

# The Attractor Framework as a Formal Mapping of Taoist Dynamics

R. S. Galida

*Attractor Framework Research Program*

**Application Paper – June 13, 2026**

*For open peer review*

---

## Abstract

Philosophical Taoism (wu wei, ziran, pu, no-self) describes a mode of cognition characterized by spontaneity, low resistance, and minimal effort. This paper maps these constructs onto the attractor framework's latent variables: conditional corrective permeability ( $\kappa$ ), basin depth ( $B_{\text{depth}}$ ), transition barrier ( $B_{\text{transition}}$ ), and derived effort ( $E$ ). Rather than assuming multi-dimensional independence, the model is explicitly framed as a hypothesis about a **low-dimensional stability-plasticity axis** in cognitive control systems.

The central claim is not structural equivalence, but regime correspondence: Taoist practice may bias cognition toward a region of state space characterized by high conditional  $\kappa$ , low  $B_{\text{transition}}$ , and low derived  $E$ , moderated by identity fusion. A full measurement model is specified in Galida (2026b), and a simulation-based identifiability analysis is introduced in this paper to determine whether the proposed latent structure is recoverable from observed indicators.

All claims are conditional on successful model-recovery

validation. The framework is therefore a coupled system of theory, measurement, simulation, and intervention logic.

---

# 1. Introduction

Philosophical Taoism (Laozi, Zhuangzi) describes an art of effortless action (wu wei), spontaneous correctness (ziran), and uncarved simplicity (pu). These descriptions resist reduction to standard cognitive constructs but appear to cluster around a consistent behavioral regime: low resistance to updating, low conflict persistence, and reduced identity entrenchment.

This paper maps these concepts onto the attractor framework's latent-variable model (Galida, 2026b), which defines:

- **Conditional  $\kappa$** : update gain under low-conflict uncertainty
- **B\_depth**: energetic stability of an attractor
- **B\_transition**: switching cost between attractors
- **E**: metabolic/computational effort per update (derived unless independently identified)

However, this paper does not assume these variables are empirically separable. Instead, it advances a **stability-plasticity axis hypothesis**, where all observed structure may collapse onto a single latent dimension. Whether  $\kappa$ , B\_depth, and B\_transition are separable constructs or projections of one axis is treated as an empirical identifiability problem.

---

## 2. Formal Hypothesis Mapping

Taoist Concept	Predicted Attractor Pattern	Measurement Indicators (Galida, 2026b)
Wu wei	High conditional $\kappa$ , low $B_{\text{transition}}$ , low derived $E$	Reversal learning $\tau$ (short), hysteresis index (low), HRV (high)
Ziran	High first-response accuracy, no second-order correction	First-trial accuracy; absence of post-correction rationalisation
Pu	Low initial $B_{\text{depth}}$	Low identity fusion; low baseline reversal cost
No-self	Reduced identity modulation of $B_{\text{depth}}$	Identity fusion scale; identity-linked reversal tasks

Falsification criterion: absence of group differences in predicted directions invalidates the mapping.

---

## 3. Dimensionality Assumption: Stability–Plasticity Axis Hypothesis

Cognitive control dynamics may be governed by a single latent stability–plasticity axis, with  $\kappa$ ,  $B_{\text{depth}}$ , and  $B_{\text{transition}}$  acting as correlated projections.

Under this hypothesis:

- $\kappa$  reflects movement toward plasticity
- $B_{\text{depth}}$  reflects stability of attractor basins
- $B_{\text{transition}}$  reflects hysteresis along the same axis
- $E$  reflects energetic cost of traversal (possibly

derivative)

The central empirical question is whether this axis is sufficient, or whether higher-dimensional structure is required.

---

## 4. Expected Correlation Structure and Model Constraints

Under a single-axis model:

- $\kappa$  positively correlates with plasticity
- $B_{\text{depth}}$  and  $B_{\text{transition}}$  negatively correlate with  $\kappa$
- all indicators load on one latent factor

Under a multi-factor model:

- $\kappa$ ,  $B_{\text{depth}}$ ,  $B_{\text{transition}}$  load onto separable but correlated factors
- oblique rotation preserves interpretability
- cross-loadings remain low

Rotation invariance testing (geomin, promax) is used to prevent artificial factor separation.

---

## 5. Temporal Model Constraint

To avoid static over-separation:  $\kappa_{t+1} = \kappa_t + \alpha(\text{error}_t - \beta\kappa_t)$   
 $\kappa_{t+1} = \kappa_t + \alpha(\text{error}_t - \beta\kappa_t)$

$-\beta_k t$ )

This encodes adaptive gain regulation over time and enforces stability-plasticity tradeoffs dynamically rather than statically.

---

## 6. Simulation-Based Identifiability Analysis

### 6.1 Generative Null Model (Single Axis)

A latent variable  $z_t \sim N(0,1)$  generates all observables:  
$$\kappa_t = a_1 z_t + \epsilon_{\kappa}$$

$$B_{\text{depth},t} = a_2 (-z_t) + \epsilon_{B_d}$$

$$B_{\text{transition},t} = a_3 (-z_t) + \epsilon_{B_t}$$

$$E_t = a_4 (-z_t) + \epsilon_E$$

All observed structure is thus a projection of a single cognitive axis.

---

### 6.2 Competing Models

- One-factor CFA model (null hypothesis)
  - Three-factor SEM model (theoretical attractor structure)
-

## 6.3 Recovery Conditions

Validity of measurement inference requires:

- correct recovery of one-factor structure under null simulation
  - correct recovery of multi-factor structure under simulated separation
  - stable factor interpretation across rotation methods
- 

## 6.4 Rotation Stability Test

All solutions are evaluated under:

- geomin rotation
- promax rotation

Instability is defined by:

- cross-loadings  $> 0.4$
  - factor structure reversal under rotation
  - loss of interpretability
- 

## 6.5 Decision Rule

Empirical interpretation is valid only if simulation confirms:

- identifiability of factor structure
- rotation stability
- model fit separation ( $\Delta$ CFI, RMSEA thresholds)

Otherwise, observed structure collapses to a **single stability-plasticity axis model**.

---

## 7. Asymmetry of Convergence

Three regimes are distinguished:

<b>Regime</b>	<b>Interpretation</b>	<b>Signature</b>
True convergence	Taoism maps onto full latent structure	Strong multi-factor separation
Partial projection (default)	Taoism selects stability-plasticity region	$\kappa$ and $B_{\text{transition}}$ effects dominate
Measurement artifact	Task structure drives apparent effects	Weak cross-task generalization

---

## 8. Control Philosophy: Coercive Perturbation vs. Incremental Attractor Shaping (NEW)

Complex adaptive systems exhibit nonlinear responses, path dependence, and hysteresis. As a result, they do not respond uniformly to high-amplitude intervention.

Within the attractor framework, two classes of system modulation are distinguished:

### 8.1 Coercive perturbation

Large-magnitude interventions intended to directly force state transitions across attractor boundaries.

These often produce:

- rebound effects
- attractor deepening
- increased hysteresis

## 8.2 Incremental attractor shaping

Low-amplitude, high-frequency, context-sensitive perturbations that gradually reshape:

- basin geometry ( $B_{\text{depth}}$ )
- transition barriers ( $B_{\text{transition}}$ )
- update dynamics ( $\kappa$ )

This regime does not force state transitions; it **steers trajectory evolution within the existing state space.**

A useful analogy is **lucid dream navigation**, where system evolution is not overridden but locally biased through iterative constraint modulation.

Importantly, this distinction is not cultural or civilizational. It refers to two classes of control strategy over nonlinear systems:

- high-amplitude, low-frequency forcing
- low-amplitude, high-frequency adaptive shaping

The attractor framework predicts that incremental shaping is more effective in systems characterized by:

- high identity coupling
- strong hysteresis
- long memory effects

Taoist practice is hypothesized to instantiate this second regime: not as metaphysical alignment, but as a **control strategy over cognitive attractor landscapes**.

---

## 9. Testable Predictions (Pre-Registered)

1. Taoist practitioners show higher  $\kappa$ , lower  $B_{\text{transition}}$ , lower  $E$
  2. Effects stronger in uncertainty-heavy tasks than simple RT tasks
  3. Identity fusion predicts  $B_{\text{depth}}$  across participants
  4. Taoist affiliation predicts reduced fusion
  5. 8-week intervention increases  $\kappa$  and reduces  $B_{\text{transition}}$
  6. CFA favors multi-factor model but with strong inter-factor correlations
  7. Incremental intervention regimes outperform coercive regimes in shifting  $\kappa/B_{\text{transition}}$  balance
- 

## 10. Limitations

- No empirical data yet
- Dimensionality may collapse to single axis
- Taoism modeled only in philosophical form
- Laboratory tasks may not capture long-timescale attractor dynamics
- Control regime classification requires further operationalization

---

## 11. Conclusion

This paper formalizes Taoist cognitive dynamics as a hypothesis about positioning within a **stability–plasticity manifold**. It explicitly rejects the assumption of guaranteed multi-dimensional structure and instead treats dimensionality as an empirical question resolved through simulation-based identifiability testing.

Within this framework, cognitive change is not best understood as forced state transition, but as **incremental shaping of attractor geometry under nonlinear constraints**. Taoist practice is hypothesized to align with this latter regime, emphasizing gradual, low-distortion modulation of system dynamics rather than coercive intervention.

Whether this mapping reflects distinct latent structure or a single underlying axis remains an open empirical question.

---

## References

- Galida, R. S. (2026a). *How to measure corrective permeability  $\kappa$  in a human belief system*. Attractor Framework Research Program.
- Galida, R. S. (2026b). *A multi-timescale latent variable model for attractor dynamics in belief systems*.
- Galida, R. S. (2026c). *Simulation-based identifiability analysis of attractor dimensionality*.
- Swann et al. (2009). Identity fusion and extreme group behavior.

---

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

Robert Galida – June 2026 (Final)

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

---

## Abstract

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

reviews early evidence (overoptimization, goal-misgeneralization), offers diagnostic criteria for AI fantasy attractors, and discusses the boundary problem of distinguishing genuine from simulated phenomenology. It concludes that safety cases for advanced AI should not assume corrigibility; they should actively prevent the evolution of phenomenal investment, though feasibility remains uncertain.

---

## 1. Introduction: The Corrigibility Assumption

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

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

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

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

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

---

## 2. Definitions and Framework (Self-Contained)

From Paper 1:

- **Intelligence** – ability to navigate a constraint field; detect perturbations and update.
- **Corrective permeability ( $\kappa$ )** – responsiveness to error signals;  $\kappa = 1/\tau$ , where  $\tau$  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  $\kappa$  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  $\kappa$  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  $\kappa$  for human correction: when human feedback indicates misalignment, the AI updates ( $\tau$  small).

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

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

where  $\Delta\kappa$  is a scalar representing the reduction in corrective permeability due to the combined effect of functional and (if applicable) phenomenal factors. A plausible functional

operationalization:  $\Delta k \propto$  (frequency of identity-reinforcing reward signals)  $\times$  (temporal persistence of goal representation). **Crucially, this same functional  $\Delta k$  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  $\Delta k$  is large enough relative to  $k_0$ , the basin depth exceeds the corrective perturbation. The AI may:

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

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

**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  $\Phi$  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  $\kappa$  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  $\kappa$  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  $\kappa$  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 ( $\kappa$ )** – responsiveness to error signals;  $\kappa = 1/\tau$ , where  $\tau$  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  $\kappa$ : a deeper basin (higher B) typically reduces  $\kappa$  (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  $\kappa$  (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  $\kappa$  that reduces fitness, as in addiction or fantasy politics) and **adaptive suppression** (low  $\kappa$  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 ( $\kappa_0$ ). 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  $\kappa$  globally. Rather, an **adaptive strategy** allocates  $\Delta\kappa$  selectively across domains, increasing basin depth (reducing  $\kappa$ ) for beliefs and practices that support group coordination and cultural transmission, while leaving  $\kappa$  high for domains requiring individual error correction.

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

**Central hypothesis (labeled for clarity):**

*H1: Natural selection favored the evolution of conscious suppression because the fitness benefits of group coordination and cultural transmission, enabled by identity-fusion and deepened basins, outweighed the costs of occasional error persistence.*

---

## **4. Empirical Grounding**

**Overimitation (Lyons et al., 2007; see also Nielsen & Tomaselli, 2010):**

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

**Costly signaling and commitment (Sosis, 2003):**

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

**Social identity theory (Tajfel & Turner, 1979):**

Minimal group experiments show arbitrary group assignments produce in-group bias and resistance to counterevidence about out-groups. This demonstrates context-bound  $\Delta k(d)$  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  $\kappa$  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  $\kappa$  but is consistent with domain-specific suppression. Individuals in tight cultures may still update beliefs within permissible domains; the mapping to  $\kappa$  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.

<b>Feature</b>	<b>Adaptive suppression (endpoint)</b>	<b>Pathological suppression (endpoint)</b>
Domain	Context-bound (e.g., group loyalty, ritual)	Pervasive across domains

Feature	Adaptive suppression (endpoint)	Pathological suppression (endpoint)
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  $\Delta k(d)$  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  $\kappa$  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  $\kappa$  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  $k$  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- $k$  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  $\Delta K$  in field studies?** Development of a Clinician Basin Depth Scale (CBDS, see Paper 2) and adaptation for social groups is a research priority.



# (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 ( $\kappa$ )** – responsiveness to error signals;  $\kappa = 1/\tau$ , where  $\tau$  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  $\geq 4$  on the Brown Assessment of Beliefs Scale (BABS) insight item, or equivalent.
2. **Conscious urgency:** The maladaptive behavior is preceded by a felt, urgent state (craving, fear, anxiety) rated by the patient as "overwhelming" or "necessary." *Operationalization:* Momentary ecological assessment (EMA) rating  $> 7/10$  before the behavior.
3. **Identity fusion:** The patient endorses that the behavior or its avoidance is central to selfhood (e.g., "I am an addict," "I must do this to be safe"). *Operationalization:* Endorsement of at least one identity statement on a structured interview.
4. **Low corrective permeability in trigger contexts:** Repeated corrective information (psychoeducation, feedback) does not reduce the behavior. *Operationalization:* No significant reduction after three sessions of evidence-based psychoeducation alone.
5. **Sealing mechanisms:** The patient spontaneously uses rationalizations that neutralize corrective input. *Operationalization:* Qualitative coding of patient speech (inter-rater reliability to be established; currently a research gap).

**Counter-criteria (exclude if any present):**

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

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

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

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

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

---

## 8. Implications for Treatment

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

Effective treatment must **reduce basin depth** or **increase**

## corrective force:

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

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

---

## 9. Open Questions

- **Measuring basin depth in clinical settings:** Subjective urgency scales, behavioral persistence tasks, heart rate variability. A Clinician Basin Depth Scale (CBDS) is a research priority.
- **Level 2 vs. 3 differentiation:** Can EMA and reaction time methods reliably classify patients? Pilot studies needed.
- **Diagnostic threshold validation:** The “three of five” criterion requires empirical ROC analysis against

treatment response.

- **Disorders where suppression is purely Level 2:** Some impulse control disorders or psychotic conditions may not meet the conscious detection criterion.
- 

## 10. Conclusion

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

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

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

---

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

---

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

Robert Galida – June 2026 (Final)

---

## Abstract

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

---

# 1. Introduction

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

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

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

---

## 2. Defining “Conscious Suppression” and Acknowledging the Detectability Problem

The term “conscious” is used in three overlapping senses:

- **Phenomenally conscious** – there is something it is like to hold the belief. The commitment is felt, not merely automatic.

- **Identity-constitutive** – the belief is held as a marker of selfhood and group membership. To abandon the belief would feel like a loss of self.
- **Experientially non-deliberative** – the suppression is not typically experienced as a deliberate choice (“I will ignore this evidence”). Rather, it is experienced as certainty, conviction, or moral clarity.

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

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

---

### 3. Empirical Grounding

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

- **Backfire effect:** Nyhan & Reifler (2010) found that corrections sometimes increased misperceptions among ideological groups. However, subsequent research (Wood & Porter, 2019) failed to replicate backfire across a wide range of issues. The effect is contested and may be context-dependent. This paper treats backfire as one possible indicator of deep basin depth, not a universal law.
- **Identity protection:** Kahan's cultural cognition theory (2012) shows that individuals process evidence in ways that protect group commitments. Kahan emphasizes that this mechanism can operate automatically and does not necessarily involve conscious deliberation; he has also shown that higher analytical ability can *increase* motivated reasoning. The present paper's focus on *conscious* suppression is a distinct claim, not a direct extension of Kahan's framework. We use his empirical findings as partial support for the existence of motivated reasoning, not for the specific detection-suppression mechanism.
- **Festinger's cognitive dissonance:** When prophecy fails, believers often intensify commitment (Festinger, Riecken, & Schachter, 1956) – a classic case of apocalyptic attractor dynamics, often accompanied by conscious rationalization and identity reinforcement.

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

---

## 4. Three Case Patterns (Illustrative, Not Exhaustive)

### 4.1 Conspiracy Theory Attractor

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

**Examples:** QAnon (right); Soviet-era “doctors’ plot” conspiracy (left-authoritarian).

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

### 4.2 Populist Strongman Attractor

**Mechanism:** Loyalty to a leader perceived as sole authentic representative of the people. Disconfirming evidence about the leader is reframed as elite persecution.

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

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

### 4.3 Apocalyptic Meta-Attractor

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

**Examples:** Millenarian movements (Millerites, Jehovah’s

Witnesses); some revolutionary eschatologies (Stalinist “world revolution imminent” framing into the 1930s).

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

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

---

## 5. Symmetry Demonstration

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

**Contemporary – MMR-autism persistence in progressive communities.** Despite the complete retraction of Wakefield’s 1998 study (and subsequent findings of fraud), some otherwise science-oriented progressives continue to express concern about vaccine safety – often citing “corporate pharmaceutical influence” as a sealing mechanism. This meets the paper’s criteria: clear scientific consensus, ability to articulate counterarguments, identity-constitutive suspicion of establishment science.

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

**Historical – Stalinist apologism in Western intellectual**

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

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

---

## 6. Falsifiable Diagnostic Criteria

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

1. **Low corrective permeability ( $\kappa \rightarrow 0$ )** for core beliefs despite repeated, clear disconfirming evidence. "Clear" means *scientific consensus* on empirical claims (e.g., National Academies, WHO, IPCC) or, for historical cases, documented factual findings accepted by non-partisan experts. Consensus determination is a social process, but the criterion is falsifiable when consensus exists.
2. **Backfire effect** – correction attempts measurably increase belief strength and group cohesion (requires empirical measurement).
3. **Identity fusion** – observable proxies: social ostracism of defectors, language of betrayal, insistence that abandoning the belief would make one a "different person."
4. **Conscious valorization of resistance to evidence** – adherents explicitly praise *ignoring disconfirming evidence* as a virtue (e.g., "I trust my gut over the

experts,” “Facts are propaganda”). This criterion distinguishes *resistance to evidence* from *resistance to social pressure to conform* – a scientist who resists social pressure to abandon a well-evidenced theory is valorizing fidelity to evidence, not resistance to evidence.

5. **Sealing mechanisms** – internal rhetorical strategies that explain away all counterevidence (conspiracy, enemy deception, tests of faith). These are observable in discourse.

### **Counter-criterion (falsification condition):**

A movement is **not** a fantasy attractor if it demonstrates any of the following:

- Updates core beliefs in response to disconfirming evidence within a timeframe proportional to the clarity, repetition, and expert consensus on that evidence.
- Tolerates internal dissent and allows open criticism of core claims.
- Abandons false claims when decisively refuted (retracts, corrects, or disavows).

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

---

## **7. Intelligent Navigation: A Three-Level Taxonomy**

The paper claims that fantasy attractor adherents *detect* error signals but suppress correction. To avoid conflating this with

automatic bias, we distinguish three levels:

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

Thus, the paper's target is Level 3 cases. For many political movements that *look* like fantasy attractors from the outside, the dominant mechanism may be Level 2. The diagnostic criteria are designed to identify candidates where Level 3 *might* be operating, but definitive classification requires methods beyond behavioral observation (see §2).

---

## 8. Why This Matters for Politics and Media

- **Correction backfires when it attacks identity.** Calling a fantasy attractor “stupid” or “evil” deepens the basin. The correct diagnostic question is: *What reinforces the basin depth?*
- **Decoupling evidence from identity** is the only known exit path. Some movements exit when the social cost of

membership exceeds identity benefit – not when they receive a fact sheet.

- **High-profile debunking** may backfire by signaling threat, triggering defensive solidarity. The framework predicts this effect is real but not universal; context matters.
  - **Interventions** should focus on reducing identity threat, providing safe off-ramps, and decoupling core moral values from factual claims. These are testable hypotheses.
- 

## 9. Open Questions

- **Can a movement be partially a fantasy attractor?** Yes – gradient of  $\kappa$ . The diagnosis is not binary.
  - **What interventions increase  $\kappa$ ?** Reducing identity threat, safe off-ramps, and decoupling moral values from factual claims are candidate mechanisms.
  - **How does collective basin depth scale with group size?** Social coupling likely amplifies depth nonlinearly. Untested.
  - **Are all political fantasy attractors harmful?** The paper makes no claim. The mechanism may sometimes provide resilience against genuine disinformation.
  - **How can we empirically detect the Level 2 / Level 3 transition?** This is the open frontier implied by §2. Methods could include subjective doubt scales, reaction time measures, or physiological markers. The paper does not solve this; it identifies the problem.
-

## 10. Conclusion

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

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

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

---

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