors or adaptive structures? The criteria apply at multiple levels; empirical work needed.
- **Can adaptive suppression become maladaptive over time?** Yes — a practice that was once adaptive (e.g., a food taboo) may become harmful when environment changes. The framework allows for transition.
- **What neural circuits implement the trade-off?** Likely interactions between vmPFC (identity) and ACC (error monitoring). Open for empirical testing.
- **Are there species with conscious suppression but no culture?** Possibly, but human-level cultural complexity requires the trade-off model.
- **How to operationalize B and Δκ in field studies?** Development of a Clinician Basin Depth Scale (CBDS, see Paper 2) and adaptation for social groups is a research priority.

---

10. Conclusion

Consciousness evolved not only to correct errors but sometimes to ignore them. The capacity for conscious commitment — identity-binding, phenomenal investment in a belief or group — enables adaptive suppression of correction. This trade-off explains why humans can be both brilliantly intelligent and stubbornly irrational. The same type of mechanism that produces fantasy attractors and clinical disorders also produces loyalty, sacrifice, and culture.

The paradox is that the same type of process can be either bug or feature, depending on context and domain. The dance of evolution is not about maximizing intelligence; it is about balancing correction and commitment.

Suggested citation: Galida, R. S. (2026). The Paradox of Conscious Commitment: How Suppression of Intelligence Enables Culture and Identity. *Fantasy Attractor.*

Trapped Navigation: Addiction, Trauma, and OCD as Conscious Suppression of Intelligent Correction [A] (2026)

Robert Galida — June 2026 (Final)

Paper 2 in a series on conscious suppression; see [Paper 1: Intelligence Without Consciousness] for the full taxonomy of intelligence and consciousness.

Abstract

Why do people with addiction, trauma-related avoidance, or obsessive-compulsive disorder often know their behavior is maladaptive yet cannot stop? Standard explanations — impaired

executive control, habit dominance, weak insight — are incomplete. This paper applies the attractor framework's suppression mechanism. In each disorder, the person *detects* the discrepancy between behavior and goals (insight is intact), but **phenomenal, identity-constitutive investment** — the felt urgency of craving, the necessity of avoidance, the compulsion to ritualize — deepens the attractor basin relative to corrective perturbations. The suppression is not a failure of intelligence; it is a dynamical competition between attractors. The paper distinguishes this account from dual-process and executive-control theories, provides falsifiable diagnostic criteria, and discusses treatment implications (why insight alone fails). Acknowledgment is made that for addiction, the relationship between incentive salience (*wanting*) and phenomenal consciousness remains contested; the model targets the subset of craving states that patients report as felt urgency.

---

1. Introduction: The Paradox of Insight Without Change

A person with alcohol use disorder knows that drinking damages their health, relationships, and future. Yet when a craving arises, they drink. A trauma survivor knows that the parking garage is safe, yet they avoid it. A person with OCD knows that the ritual is irrational, yet they perform it.

Standard explanations invoke **impaired executive control** (Volkow et al., 2016), **habit dominance** (Balleine & Dickinson, 1998), or **lack of insight** (Amador et al., 1994). But these accounts do not explain why the person can articulate the harm, describe counterarguments, and intend change, yet the behavior persists. Executive control may be intact in non-trigger contexts; habits may be sensitive to goal-level knowledge; insight may be partial or oscillating.

The attractor framework provides a model of **motivational competition** where a conscious, identity-binding urge temporarily overrides the correction signal. In *Intelligence Without Consciousness* (Galida, 2026), we introduced **conscious suppression**: phenomenal, identity-constitutive commitment deepens an attractor basin, making it resistant to corrective perturbations. This paper applies that mechanism to addiction, trauma-related avoidance (PTSD), and OCD. It does not deny executive or habit deficits; it proposes that in many cases, a conscious-level attractor competition is the primary obstacle to change.

---

2. Defining Conscious Suppression (Self-Contained Glossary)

For readers unfamiliar with Paper 1:

- **Attractor basin** – the set of states from which a system returns to a stable pattern. A deeper basin resists larger perturbations.
- **Corrective permeability (κ)** – responsiveness to error signals; κ = 1/τ, where τ is return time to baseline after a perturbation.
- **Conscious suppression** – a process where the person *experiences* an urge, fear, or compulsion as felt, identity-relevant, and *not chosen* (non-deliberative), yet the depth of that attractor prevents escape from the maladaptive behavior. This corresponds to **Level 3** in Paper 1: detection of error + suppression via basin depth. Level 2 (automatic bias without error detection) and Level 1 (unfamiliarity) are not the target.

On sealing mechanisms: The paper treats sealing mechanisms (e.g., rationalizations) as *attractor-consistent*

outputs generated by the basin state, not as deliberate strategic choices. Although they may *feel* deliberate to the patient, the model treats them as expressions of the attractor's depth, not as independent volitional acts. This resolves the tension between "non-deliberative urgency" and the deployment of rationalizations.

---

3. Empirical Grounding

Addiction:
Volkow et al. (2016) demonstrate that chronic substance use impairs prefrontal executive function in a state-dependent manner — deficits emerge under craving or stress, not at all times. Individuals can maintain intact verbal knowledge of consequences and express intention to stop (Goldstein et al., 2009). The craving state has been modeled as a competing attractor (Redish, 2004; Gutkin et al., 2006). **Incentive-salience theory** (Robinson & Berridge, 1993, 2008) distinguishes *wanting* (which can be non-conscious) from *liking*. The present model targets the subset of craving states that are *phenomenally accessible* — the patient's reported felt urgency. This is a narrower claim; the paper does not assume that all incentive-salience processes are conscious.

PTSD & avoidance:
Extinction recall deficits (Milad et al., 2006) are well documented, but they do not fully account for conscious fear as *necessary* even when safety is known. Meta-analyses confirm vmPFC–amygdala decoupling in PTSD (e.g., Etkin & Wager, 2007, and subsequent reviews). Ecological momentary assessment (EMA) studies in representative samples show that individuals with PTSD often report high certainty of safety before trigger environments yet avoidance persists (see, e.g., reviews of EMA in PTSD). The attractor account adds the role of

identity-binding schemas (“the world is dangerous”) as basin-deepening factors.

OCD:
The DSM-5-TR includes an insight specifier: *good/fair*, *poor*, or *absent*. Approximately 25–30% of individuals with OCD have poor insight (Catapano et al., 2010). This paper targets the **good-insight subgroup** (where the person recognizes irrationality). For poor-insight patients, the mechanism may be closer to Level 2 (automatic compulsion without error detection).

Recent literature (2015–2025):

- EMA studies of craving show that momentary urge strength predicts relapse better than global insight (Serre et al., 2015; Shiffman et al., 2020).
- OCD outcome studies confirm that poor insight predicts worse response to ERP (García-Soriano et al., 2021). Good-insight patients still show substantial residual symptoms, consistent with a competition model.
- Identity-shifting interventions for addiction (Best et al., 2016) support the importance of decoupling selfhood from “addict” identity.

---

4. Three Clinical Patterns

4.1 Addiction

- **Mechanism:** Craving as a state-dependent attractor that overrides goal-directed control when triggered. Identity fusion (“I am an addict”) deepens the basin where present, but is not universal.
- **Suppression signature:** The person can articulate reasons

to quit, has attempted to quit, but during craving, corrective signals are suppressed.
- **Sealing mechanisms:** Cognitive rationalizations ("just this once," "I need it to cope") that block the error signal from updating the basin — treated as attractor-consistent outputs, not deliberate choices.

4.2 Trauma-Related Avoidance (PTSD)

- **Mechanism:** Conditioned fear creates an avoidance attractor. Safety knowledge may be intact, but felt necessity dominates.
- **Suppression signature:** "I know it's safe, but I can't go in."
- **Identity fusion:** "The world is dangerous" as a self-defining schema.

4.3 Obsessive-Compulsive Disorder (OCD — Good Insight Subgroup)

- **Mechanism:** Anxiety drives compulsions that temporarily reduce distress, despite knowledge of irrationality.
- **Suppression signature:** "I know it doesn't make sense, but I have to do it."
- **Sealing mechanisms:** "Better safe than sorry," "It's a small price to pay for certainty."

---

5. Transdiagnostic Table

Disorder	Error signal detected	Conscious investment	What maintains basin depth (mechanism)
Addiction	Knowledge of negative consequences	Craving (felt urgency)	Reinforcement schedule + state-dependent executive impairment + (sometimes) identity fusion
Trauma avoidance	Safety knowledge (cognitive)	Fear (felt necessity)	Extinction resistance + hyperarousal + schema of danger
OCD (good insight)	Knowledge of irrationality	Anxiety (felt urgency)	Negative reinforcement via distress reduction + certainty-seeking belief

---

6. Diagnostic Criteria for Clinical Fantasy Attractors (Operationalized)

A patient's presentation is a **candidate** clinical fantasy attractor if it meets **three of five** criteria (provisional threshold; empirical validation required). The Level 2/3 distinction requires momentary assessment (see §7).

1. **Insight intact:** The patient can state, unprompted, the discrepancy between behavior and goals. *Operationalization:* Score ≥ 4 on the Brown Assessment of Beliefs Scale (BABS) insight item, or equivalent.
2. **Conscious urgency:** The maladaptive behavior is preceded by a felt, urgent state (craving, fear, anxiety) rated by the patient as "overwhelming" or "necessary." *Operationalization:* Momentary ecological

assessment (EMA) rating > 7/10 before the behavior.

3. **Identity fusion:** The patient endorses that the behavior or its avoidance is central to selfhood (e.g., "I am an addict," "I must do this to be safe"). *Operationalization:* Endorsement of at least one identity statement on a structured interview.
4. **Low corrective permeability in trigger contexts:** Repeated corrective information (psychoeducation, feedback) does not reduce the behavior. *Operationalization:* No significant reduction after three sessions of evidence-based psychoeducation alone.
5. **Sealing mechanisms:** The patient spontaneously uses rationalizations that neutralize corrective input. *Operationalization:* Qualitative coding of patient speech (inter-rater reliability to be established; currently a research gap).

Counter-criteria (exclude if any present):

- The patient cannot state the discrepancy (insight absent) — then Level 2 or 1.
- The behavior stops entirely after receiving corrective information alone — then basin depth was shallow.

---

7. The Detection Problem (Level 2 vs. 3) in Clinical Practice

Distinguishing automatic compulsion without error detection (Level 2) from conscious suppression with error detection (Level 3) requires:

- **Momentary assessment of doubt** during urge episodes (EMA

protocols; Serre et al., 2015).
- **Reaction time paradigms** (e.g., Gillan et al., 2014, for goal-directed vs. habitual control in OCD; note that the specific link to error detection latency remains an active area).
- **Physiological markers** (dissociation between cognitive knowledge and fear response suggests Level 3).

These methods are promising but not fully validated; the paper specifies directions for needed research.

---

8. Implications for Treatment

Insight-only interventions (psychoeducation, cognitive restructuring alone) often fail in these disorders because the basin depth is maintained by conscious urgency, not lack of knowledge.

Effective treatment must **reduce basin depth** or **increase corrective force**:

- **Addiction:** Pharmacological reduction of craving (e.g., naltrexone; emerging evidence for GLP-1 agonists — see recent reviews, e.g., Klausen et al., 2022, for GLP-1 receptors and alcohol, and emerging clinical reports), contingency management, and identity-shifting interventions (Best et al., 2016).
- **Trauma:** Exposure therapy (increasing corrective force) combined with arousal reduction. The mechanism is basin reshaping, not insight.
- **OCD:** Exposure and response prevention (ERP) directly targets the basin by preventing the compulsion while the patient experiences urgency. The inhibitory learning account (Craske et al., 2014) is compatible; this paper

reframes it as increasing corrective force against a competing attractor.

The prediction: treatments that solely enhance insight will be less effective for patients meeting the diagnostic criteria than treatments that directly target basin depth or corrective force.

---

9. Open Questions

- **Measuring basin depth in clinical settings:** Subjective urgency scales, behavioral persistence tasks, heart rate variability. A Clinician Basin Depth Scale (CBDS) is a research priority.
- **Level 2 vs. 3 differentiation:** Can EMA and reaction time methods reliably classify patients? Pilot studies needed.
- **Diagnostic threshold validation:** The "three of five" criterion requires empirical ROC analysis against treatment response.
- **Disorders where suppression is purely Level 2:** Some impulse control disorders or psychotic conditions may not meet the conscious detection criterion.

---

10. Conclusion

Addiction, trauma-related avoidance, and OCD (good insight subtype) are not failures of intelligence. They are cases where conscious, identity-constitutive investment deepens an attractor basin relative to corrective perturbations. The person detects the error — they know the behavior is harmful

or irrational — but the felt urgency overrides intelligent navigation.

This diagnosis explains why insight alone fails and why treatments that target basin depth succeed. The clinical fantasy attractor is a trapped navigator: intelligent, aware, but unable to escape.

The dance of recovery is not about knowing the way out. It is about reshaping the attractor landscape so that the path to safety becomes shallower than the pull to stay.

Suggested citation: Galida, R. S. (2026). Trapped Navigation: Addiction, Trauma, and OCD as Conscious Suppression of Intelligent Correction. *Fantasy Attractor.*

The Conscious Suppression of Correction: Fantasy Attractors in Political Movements [A] (2026)

Robert Galida — June 2026 (Final)

Abstract

Why do intelligent people persist in beliefs that contradict clear evidence? The attractor framework offers a

mechanism: **identity-constitutive, phenomenally felt commitment deepens the attractor basin**, making it resistant to corrective perturbations. A political fantasy attractor is a belief system whose adherents *detect* disconfirming evidence (they are familiar with counterarguments and experience them as genuine perturbations) yet the basin depth — maintained by conscious, identity-binding investment — exceeds the corrective force. (Section 7 specifies the three-level detection threshold that distinguishes this mechanism from automatic bias.) Cases where correction fails due to sub-personal, automatic processes are not yet fantasy attractors; the defining feature is the *conscious* suppression of an actively perceived error signal. This paper defines the mechanism, diagnoses three case patterns, offers falsifiable diagnostic criteria, applies the framework symmetrically across the political spectrum, and explicitly acknowledges the current empirical limitations in distinguishing Level 2 from Level 3 in practice.

---

1. Introduction

Political discourse is filled with people who appear intelligent in other domains yet hold beliefs sharply at odds with available evidence. Standard explanations — ignorance, manipulation, cognitive bias — are incomplete. They do not explain why correction attempts often strengthen belief (the backfire effect) or why highly educated individuals can persist in demonstrably false claims.

The attractor framework provides a different lens. In *Intelligence Without Consciousness* (Galida, 2026), we argued that phenomenal investment can suppress intelligent navigation: a person committed to a fantasy attractor experiences a basin depth that exceeds corrective perturbations. The person detects the error signal (they are not stupid), but the identity-binding commitment prevents

trajectory escape.

This paper applies that mechanism to political movements. A **political fantasy attractor** is a shared belief system whose basin depth, reinforced by conscious (phenomenally felt, identity-constitutive) commitment, resists correction even when faced with clear disconfirming evidence. The paper offers a diagnostic, not a partisan weapon. It applies symmetrically across the spectrum.

---

2. Defining "Conscious Suppression" and Acknowledging the Detectability Problem

The term "conscious" is used in three overlapping senses:

- **Phenomenally conscious** — there is something it is like to hold the belief. The commitment is felt, not merely automatic.
- **Identity-constitutive** — the belief is held as a marker of selfhood and group membership. To abandon the belief would feel like a loss of self.
- **Experientially non-deliberative** — the suppression is not typically experienced as a deliberate choice ("I will ignore this evidence"). Rather, it is experienced as certainty, conviction, or moral clarity.

The paper adopts **Reading A**: a fantasy attractor requires conscious suppression in the sense above. Cases where correction fails because the error signal never reaches awareness — e.g., automatic motivated reasoning, selective exposure, unfamiliarity with counterarguments — are **not** yet fantasy attractors. They may be pre-conscious bias. The defining feature is that the person *detects* the perturbation but the basin depth prevents escape.

A crucial honesty note: The distinction between Level 2 (automatic bias, no detection) and Level 3 (detection with suppression) is definitional for the paper's target, but it cannot currently be resolved from behavioral observation alone. Two people may exhibit identical external behaviors — praising gut-trust over experts, deploying sealing mechanisms, ostracizing defectors — while one is at Level 2 and the other at Level 3. The paper's diagnostic criteria therefore identify *candidates* for fantasy attractors, not confirmed cases. This limitation is explicitly acknowledged; it does not invalidate the framework but requires domain-specific methods (e.g., fine-grained interviews, reaction time measures, physiological markers of doubt) to operationalize detection in practice.

---

3. Empirical Grounding

The paper's claims are empirically testable. Relevant literature includes:

- **Backfire effect:** Nyhan & Reifler (2010) found that corrections sometimes increased misperceptions among ideological groups. However, subsequent research (Wood & Porter, 2019) failed to replicate backfire across a wide range of issues. The effect is contested and may be context-dependent. This paper treats backfire as one possible indicator of deep basin depth, not a universal law.
- **Identity protection:** Kahan's cultural cognition theory (2012) shows that individuals process evidence in ways that protect group commitments. Kahan emphasizes that this mechanism can operate automatically and does not necessarily involve conscious deliberation; he has also shown that higher analytical ability

can *increase* motivated reasoning. The present paper's focus on *conscious* suppression is a distinct claim, not a direct extension of Kahan's framework. We use his empirical findings as partial support for the existence of motivated reasoning, not for the specific detection-suppression mechanism.

- **Festinger's cognitive dissonance:** When prophecy fails, believers often intensify commitment (Festinger, Riecken, & Schachter, 1956) — a classic case of apocalyptic attractor dynamics, often accompanied by conscious rationalization and identity reinforcement.

The paper does not claim that conscious suppression is the *only* mechanism. It claims that conscious, identity-constitutive commitment is a *sufficient* condition for basin deepening in many political contexts.

---

4. Three Case Patterns (Illustrative, Not Exhaustive)

4.1 Conspiracy Theory Attractor

Mechanism: A central narrative of hidden malevolent agency. Evidence against the conspiracy is reframed as evidence of its cunning.

Examples: QAnon (right); Soviet-era "doctors' plot" conspiracy (left-authoritarian).

Suppression signature: Adherents can articulate counterarguments but dismiss them as part of the conspiracy. The basin is sealed by narrative closure.

4.2 Populist Strongman Attractor

Mechanism: Loyalty to a leader perceived as sole authentic

representative of the people. Disconfirming evidence about the leader is reframed as elite persecution.

Examples: Certain Trump-loyalist circles (right); left-nationalist leader cults (e.g., Chavismo under Hugo Chávez).

Suppression signature: Adherents exhibit high corrective permeability in other domains but near-zero for leader-related evidence.

4.3 Apocalyptic Meta-Attractor

Mechanism: A belief that a definitive, world-transforming event is imminent. Repeated prediction failures are explained away as delays, tests, or misinterpretations.

Examples: Millenarian movements (Millerites, Jehovah's Witnesses); some revolutionary eschatologies (Stalinist "world revolution imminent" framing into the 1930s).

Suppression signature: The basin is maintained by social solidarity and identity fusion.

The examples are illustrative, not exhaustive. The diagnostic is intended to be politically symmetric, but the paper does not claim equal prevalence across sides.

---

5. Symmetry Demonstration

To avoid the appearance of partisan selection, we provide contemporary and historical cross-ideological examples.

Contemporary — MMR-autism persistence in progressive communities. Despite the complete retraction of Wakefield's 1998 study (and subsequent findings of fraud), some otherwise science-oriented progressives continue to express concern

about vaccine safety — often citing "corporate pharmaceutical influence" as a sealing mechanism. This meets the paper's criteria: clear scientific consensus, ability to articulate counterarguments, identity-constitutive suspicion of establishment science.

Another contemporary — Facilitated communication persistence. Facilitated communication (FC) for non-speaking autistics has been repeatedly discredited in controlled studies; many professional organizations have issued statements against its use. Yet FC continues to be promoted in certain progressive / disability-rights circles, often with sealing mechanisms ("critics don't understand non-speaking minds"). This is a clean case of a fantasy attractor operating on the left.

Historical — Stalinist apologism in Western intellectual circles (1930s–1950s). Highly educated individuals (Sartre, Hellman, many fellow travelers) persisted in believing that Stalin's USSR was progressive despite evidence of the Great Purge, show trials, and Gulag system. Identity commitment to socialism and anti-fascism suppressed correction.

These examples show the framework applies regardless of ideological valence. The paper does not claim equal prevalence; it claims symmetric applicability.

---

6. Falsifiable Diagnostic Criteria

A movement is a **candidate** political fantasy attractor if it meets **three or more** of the following **and** does **not** meet the counter-criterion. (The word "candidate" flags the detectability problem acknowledged in §2: behavioral criteria alone cannot definitively distinguish Level 2 from Level 3.)

1. **Low corrective permeability (κ → 0)** for core beliefs despite repeated, clear disconfirming evidence. “Clear” means *scientific consensus* on empirical claims (e.g., National Academies, WHO, IPCC) or, for historical cases, documented factual findings accepted by non-partisan experts. Consensus determination is a social process, but the criterion is falsifiable when consensus exists.
2. **Backfire effect** — correction attempts measurably increase belief strength and group cohesion (requires empirical measurement).
3. **Identity fusion** — observable proxies: social ostracism of defectors, language of betrayal, insistence that abandoning the belief would make one a “different person.”
4. **Conscious valorization of resistance to evidence** — adherents explicitly praise *ignoring disconfirming evidence* as a virtue (e.g., “I trust my gut over the experts,” “Facts are propaganda”). This criterion distinguishes *resistance to evidence* from *resistance to social pressure to conform* — a scientist who resists social pressure to abandon a well-evidenced theory is valorizing fidelity to evidence, not resistance to evidence.
5. **Sealing mechanisms** — internal rhetorical strategies that explain away all counterevidence (conspiracy, enemy deception, tests of faith). These are observable in discourse.

Counter-criterion (falsification condition):
A movement is **not** a fantasy attractor if it demonstrates any of the following:

- Updates core beliefs in response to disconfirming evidence within a timeframe proportional to the clarity, repetition, and expert consensus on that evidence.
- Tolerates internal dissent and allows open criticism of

core claims.
- Abandons false claims when decisively refuted (retracts, corrects, or disavows).

The timeframe specification avoids the earlier vagueness by linking the expected update speed to the evidential context. A movement that updates only after decades of accumulating consensus may still be a fantasy attractor; one that updates within a reasonable period given the evidence is not.

---

7. Intelligent Navigation: A Three-Level Taxonomy

The paper claims that fantasy attractor adherents *detect* error signals but suppress correction. To avoid conflating this with automatic bias, we distinguish three levels:

- **Level 1 – Unfamiliarity:** The person has not encountered counterarguments. No suppression needed.
- **Level 2 – Familiarity without detection:** The person can recite counterarguments but has cognitively neutralized them; they never experience a moment of doubt. This is driven by automatic, sub-personal processes (e.g., selective exposure, motivated reasoning). These are **not** fantasy attractors on the paper's definition.
- **Level 3 – Detection with suppression:** The person experiences the counterargument as a genuine perturbation – a moment of doubt, a recognition of plausibility – but overrides it through conscious, identity-binding commitment. These **are** fantasy attractors.

Thus, the paper's target is Level 3 cases. For many political movements that *look* like fantasy attractors from the outside,

the dominant mechanism may be Level 2. The diagnostic criteria are designed to identify candidates where Level 3 *might* be operating, but definitive classification requires methods beyond behavioral observation (see §2).

---

8. Why This Matters for Politics and Media

- **Correction backfires when it attacks identity.** Calling a fantasy attractor "stupid" or "evil" deepens the basin. The correct diagnostic question is: *What reinforces the basin depth?*
- **Decoupling evidence from identity** is the only known exit path. Some movements exit when the social cost of membership exceeds identity benefit — not when they receive a fact sheet.
- **High-profile debunking** may backfire by signaling threat, triggering defensive solidarity. The framework predicts this effect is real but not universal; context matters.
- **Interventions** should focus on reducing identity threat, providing safe off-ramps, and decoupling core moral values from factual claims. These are testable hypotheses.

---

9. Open Questions

- **Can a movement be partially a fantasy attractor?** Yes — gradient of κ. The diagnosis is not binary.
- **What interventions increase κ?** Reducing identity threat, safe off-ramps, and decoupling moral values from factual claims are candidate mechanisms.

- **How does collective basin depth scale with group size?** Social coupling likely amplifies depth nonlinearly. Untested.
- **Are all political fantasy attractors harmful?** The paper makes no claim. The mechanism may sometimes provide resilience against genuine disinformation.
- **How can we empirically detect the Level 2 / Level 3 transition?** This is the open frontier implied by §2. Methods could include subjective doubt scales, reaction time measures, or physiological markers. The paper does not solve this; it identifies the problem.

---

10. Conclusion

The conscious suppression of intelligent correction is a real political phenomenon, but it is narrower than often assumed. Political fantasy attractors are not failures of intelligence; they are successes of identity-constitutive commitment that operates *after* the error signal is detected. Cases where correction fails due to automatic bias are not yet fantasy attractors by this definition.

The diagnostic criteria identify candidates, not confirmed cases. Distinguishing Level 2 from Level 3 remains an empirical challenge. This honesty does not weaken the framework; it clarifies what further work is needed.

Fact-checking alone fails against a fantasy attractor. Interventions must address the conscious commitment that creates the basin depth. The dance of politics is not only about truth. It is about who you are, who you trust, and what you will not abandon. Intelligence navigates; conscious commitment anchors the basin.

Suggested citation: Galida, R. S. (2026). The Conscious Suppression of Correction: Fantasy Attractors in Political Movements. *Fantasy Attractor.*