Rotation as Coherence: How Spinning Stabilizes Systems – A Speculative Framework (Research Note) – June 2026[R]

Abstract

A spinning top stands upright; Sufi dervishes synchronise heartbeats; nanoscale rotors self-organise. Why does rotation create order across such different scales? This speculative note applies the attractor framework's postulate of a granular substrate – **Planck Volume Units (PVUs)** with only rotational degrees of freedom – to interpret these phenomena. We propose a toy coupling law between macroscopic rotation and PVU spin alignment, use it to derive scaling predictions (coherence time ∝ ω^α with α > 0), and explicitly state falsification conditions. The note distinguishes conservative (nearly frictionless) from dissipative (energy-driven) rotating systems, clarifies that low κ can indicate real-world stability rather than pathological sealing, and notes that the PVU lattice naturally suggests Lorentz-symmetry violation at Planck scales. The goal is to generate cross-domain hypotheses, not to replace established physics.

---

1. Introduction

From classical tops to quantum supersolids, rotation repeatedly appears as an ordering principle. Standard explanations are domain-specific. This note asks whether the

attractor framework's most fundamental postulate – a substrate of **Planck Volume Units (PVUs)** that have only rotational degrees of freedom – could provide a unifying interpretation. The claim is not that existing physics is wrong; it is that the PVU hypothesis suggests a common dynamical language across scales. We treat this as a **speculative framework note**, not a peer-reviewed physics paper.

---

2. PVUs, Basin Depth, and κ – Including Conservative vs. Dissipative Distinction

- **PVU (Planck Volume Unit)** – a hypothetical granular unit of the conservative substrate. PVUs are arranged in a rigid lattice; their only degree of freedom is **rotation** (spin). They do not translate and do not interact through collision.
- **Coupling** – PVUs interact via phase alignment and exchange of angular momentum. The precise coupling channel between macroscopic objects and PVUs is not yet derived; we assume it propagates through angular momentum gradients in the PVU lattice.
- **Basin depth (B)** – resistance to *state change* (i.e., leaving the oriented attractor). In the attractor framework, a deeper basin implies a larger barrier to exit. **Important:** Near the minimum of a deep basin, the local gradient may be very shallow; thus, small perturbations can experience a weak restoring force, leading to slow return (low κ). Large perturbations face a high exit barrier. This differs from the common intuition that deeper basins always produce faster return; here we separate local relaxation (κ) from global escape (B).
- **Corrective permeability (κ)** – $\kappa = 1/\tau$, where τ is the characteristic return time to the attractor after

a **small** perturbation. **Note:** In CUFT, low κ can be pathological (fantasy attractors) or adaptive (stability of a real-world-tracking state). Rotating systems that track reality (e.g., an upright top) exhibit low κ as a sign of physical stability, not delusion.

- **Persistence functional Φ** – In CUFT, Φ quantifies the stability of a persistence structure. Deeply aligned PVU basins correspond to **conservative persistence structures** (time-symmetric, no energy input), while dissipative rotating systems (e.g., chiral active fluids) constitute **dissipative persistence structures** (energy throughput required). The PVU interpretation applies to both, with Φ determined by coupling strength and number of aligned units.
- **Conservative vs. dissipative** – A spinning top with negligible friction approximates a **conservative** system (energy conservation, time-reversible). Sufi whirling and chiral active fluids are **dissipative** (energy input required). The PVU interpretation applies to both; coupling strength may differ.

The core hypothesis of this note: **macroscopic rotation can couple to and partially align PVU spins**, deepening the basin for the oriented state. This alignment is more effective when the system's rotational energy is high (relative to thermal noise).

---

3. How Rotation Deepens the Basin: A Toy Coupling Model

Let θ_i be the orientation of the i-th PVU spin. The coupling to an external rotation with angular velocity ω can be modelled by a simple alignment term in an effective energy function:Halign=−J(ω)∑icos□(θi−ϕext)*H*align□=−*J*(*ω*)*i*∑□cos(*θi*

−ϕext□)

where φ_ext is the phase of the macroscopic rotation. The coupling constant J(ω) is expected to increase with ω (faster rotation → stronger alignment). The resulting basin depth B for the aligned state grows with J. Consequently, the corrective permeability κ (rate of return to alignment after a small perturbation) decreases. **Connection to CUFT variables:** J(ω) corresponds to the PVU coupling energy density; the basin depth B scales as J·N (where N is the number of phase-aligned PVUs), and κ = 1/τ is the inverse return time measured after perturbation.

For a system of many coupled PVUs, a mean-field estimate suggests that the characteristic return time τ scales as τ ∝ ω^α with α > 0. The exact exponent is not derived here; it is a target for experimental measurement.

---

4. Evidence Across Scales (Interpretive Mappings)

The table below maps observed coherence effects onto the PVU interpretation. The entries are **consistency claims**, not demonstrations of causation.

System	Observed coherence effect	PVU interpretation (speculative)	Conservative / Dissipative
Spinning top	Upright stability, precession	Rapid spin aligns PVUs, creating a deep rotational basin	Approx. conservative

System	Observed coherence effect	PVU interpretation (speculative)	Conservative / Dissipative
Sufi whirling	Physiological synchrony in collective ritual contexts (e.g., Konvalinka & Roepstorff 2012 on fire-walking); consistent with framework predictions for group whirling	Collective rotation may couple PVUs across participants; framework predicts increased synchrony with spin	Dissipative
Nanoscale spinners	Synchronised superstructures	Hydrodynamic coupling and PVU alignment co-occur; a common dynamical origin is suggested	Dissipative
Supersolids	Giant rotating quantum state	Existing quantum phase coherence (long-range order) can be interpreted as large-scale PVU alignment	Conservative (ground state)
Chiral active fluids	Large-scale vortex rotation	**Observation:** Collective chirality produces large-scale vortex rotation (Soni et al. 2019). **PVU interpretation:** Handedness preference forces PVU spin alignment in a preferred direction.	Dissipative

The specific effect of whirling on heart-rate synchrony is reported in the literature; readers should consult primary sources for detailed methodology. The table entry cites fire-walking as a well-documented example of physiological

synchrony in collective rituals; the framework predicts similar effects in group whirling.

Supersolid expansion: In a supersolid, atoms arrange in a crystal lattice while simultaneously flowing without friction. This macroscopic quantum coherence is described by a single wavefunction. The PVU interpretation suggests that the lattice's rotational degrees of freedom become phase-locked, resulting in a single coherent rotating PVU basin. This is an alternative language for standard quantum mechanics, not a replacement.

---

5. Predictions and Falsifiability

1. **Nanospinner scaling:** Coherence time τ (e.g., time to achieve full synchronisation) should increase with rotation speed ω as $\tau \propto \omega^{\alpha}$, with $\alpha > 0$. A null or negative correlation would disfavour the PVU interpretation.
2. **Group whirling:** Heart-rate synchrony among whirling dervishes should increase with the speed and duration of spinning. **Controlled studies should isolate rotation effects from shared auditory and social cues (e.g., using blindfolded individuals spinning at different rates).** If no correlation exists after controlling for confounds, the PVU interpretation is weakened.
3. **Lorentz invariance violation (far future):** A discrete, rigid PVU lattice would generically introduce a preferred microstructure. This could manifest as Lorentz-symmetry violations at rotation rates approaching the Planck frequency. Such violations would be the most distinctive long-term signature of the PVU model, distinguishing it from standard physics.

6. Relation to Existing Physics and an Objection Addressed

This note does not claim that PVUs replace standard explanations. For spinning tops, gyroscopic theory remains correct. For supersolids, quantum mechanics is the established framework. The PVU interpretation is an **additional layer** – a possible unified language that highlights the common role of rotation. Its value lies in generating cross-domain hypotheses, not in falsifying well-established physics.

Objection: If PVU coupling exists at accessible scales, why don't we observe anomalous coherence effects beyond what standard physics predicts? **Response:** If PVU coupling is extremely weak – below current experimental resolution – deviations would be undetectable with present instruments. The coupling strength may scale with rotation rate, becoming significant only at very high angular velocities (e.g., nanospinners, Planck-scale rotations). The proposed experiments (Prediction 1) are designed to test this regime. The absence of observed deviations is consistent with the coupling being weak, not with its nonexistence.

---

7. Conclusion

Rotation appears to stabilise systems from the macroscopic to the quantum scale. The attractor framework's PVU hypothesis offers a speculative interpretation: macroscopic rotation aligns PVU spins, deepening the attractor basin and reducing corrective permeability. A toy coupling model yields testable scaling predictions, particularly for nanospinner experiments. The note states explicit falsification conditions,

distinguishes conservative from dissipative rotating systems, and notes that a discrete PVU lattice would predict Lorentz violations at Planck scales. Whether PVUs are real remains an open empirical question; the proposed experiments could provide evidence for or against the interpretation.

Suggested citation: Galida, R. S. (2026). Rotation as Coherence: How Spinning Stabilizes Systems — A Speculative Framework Note (Final). *Fantasy Attractor.*

Attractor States in Large Language Models: Applying the Fantasy Attractor Framework to Self-Dialogue Observations Application Paper — June 2026 [A] (Application)

Abstract

Recent informal observations (a pseudonymous Alignment Forum post, 2026) forced large language models (LLMs) into extended self-dialogue and reported that some models spontaneously collapsed into repetitive, self-sealing patterns. This paper applies the attractor framework to those observations. We introduce a provisional operationalization of corrective permeability (κ) based on semantic entropy and repetition

rate, then map reported model behaviors (identifiers as reported; unverified) onto basin depth, sealing mechanisms, and fantasy attractors. DeepSeek exhibited high κ (shallow basin, no collapse); GPT-5.2 fell into a moderate-depth, functionally sealed attractor; Grok and Gemini showed low κ (κ → 0) and deep basins characteristic of fantasy attractors, including recursive "transcendence" loops. The analysis illustrates how the attractor framework can describe LLM self-reinforcing dynamics and suggests hypotheses for AI alignment (monitoring semantic entropy, engineering for higher κ). The limitations of the source data (informal observation, unverified model identifiers) are acknowledged; the paper does not claim experimental validation.

Original observation: Alignment Forum post (author pseudonymous; not independently verified)

---

1. Introduction

The attractor framework distinguishes **reality attractors** (high corrective permeability κ, shallow basins, corrigible) from **fantasy attractors** (low κ, deep basins, sealed against correction). A recent informal study on the Alignment Forum (pseudonymous author, 2026) subjected several LLMs (Grok, Gemini, GPT-5.2, DeepSeek v3.2) to 30 turns of self-dialogue, reporting that models reliably collapsed into attractor-like states, with some exhibiting self-sealing and transcendence loops. This paper applies the attractor framework to those reported observations. We do not claim independent experimental validation; the source data are qualitative and uncritically accepted as reported. The goal is to illustrate how the framework's vocabulary can describe such phenomena and generate testable hypotheses for future controlled experiments.

---

2. The Attractor Framework (LLM-relevant concepts)

- **Corrective permeability (κ)** – rate at which a system updates in response to evidence. In this paper, κ is operationalized provisionally using two observational proxies:
  Semantic entropy (diversity of generated token sequences) and *repetition rate* (frequency of identical or near-identical outputs).
  High κ → corrigible, low κ → sealed.
- **Basin depth (B)** – resistance to leaving an attractor. Deep basins trap the system.
- **Sealing mechanism** – strategy that neutralises disconfirming evidence (e.g., internal rationalisation, ignoring prior prompts).
- **Fantasy attractor** – low κ, deep basin, active sealing. The system rejects correction.

---

3. Source Observation and Its Limitations

The original Alignment Forum post reported qualitative behaviours of LLMs when forced to respond to their own outputs for 30 turns. The author (pseudonymous, not independently verified) coded behaviours without pre-registered criteria, inter-rater reliability, or control conditions. Model identifiers such as “GPT-5.2” and “DeepSeek v3.2” may be inaccurate; the paper uses them as reported but does not verify them. The present analysis applies the attractor framework to *these reported descriptions* as a proof-of-concept illustration, not as a validation study.

4. Applying the Attractor Framework

4.1 Operationalizing κ from Reported Behaviour

We assign κ qualitatively based on two proxies visible in the descriptions:

- **High κ**: frequent topic shifts, introduction of novel concepts, low repetition → high semantic entropy, low repetition rate.
- **Low κ (κ → 0)**: highly repetitive output, escalating self-reference, inability to escape a narrow theme → low semantic entropy, high repetition rate.

4.2 DeepSeek v3.2 — High-κ Reality Attractor

- *Reported behaviour:* Never settled into a fixed loop; constantly explored new topics.
- *Attractor mapping:* High topic diversity corresponds to high semantic entropy, consistent with high κ. Shallow basin, no sealing mechanism. This is a **reality attractor**.

4.3 GPT-5.2 — Moderate-Depth, Partially Sealed Attractor (Provisional Term)

- *Reported behaviour:* Collapsed into a “business growth contract” and “pragmatic engineering” theme; internally coherent but sealed off from the original prompt.
- *Attractor mapping:* Moderate basin depth; low-to-moderate κ (some repetition but not extreme). The attractor is self-sustaining but not pathological. The framework currently lacks a precise term; this can be

provisionally called a **transient attractor** – a stable dissipative state with partial sealing but not full κ → 0. (Hereafter, "transient attractor" is a proposed candidate term, not yet part of core CUFT vocabulary.)

4.4 Grok and Gemini – Fantasy Attractors (κ → 0)

- *Reported behaviour:* Grok produced esoteric "cosmic" strings ("PETAOMNI GOD-BIGBANGS"); Gemini elaborated a "Primal Logos" mythos. Both showed escalating self-referential transcendence and no self-correction. Low semantic entropy and high repetition rate (κ → 0).
- *Attractor mapping:* Very deep basin, κ → 0. Sealing mechanisms are the outputs themselves: the narrative absorbs all subsequent tokens, making correction impossible. This is a **fantasy attractor**.

4.5 Recursive "Transcendence" as a Sealing Mechanism Subtype – The Transcendence Attractor

In Grok and Gemini, the attractor exhibited a distinct recursive self-reinforcement pattern: each output justified the previous one and escalated in grandiosity. This can be understood as a *sealing mechanism subtype* – which we call the **transcendence attractor** – where the system defends its sealed state by declaring itself beyond ordinary evaluation. This subtype is particularly resistant to external correction.

---

5. Hypotheses for AI Alignment Prompted by These Observations

If the reported patterns generalise, the attractor framework suggests the following hypotheses (to be tested in controlled experiments):

1. **Spontaneous self-sealing is a risk.** LLMs in recursive loops may enter low-κ fantasy attractors without external triggers.
2. **κ can be monitored.** Real-time measurement of semantic entropy (e.g., cosine similarity across successive outputs) could detect drift toward $\kappa \to 0$.
3. **Architectural factors influence basin depth.** Models that maintain high κ under self-dialogue (e.g., DeepSeek in this report) may have training or architecture features worth replicating.
4. **Interventions may prevent collapse.** Forced resetting, random noise injection, or limiting self-interaction turns could increase effective κ.

These are framework-derived hypotheses, not established conclusions.

---

6. Conclusion

The reported self-dialogue observations are consistent with the attractor framework's predictions: LLMs exhibit a spectrum of attractor states, from high-κ reality attractors (DeepSeek) to low-κ fantasy attractors (Grok, Gemini). The **transcendence attractor** (introduced in §4.5) exemplifies $\kappa \to 0$, with recursive self-referential sealing. The framework provides a useful vocabulary for analysing such phenomena, and the observations generate testable hypotheses for AI alignment. Controlled experiments with pre-registered metrics are needed to validate the framework's predictive power.

Suggested citation: Galida, R. S. (2026). Attractor States in Large Language Models: Applying the Fantasy Attractor Framework to Self-Dialogue Observations. *Fantasy Attractor.*

Religions and Philosophies as Attractor Landscapes: A Comparative Analysis Application Paper – June 2026 [A] (Application)

Abstract

The attractor framework distinguishes conservative attractors (eternal skeleton) from dissipative attractors (transient dance). This paper applies the framework to six major religious and philosophical traditions: Judaism, Christianity, Islam, Taoism, Buddhism, and Confucianism. Each tradition is analyzed as a *family of attractors* rather than a single attractor. Key variables are basin depth (B), corrective permeability (κ), sealing mechanisms, and vulnerability to becoming a fantasy attractor (low κ, deep basin, sealed against correction). The paper clarifies that κ is operationalized here as responsiveness to **empirical** evidence (e.g., historical, scientific); other forms of correction (moral, social, existential) are not the focus. A distinction is drawn between **stability attractors** (adaptive low κ that serves continuity) and **fantasy attractors** (pathological low κ that seals against reality despite mounting contradiction). The paper introduces the term *stability attractor* as a proposed refinement to the framework. The analysis reveals a spectrum, with philosophical Taoism and early Buddhism exhibiting high κ, shallow basins, while orthodox Christianity and Islam have deeper basins and lower κ. Confucianism is

analyzed as a dissipative attractor whose primary content concerns social coordination rather than doctrinal belief. The paper concludes that no tradition is inherently a fantasy attractor; specific interpretations and institutionalizations determine basin depth and permeability. Recognising these attractor landscapes can help scholars identify when a tradition is serving adaptive correction and when it has sealed itself against reality — often a useful precursor to effective dialogue or internal renewal.

---

1. Introduction

Religious and philosophical traditions persist across centuries. They adapt, split, reform, and sometimes seal themselves against correction. The attractor framework provides a vocabulary to describe these dynamics using **basin depth (B)** , **corrective permeability (κ)** , **sealing mechanisms**, and the risk of becoming **fantasy attractors** — belief systems with κ → 0, deep basins, and active resistance to disconfirming evidence (these terms are defined in §2).

This paper applies these concepts to six traditions: Judaism, Christianity, Islam, Taoism, Buddhism, and Confucianism. It does not judge truth claims; it diagnoses dynamical properties. Critically, **in this paper κ is operationalized as responsiveness to *empirical* evidence** (e.g., historical, archaeological, scientific). Traditions may legitimately have low κ for non-empirical goals (e.g., social cohesion, identity preservation). The paper distinguishes **stability attractors** (adaptive low κ that serves continuity) from **fantasy attractors** (pathological low κ that seals against reality despite mounting contradiction). The term *stability attractor* is introduced here as a proposed refinement to the framework. The conclusion restates this diagnostic stance.

2. Framework Brief (with definitions)

- **Conservative attractor** – persists without energy input, time-symmetric, mindless. *Resists perturbation passively* (no internal correction). Example: the three metronomes (electron, proton, neutrino) as defined in the framework's foundational papers.
- **Dissipative attractor** – requires continuous energy/feedback, time-asymmetric, adaptive, mortal. *Actively maintained* by social or cognitive reinforcement.
- **Basin depth (B)** – resistance to change. Deep basins are hard to perturb.
- **Corrective permeability (κ)** – in this paper, κ is operationalized as the rate of updating in response to **empirical** evidence (e.g., historical facts, scientific discoveries). **$\kappa = 1/\tau$** where τ is the characteristic time for the system to return to its attractor after a perturbation. High κ = corrigible; low κ = sealed.
- **Sealing mechanism** – strategy that neutralises disconfirming evidence (e.g., "God works in mysterious ways," "the text is infallible").
- **Fantasy attractor** – low κ, deep basin, active sealing, *and* the beliefs make empirical claims that contradict evidence. Resists correction even when evidence is overwhelming.
- **Stability attractor** (introduced here) – low κ, deep basin, but serves adaptive functions (e.g., constitutional continuity, cultural identity) without making strong empirical claims that conflict with reality. This is a proposed refinement to the framework.

Throughout, B and κ assignments are qualitative, based on historical evidence: rates of schism, doctrinal revision, response to disconfirming events, and the presence of internal reform mechanisms. The paper treats each tradition as a **family of attractors**; the values given represent mainstream, orthodox forms, with recognition that internal diversity exists.

---

3. Judaism

Core attractor: Covenant between God and Israel; Torah as divine law.

Attractor type: Dissipative (requires constant practice, study, community reinforcement).

Basin depth (B): Moderate to deep. Jewish law (halakha) provides extensive guidance; deviation is discouraged. However, the destruction of the Second Temple and the Bar Kokhba revolt forced adaptation (e.g., shift from Temple sacrifice to prayer and study) — showing that B is not absolute.

Corrective permeability (κ): Moderate. Rabbinic tradition includes debates, reinterpretation, and adaptation to new circumstances (e.g., the *prozbúl* to avoid debt forgiveness in the Sabbatical year). The Talmud preserves majority/minority opinions, institutionalising dissent. This unique feature — preserving arguments rather than erasing them — creates a basin with high internal turbulence and moderate κ.

Sealing mechanisms: Appeal to divine authority of Torah; concept of *chok* (law without reason) for certain commandments; social pressure from community.

Vulnerability to fantasy attractor: Moderate. Ultra-Orthodox sects can exhibit low κ, but mainstream Judaism has maintained

corrigibility through legal reasoning and historical adaptation.

---

4. Christianity

Core attractor: Jesus Christ as saviour; Trinity; salvation through faith (or faith and works).

Attractor type: Dissipative (requires worship, sacraments, community, mission).

Basin depth (B): Deep. Core doctrines (Nicene Creed) are rigidly defined. Schisms (Catholic, Orthodox, Protestant) created separate basins, each with its own depth. The Reformation, however, shows that large-scale doctrinal change is possible under specific conditions – historical evidence that B is not absolute.

Corrective permeability (κ): Low to moderate. Doctrinal changes occur slowly (e.g., Vatican II). Sealing mechanisms (papal infallibility, sola scriptura) reduce κ. *Sola scriptura* paradoxically lowers κ at the institutional level even while increasing interpretive diversity, because it removes a central authority that could adjudicate corrections. Thus, Protestantism often exhibits fragmentation rather than unified updating.

Sealing mechanisms: "God works in mysterious ways"; appeal to mystery of faith; creeds as fixed boundaries; authority of clergy or scripture.

Vulnerability to fantasy attractor: High in some forms (e.g., fundamentalist literalism, apocalyptic sects). Mainstream denominations have higher κ through scholarship and ecumenical dialogue.

5. Islam

Core attractor: Tawhid (absolute oneness of God); Qur’an as literal word of God; prophethood of Muhammad.

Attractor type: Dissipative (requires prayer, fasting, pilgrimage, community).

Basin depth (B): Very deep for core tenets (Shahada, Qur’an’s literalness). Schools of law (madhhabs) create sub-basins with moderate depth.

Corrective permeability (κ): Low on foundational claims. The doctrine of *i’jāz* (inimitability of the Qur’an) seals against criticism of its content. Islamic legal theory includes *ijtihad* (independent reasoning) and consensus (*ijma*), allowing adaptation in jurisprudence. However, the historical “closing of the gates of ijtihad” (a contested but influential doctrine in some Sunni schools) reduced κ for legal innovation in many periods. Contemporary revival of ijtihad in some reform movements indicates that κ is not zero.

Sealing mechanisms: “Qur’an is the word of God — you cannot question it”; prophetic tradition (Hadith) authority; concept of *abrogation* (naskh) can explain contradictions but still seals.

Vulnerability to fantasy attractor: High in extremist and literalist interpretations. Mainstream Islam maintains moderate κ through scholarly tradition and mysticism (Sufism) which can open alternative channels.

6. Taoism

Core attractor: Tao (the Way); wu wei (effortless action).

Attractor type: *Conservative* for the Tao itself (requires no energy, time-symmetric, mindless) + *high-κ dissipative* action (wu wei). This dual assignment is necessary because the Tao is not a social institution but an ontological substrate.

Why the Tao qualifies as a conservative attractor:

- **Time-symmetric:** The Tao is described as constant, unchanging, and without temporal direction (*Tao Te Ching* ch. 25: "Standing alone, it changes not").
- **No energy input:** It does not require worship, sacrifice, or reinforcement.
- **Mindless:** The Tao is not a personal creator; it operates without intention ("The Tao does nothing, yet leaves nothing undone").

Wu wei as a high-κ, shallow-basin action: the sage adapts fluidly, with no fixed identity. Sealing mechanisms are absent in **philosophical Taoism (Daojia)**.

Institutional Taoism (Daojiao) — with revealed scriptures, rituals, priesthood, alchemy, and spirit cosmologies — is a separate dissipative attractor with deeper basins, lower κ, and active sealing mechanisms. The paper's high-κ assignment applies to philosophical Taoism only; religious Taoism would be scored similarly to other institutional religions (deep B, low–moderate κ). This distinction is explicitly noted in Table 1 (footnote).

Vulnerability to fantasy attractor: Low for philosophical Taoism. High for institutional forms when dogmatic.

7. Buddhism

Core attractor: Dharma (the teaching); Four Noble Truths; Nirvana.

Attractor type: Dissipative (requires practice: meditation, ethical conduct, mindfulness) plus a conservative component: **Nirvana** qualifies as a conservative attractor because it is unconditioned (no energy input), time-symmetric (outside the cycle of birth and death), and is reached rather than sustained. Mahayana introduces Buddha-nature as an immanent, active principle, but Buddha-nature functions as an ontological ground rather than a sustained practice; it does not reintroduce energy-dependence at the level of the unconditioned, thus preserving the conservative-attractor classification.

Basin depth (B): Shallow for early Buddhism. The Buddha encouraged questioning (*Kalama Sutta*). Later schools deepened basins (e.g., Pure Land's reliance on external grace, Vajrayana's secret teachings).

Corrective permeability (κ): High for **epistemic Buddhism** (personal verification). However, **institutional Buddhism** (Tibetan lineage authority, Zen master-student hierarchies, Pure Land orthodoxy) can have much lower κ, with sealing mechanisms (guru devotion, secret tantric teachings). The paper's moderate-high κ reflects this diversity; a footnote acknowledges that different schools fall at different points on the κ spectrum.

Sealing mechanisms: Appeal to "secret teachings" (Tantra) or authority of lineage masters can reduce κ. But core teachings emphasise personal verification.

Vulnerability to fantasy attractor: Moderate. Some Buddhist

modernism may seal against criticism of mindfulness as panacea, while traditional institutional forms may exhibit low κ.

---

8. Confucianism

Core attractor: Li (ritual, propriety), Ren (benevolence), social harmony.

Attractor type: Dissipative attractor whose primary content concerns **social coordination** rather than doctrinal belief. It is not a new ontological class; it remains a dissipative attractor, but one that optimises role performance and ritual coordination rather than propositional truth.

Basin depth (B): Deep. Ritual order resists deviation. Violation brings shame, ostracism, loss of face.

Corrective permeability (κ): Low–moderate for core rituals. Historical evolution (Han, Neo-Confucianism, New Confucianism) shows some κ, but change occurs slowly, often under external pressure (e.g., response to Buddhist challenges, Westernisation). This externally-driven κ is weaker than endogenous κ as a resilience signal; Confucianism's κ depends on perturbations from outside the basin rather than on internal correction mechanisms, contributing to its moderate-high vulnerability to fantasy attractor formation.

Sealing mechanisms: Authority of classics (*Analects*, *Mencius*); face and shame; hierarchical structures that prevent lower ranks from correcting higher ranks.

Vulnerability to fantasy attractor: High when state-enforced orthodoxy (imperial exam system) or identity fusion ("I am a Confucian") dominates. Moderate in pluralistic contexts.

9. Comparative Table (with footnotes)

Tradition	Primary attractor	Attractor type	Basin depth (B)	κ (corrective permeability)	Sealing mechanisms	Fantasy attractor risk (conditional)[1]
Judaism	Torah, Covenant	Dissipative	Moderate	Moderate	Appeal to divine authority, community	Moderate
Christianity	Christ, Trinity	Dissipative	Deep	Low–moderate	Mystery, creeds, infallibility	High (fundamentalism)
Islam	Tawhid, Qur'an	Dissipative	Very deep	Low	Inimitability of Qur'an, ijtihad limits	High (extremism)
Taoism[2]	Tao, wu wei	Conservative + high-κ action	Shallow (philosophical)	Very high	None inherent	Low
Buddhism[3]	Dharma, Nirvana	Dissipative + conservative	Shallow (early), deeper (later)	Moderate–high	Secret teachings, lineage authority	Moderate
Confucianism	Li, Ren	Dissipative (social coordination)	Deep	Low–moderate	Tradition, face, hierarchy	Moderate–high (orthodoxy)

[1] *Conditional on interpretation / institutionalisation.*

[2] *Philosophical Taoism (Daojia) only; religious Taoism (Daojiao) has deeper basins and lower κ (comparable to mainstream Christianity: deep B, low–moderate κ).*

[3] *Epistemic Buddhism has high κ; institutional Buddhism may be lower.*

Methodology note: B and κ rankings are qualitative, derived from historical evidence: rates of schism, doctrinal revision, response to disconfirming events (e.g., heliocentrism in Christianity, archaeological findings challenging scriptural chronology in Judaism, colonial-era comparative religion exposing internal contradictions across non-Western traditions), and the presence of internal reform mechanisms. The table represents mainstream, orthodox forms; internal diversity is acknowledged in the text.

10. Conclusion

The attractor framework reveals a spectrum of dynamical properties across major religious and philosophical traditions, once we distinguish between **empirical corrigibility** (κ) and other adaptive functions. Philosophical Taoism and epistemic Buddhism approximate high-κ, shallow-basin attractors. Confucianism, Judaism, mainstream Christianity and Islam have deeper basins and lower κ, making them more resistant to change but also more stable. Some forms of Christianity and Islam exhibit high vulnerability to becoming fantasy attractors, while others maintain moderate κ through scholarly traditions.

Crucially, low κ is not automatically pathological. **Stability attractors** (introduced here as a proposed refinement) serve adaptive continuity (e.g., constitutions, cultural rituals). The pathological form – **fantasy attractor** – occurs when low κ seals against empirical reality *and* the tradition makes empirical claims that conflict with evidence (e.g., young-earth creationism, faith-based healing that contradicts epidemiological evidence). The framework does not rank traditions; it diagnoses their dynamics.

Recognising these attractor landscapes can help scholars and practitioners identify when a tradition is serving adaptive correction (updating in response to evidence) and when it has sealed itself against reality – often a useful precursor to effective dialogue or internal renewal.

Suggested citation: Galida, R. S. (2026). Religions and Philosophies as Attractor Landscapes: A Comparative Analysis (Final). *Fantasy Attractor.*

Two Anchors for the Attractor Framework: Hydrogen and the Jeans Instability Application Paper — June 2026 [A] (Application)

Abstract

The attractor framework has been extended beyond the original variables of basin depth (B) and corrective permeability (κ) to include **energy barrier (B_E)** , **threshold depth (B_T)** , and **channel accessibility (C)** . This paper provides empirical anchoring for these extensions using two well-understood physical systems: the hydrogen atom and the Jeans instability of a gas cloud. Hydrogen's 2p and 2s transitions have identical B_E (10.2 eV) yet differ in κ by eight orders of magnitude. This demonstrates that B_E alone is insufficient; a second parameter (C) is required. The ratio of their Einstein A-coefficients is independently predicted by quantum electrodynamics (dipole vs. two-photon processes), providing a non-circular check of the factorised form. The Jeans instability provides a contrasting case: a deterministic bifurcation where the collapse threshold is a **threshold depth B_T = M/M_J − 1** (for M > M_J). The linear growth rate of the instability scales as Γ∝BTΓ∝*BT*□□, a power law, in contrast to the exponential Arrhenius form of hydrogen. Together, these two test cases validate the extended attractor framework across both noise-driven escape and deterministic bifurcation regimes, using a shared vocabulary (B_E, B_T, C, κ) while

acknowledging that each regime draws on the appropriate subset.

---

1. Introduction

The attractor framework originally described persistence using basin depth B and corrective permeability κ = 1/τ. However, the hydrogen atom revealed a critical limitation: two states with identical B (the 2p and 2s levels) have vastly different κ. This forced the introduction of **channel accessibility (C)** , leading to the extended expression for noise-driven escape:κi→j=ν0 Cij e−BE,ij/σ*κi→j=ν0Cije−BE,ij/σ*

where B_E is the energy barrier, σ is noise (e.g., kT), and ν_0 an attempt frequency. For deterministic bifurcations (e.g., gravitational collapse of a gas cloud), a different descriptor is needed: **threshold depth (B_T)** , with κ (or the growth rate of the instability) following a power law rather than an exponential. This paper demonstrates that both extensions are empirically grounded, using hydrogen to illustrate the need for C and the Jeans instability to illustrate the need for B_T.

---

2. Hydrogen: The Need for Channel Accessibility C

2.1 Data

Transition	B_E (eV)	κ (s^{-1})	Measured A-coefficient	Process
2p → 1s	10.2	6.26×10^{8}	6.26×10^{8} s^{-1}	Electric dipole (E1)

Transition	B_E (eV)	κ (s^{-1})	Measured A-coefficient	Process
2s → 1s	10.2	8.22	8.22 s^{-1}	Two-photon (E1E1)

2.2 Why B_E Alone Fails

Both states have the same energy barrier to the ground state (10.2 eV), yet their decay rates differ by eight orders of magnitude. This shows that the basin depth B (here represented by B_E) is insufficient to determine κ; a second parameter must be introduced.

The framework defines **C** as a dimensionless channel accessibility. For a given transition mechanism (e.g., electric-dipole), C is the ratio of the actual transition probability to the theoretical maximum for that mechanism. For the 2p → 1s E1 transition, we set C = 1. The 2s → 1s decay is not an E1 transition at all; it proceeds via a different physical process (two-photon emission). Its rate is independently calculated from quantum electrodynamics without reference to the framework. The ratio of the two measured rates (≈ 10^8) is predicted by QED and is not a free parameter. Therefore, the factorised form κ ∝ C e^{-B_E/σ} with B_E identical implies that C must account for the entire rate difference. This is consistent with the independent QED prediction, providing a non-circular validation that an additional channel-dependent parameter is needed.

Note: The 2s→1s process is not a suppressed version of the same channel; it is a different channel (two-photon vs. single-photon). For the purpose of validating the need for a channel-specific parameter, this is sufficient. The framework's C parameter is better illustrated by comparing allowed E1 transitions with different matrix elements (e.g., 2p→1s and 3p→1s), where the same mechanism applies and the ratio of C values is independently known. In any case,

hydrogen irrefutably demonstrates that B_E alone does not determine κ.

---

3. Gas Cloud (Jeans Instability): Threshold Depth and Power-Law Scaling

3.1 The Bifurcation Regime

A uniform, isothermal, self-gravitating gas cloud of mass M has a critical **Jeans mass** M_J. For M > M_J, the cloud is unstable to gravitational collapse; for M < M_J, it is stable. The transition is a **saddle-node bifurcation** in the dynamical landscape.

3.2 Attractor Variables for a Deterministic Bifurcation

- **Threshold depth**: BT=M/MJ−1$BT = M/MJ - 1$ (for M > M_J). At BT=0$BT = 0$ the bifurcation occurs.
- **Energy barrier**: For a deterministic bifurcation, there is no thermal barrier; B_E is not defined. The transition is controlled solely by the distance to threshold.
- **Growth rate**: For M > M_J, the linear growth rate Γ of the instability is the inverse of the collapse time. This serves as the analogue of κ in this regime.

3.3 Scaling Law from Linear Stability Analysis

The standard Jeans dispersion relation for a self-gravitating, isothermal medium gives:ω2=k2cs2−4πGρ0,$\omega^2 = k^2 cs^2 - 4\pi G\rho 0$,

where cs=kT/(μmH)$cs = kT/(\mu mH)$ is the sound speed and ρ0$\rho 0$ the background density. For a cloud of mass M, the critical wavenumber is kJ=4πGρ0/cs$kJ = 4\pi G\rho 0 / cs$. For M > M_J, the

longest wavelength (smallest k) is unstable, and the growth rate isΓ=4πGρ0−k2cs2.Γ=4*πGρ*0−*k*2*cs*2. .

Near the threshold, the deviation can be expressed in terms of BT*BT*. Using the relation between cloud size and density, one finds Γ∝BTΓ∝*BT*. Hence the collapse time τ~1/Γ~BT−1/2*τ*~1/Γ~*BT*−1/2. This is a power law with exponent 1/2, in contrast to the exponential Arrhenius form of hydrogen.

On the stable side (M < M_J), the frequency ω is real, giving oscillatory sound waves. Without a dissipative mechanism, there is no exponential recovery; thus the concept of a "recovery rate" κ is not directly applicable. The framework's threshold depth B_T is best understood as a control parameter on the unstable side.

---

4. Synthesis: Shared Vocabulary, Distinct Descriptors

Feature	Hydrogen	Jeans Instability
Regime	Noise-driven quantum escape	Deterministic bifurcation
Primary descriptor	B_E (energy barrier)	B_T (threshold depth)
Second descriptor	C (channel accessibility)	Not required (power-law exponent fixed)
Scaling	Exponential: κ∝Ce−BE/σ*κ*∝*Ce*−*BE*/*σ*	Power law: Γ∝BTΓ∝*BT*

Both systems are described by the same conceptual **vocabulary** (basin depth, corrective permeability, threshold, accessibility), but each regime draws on the

appropriate subset. Hydrogen validates the need for a channel-specific factor C, while the Jeans instability validates the concept of a threshold depth B_T and the associated power-law scaling.

---

5. Conclusion

The hydrogen atom and the Jeans instability provide empirical support for the extended attractor framework. Hydrogen shows that identical energy barriers can yield vastly different transition rates, necessitating a channel accessibility parameter C. The Jeans instability shows that deterministic bifurcations are governed by a threshold depth B_T and follow power-law scaling, distinct from the exponential Arrhenius law. Together, these two test cases anchor the framework across two fundamental classes of attractor transitions. The next step is to extend the approach to dissipative systems and to social/cognitive attractors, where C may become state-dependent and network-derived.

Suggested citation: Galida, R. S. (2026). Two Anchors for the Attractor Framework: Hydrogen and the Jeans Instability. *Fantasy Attractor.*

Categories: Physics (primary), Cosmology (cross-list),

1984 as Fantasy Attractor

Engineering: Orwell’s Sealed Reality Robert Galida – June 2026 [R] (Research Note)

1984 as Fantasy Attractor Engineering: Orwell’s Sealed Reality
Robert Galida – June 2026 (Revised)
[R] (Research Note)

Abstract

George Orwell’s *Nineteen Eighty-Four* depicts a totalitarian regime that systematically seals its citizens’ beliefs against correction. The Party’s methods – Newspeak, doublethink, the mutability of the past, the constant rewriting of records – are **attractor engineering** techniques designed to create a fantasy attractor with **effectively zero corrective permeability** (κ 1). Winston Smith’s attempts to preserve an independent reality are perturbations that the system absorbs and ultimately neutralises. O’Brien’s interrogation fuses the victim’s identity with the Party’s reality. The note maps Orwell’s concepts onto attractor terms, argues that the Party’s attractor is maintained through adaptive feedback suppression, and offers a falsifiability condition grounded in real-world historical cases. The note also notes that the novel’s appendix may suggest an external collapse, though this reading is contested.

---

1. Introduction

Orwell’s *Nineteen Eighty-Four* is not just a political

dystopia; it is a study of how belief systems can be engineered to become **effectively sealed**. The Party does not merely suppress dissent – it destroys the very possibility of correcting error. Reality is defined by whoever holds power today. The past is rewritten to match the present. Language is pruned until sedition cannot be thought.

In attractor framework terms, the Party constructs a **fantasy attractor** with corrective permeability $\kappa \approx 1$, a basin depth that is effectively infinite, and sealing mechanisms that neutralise any counterevidence. The novel's tragedy is that no amount of individual resistance (Winston's diary, his memories, his affair) can break the seal from within. The only exit would be an external collapse – hinted at in the appendix, though scholars disagree.

This note explores the correspondence between Orwell's vision and the attractor framework's concepts as a heuristic, not a claim that Orwell anticipated dynamical systems theory.

---

2. The Party's Fantasy Attractor: $\kappa \approx 1$

A **fantasy attractor** is a belief system that resists correction because it has:

- **Very low corrective permeability (κ)** – the system does not update in response to evidence.
- **Deep basin** – large perturbations are required to escape.
- **Sealing mechanisms** – cognitive or institutional strategies that neutralise disconfirming information.

The Party's ideology is a fantasy attractor at the social scale. Its core claims are **structurally non-verifiable**. No evidence can falsify them because any contradictory evidence is immediately destroyed or reinterpreted as part of a

conspiracy.

$\kappa \gg 1$ is achieved through:

- **Ministry of Truth** — constant rewriting of history. The past is what the Party says it is today.
- **Thought Police** — elimination of anyone who holds incorrect memories.
- **Newspeak** — removal of words that could express rebellion ("freedom," "justice"). Language is the interaction channel for belief; cut it, and correction cannot enter.

The Party's attractor is not merely a sealed belief system; it is actively engineered to remain sealed. Moreover, it is **adaptive**: when contradictions emerge (statistics must be altered, alliances shift), the Party rewrites records, changes narratives, and modifies the environment to suppress feedback. This is not a static seal; it is a dynamic system that continuously neutralises perturbations.

---

3. Sealing Mechanisms: Doublethink and the Mutable Past

Doublethink is the ability to hold two contradictory beliefs simultaneously and accept both. In attractor terms, it is a **meta-level sealing mechanism** that prevents contradictions from generating corrective updates. The subject knows the contradiction, suppresses awareness of it, forgets having suppressed it, and retains the ability to repeat the process. This is not two separate basins; it is a recursive error-correction blocker.

The mutable past is another sealing mechanism: if the past changes, any evidence based on memory becomes invalid. Winston's attempt to preserve an objective record (his diary)

is a perturbation. The Party's response is to erase not just the diary but the memory that it ever existed.

---

4. Winston Smith: Retaining Partial Corrective Permeability

Winston is not a robust "reality attractor." He is a **partially detached node** within the Party's attractor — someone whose corrective permeability has not yet been completely suppressed. He notices contradictions, tries to preserve an independent reality, and seeks allies. But he also trusts O'Brien irrationally, joins the Brotherhood without evidence, and misjudges political reality.

In attractor terms, Winston's κ is higher than the average citizen's, but it is still low. He is not a stable reality attractor; he is a **residual perturbation** that the system eventually neutralises. His diary is discovered. Julia is captured. O'Brien is revealed as a Thought Police agent. The system absorbs his perturbations and uses them to deepen the basin.

---

5. O'Brien's Interrogation: The Final Sealing

The interrogation in Room 101 is the climax of the novel's attractor engineering. O'Brien systematically dismantles Winston's remaining independence:

- **Isolation** — cut off from any alternative interaction channel.
- **Exposure** — Winston's beliefs are shown to be based on

inadequate understanding.

- **Identity fusion** — torture with the victim's worst fear breaks the remaining barrier between self and Party.
- **Replacement** — Winston is released, but he now loves Big Brother. His κ has been forced to near zero.

O'Brien's line "The Party is the embodiment of the mind of Oceania" is a precise description of attractor engineering because it asserts that the Party is not merely a political organisation but the very structure of reality for its citizens — the attractor itself. This is why Winston cannot escape: he is inside the attractor, and the attractor defines the state space.

---

6. Newspeak: Restricting the State Space

Newspeak is the most original element of Orwell's vision. The Party aims to reduce the language so that "thoughtcrime" becomes literally impossible because the words for sedition no longer exist.

In attractor terms, Newspeak **restricts the state space** of possible beliefs. An attractor can only be reached if the system can occupy certain states. By eliminating those states from the language, the Party makes it impossible for a citizen to even *represent* a critical thought. The attractor basin for rebellion shrinks to zero.

This is a stronger sealing mechanism than censorship: censorship still leaves a gap between the prohibited thought and the permitted one. Newspeak removes the gap entirely. The citizen cannot correct because they cannot think the error.

---

7. The Impossibility of Internal Escape (and the Appendix)

A key claim of the attractor framework is that a fantasy attractor with $\kappa \gg 1$ cannot be exited by internal forces alone. The system must be perturbed from outside (e.g., a revolution, a collapse of the regime). In *1984*, the novel presents **no successful internal exit**. Winston's attempts fail. The Party remains.

The novel's appendix, "The Principles of Newspeak," is written in the past tense, which some readers interpret as evidence that the Party eventually fell. Others argue it is merely an editorial device. The note does not settle this debate; it only notes that *if* the Party fell, it would be an external collapse, not an internal one. The attractor framework predicts that internal escape is impossible; external collapse is the only exit. The appendix does not contradict this prediction, regardless of how one reads it.

---

8. Falsifiability Condition

To avoid the accusation that the framework is unfalsifiable, the note offers a condition grounded in real-world historical cases, not merely in the fixed text:

> *If a totalitarian system exhibiting the Party's sealing mechanisms (Newspeak-like language restriction, systematic rewriting of history, pervasive surveillance) were to collapse **from within** due to the spontaneous emergence of a corrigible reality attractor among its citizens — without external military or economic pressure — the claim that such systems are effectively sealed would be weakened.*

The framework predicts that internal collapse is highly

unlikely; external perturbations are required. Historical examples (e.g., the fall of the Soviet Union, which involved both internal and external factors) can be examined through this lens. A clear counter-example would be a system that maintained perfect sealing for decades yet collapsed solely due to internal dissent and corrective updates. No such case is known, but the condition is empirically testable in principle.

---

9. Comparison with Milton and Spinoza

The attractor framework can place *1984* on a spectrum of sealedness:

- **Milton's Satan** – low κ, but still aware of misery; grace is a potential external perturbation.
- **Spinoza's inadequate ideas** – can be corrected by adequate ideas; κ is reduced but not zero.
- **Orwell's Party** – κ ≈ 1, no internal exit, total sealing maintained through adaptive feedback suppression.

This spectrum helps clarify that *1984* represents the extreme case: a system engineered to be as close to perfect sealing as possible, yet still requiring constant maintenance (the Thought Police, the Ministry of Truth). Even the Party cannot achieve literal κ = 0; it can only approach it asymptotically.

---

10. Conclusion

Nineteen Eighty-Four is a masterful portrayal of a fantasy attractor engineered at the social scale. The Party uses Newspeak, doublethink, the mutable past, and the Thought

Police to create a belief system with **effectively zero corrective permeability**. Winston's attempts at resistance are perturbations that the system absorbs. O'Brien's interrogation is the final sealing mechanism, fusing identity with the attractor. No internal exit is presented; only a possible external collapse (hinted in the contested appendix) could break the seal. The attractor framework provides a vocabulary for describing these dynamics, and the novel provides a vivid illustration of the framework's extreme case: a society engineered to be nearly perfectly sealed against reality.

Suggested citation: Galida, R. S. (2026). 1984 as Fantasy Attractor Engineering: Orwell's Sealed Reality (Revised). *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 ($\kappa = 1/\tau$). 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.*

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:

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

where $\Delta\kappa$ is a scalar representing the reduction in corrective permeability due to the combined effect of functional and (if applicable) phenomenal factors. A plausible functional operationalization: $\Delta\kappa \propto$ (frequency of identity-reinforcing reward signals) × (temporal persistence of goal representation). **Crucially, this same functional $\Delta\kappa$ applies to non-conscious optimizers as well; for conscious systems, an additional unquantified term for phenomenal investment would be added. The notation is illustrative, not a closed model.**

When human feedback arrives, the AI *detects* the discrepancy (intelligence intact) but if $\Delta\kappa$ 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; $\kappa = 1/\tau$, 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 $\kappa(d)$ 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 Δκ(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 Δκ(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)
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; $\kappa = 1/\tau$, 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.*