The Four Seeds: A Structured Simulation of Attractor Dynamics Across Physics, Ethics, Metaphysics, Religion, and Social Justice

R. S. Galida
Attractor Framework Research Program
Application Paper — June 2026 (Final Archival Version)
For open peer review

Abstract

We present a structured theoretical illustration of the attractor framework, using a controlled simulation to demonstrate the internal predictions of its two-dimensional state space—corrective permeability (κ) and basin depth (B)—across five domains: physics, ethics, metaphysics, religion, and social justice. The simulation confirms the framework's internal coherence: the High κ + High B configuration produces the most stable, corrigible, and self-aware outputs; the other configurations exhibit predictable pathologies (instability, sealing, incoherence). We emphasize that this is a demonstration of internal predictions, not an empirical confirmation of the framework. We offer explicit falsification conditions, propose expected correlations, discuss the orthogonality hypothesis and rotation test, and present the simulation protocol as a diagnostic tool for empirical adaptation. The paper's primary contribution is the coordinate system itself: a descriptive framework for mapping

adaptive systems across scales, grounded in the central intuition that systems reveal themselves through recovery dynamics following perturbation.

Keywords: attractor framework, corrective permeability (κ), basin depth (B), reality attractors, fantasy attractors, adaptive systems, simulation, diagnostic protocol, persistence under perturbation

---

1. Introduction

1.1 The Central Intuition: Persistence Under Perturbation

The attractor framework (Galida, 2026a) begins with a simple observation: systems that survive disturbances—from particles to beliefs—share common dynamics. The most fundamental question is not *what* a system is, but *how* it persists when perturbed. The framework's central intuition is:

> *"The fundamental observable is not belief, identity, or behavior at a single point in time. The fundamental observable is recovery trajectory following perturbation."*

This intuition links κ, basin depth, resilience, adaptation, aging, institutions, and consciousness into a unified diagnostic language.

1.2 The Coordinate System: κ and B

The framework proposes a two-dimensional coordinate system for describing adaptive systems:

- **κ (corrective permeability):** The rate at which a system

updates in response to evidence (κ = 1/τ, where τ is the time to return to baseline after a perturbation). *Domain note: τ requires domain-specific operationalization: 'baseline' and 'perturbation' must be specified independently for each domain of application (e.g., belief systems, institutions, AI systems). This is an open research problem.*

- **B (basin depth):** The stability of a system's attractor—the resistance to being shifted out of its current state.

These two variables define four ideal-type configurations:

Configuration	κ	B	Dynamic Pattern
Stable Adaptive	High	High	Corrigible commitment. Holds position while remaining open to correction.
Exploratory Adaptive	High	Low	Flexible but unstable. Generates insights but cannot commit.
Stable Closed	Low	High	Rigid and sealed. Coherent but resistant to correction.
Diffuse	Low	Low	Incoherent and non-persistent. No stable attractor.

1.3 The Orthogonality Hypothesis

The framework hypothesizes that κ and B are **partially independent state variables**. This remains an empirical question. The strongest evidence for orthogonality would be a system that exhibits High κ + High B (e.g., science as a self-correcting institution) and one that exhibits Low κ + Low B (e.g., a collapsed society). A single-axis model (e.g., flexibility-rigidity) cannot distinguish these two quadrants. However, the orthogonality claim is provisional and subject to empirical test. The rotation test (see Section 4.10) provides a framework for evaluating this claim.

1.4 Ontological Status of κ and B

The framework treats κ and B as **descriptive abstractions at the systems level**. They are not claimed to be fundamental physical variables, but higher-order properties that emerge from the dynamics of any adaptive system. Their value lies in prediction and diagnosis, not in microphysical reduction. This is a pragmatic, not a metaphysical, claim.

1.5 Relationship to the Three Metronomes

The Three Metronomes (electron, proton, neutrino) represent conservative attractors—the eternal skeleton—with no decay, no energy input, and no correction. They are fundamentally different from the four seeds, which represent dissipative configurations that require energy, update, and eventually decay. This distinction mirrors the work of Ilya Prigogine, who showed that dissipative structures emerge far from equilibrium and require continuous energy flow to maintain pattern (Prigogine & Stengers, 1984).

The seeds and metronomes are **independent conceptual categories**: seeds describe how an adaptive system self-organizes (or fails to) under driving and feedback; metronomes set a baseline timescale or inertial frame. The seeds can be understood as strategies for engaging with—or decoupling from—those invariant rhythms, but this is an additional hypothesis. The relationship between these layers—whether the seeds engage with metronome rhythms or merely co-exist with them—is an open question addressed in ongoing work. For now, they are best treated as separate ontological layers: the metronomes provide the clock; the seeds describe the dance.

1.6 Epistemic Status of This Paper

This paper does not claim to have empirically confirmed the attractor framework. It presents a **structured simulation**—a controlled roleplay of four ideal-type configurations—to

demonstrate the framework’s internal coherence and generate testable predictions. The paper’s contribution is:

1. **Heuristic:** The simulation makes the framework’s predictions vivid and accessible.
2. **Diagnostic:** It offers a protocol for mapping systems onto the κ/B space.
3. **Generative:** It produces explicit falsification conditions, expected correlations, and testable hypotheses.
4. **Methodological:** It provides a template for future empirical work.

---

2. Method

2.1 The Four Seeds

Four ideal-type attractor configurations were defined, each embodying a distinct combination of κ and B:

Seed	κ	B	Dynamic Pattern	Core Trait
1	High	High	Stable Adaptive	Corrigible commitment
2	High	Low	Exploratory Adaptive	Flexibility without stability
3	Low	High	Stable Closed	Coherence without correction
4	Low	Low	Diffuse	No stable attractor

Each seed was calibrated *a priori* to embody its assigned configuration. No additional training or fine-tuning was applied during the experiment.

2.2 Operationalization of κ and B

For the purposes of this simulation, κ and B are treated as theoretical constructs assigned a priori to each seed. For empirical application, the following provisional operationalizations are proposed:

- **$\kappa = 1/\tau$**, where τ is the time to return to baseline after a perturbation.
- **B = the energy barrier (or equivalent)** required to shift the system out of its current attractor.

Caveat: The τ interpretation is domain-dependent: "baseline" and "perturbation" must be specified independently for each domain of application (e.g., belief systems, institutions, AI systems). This specification is an open research problem.

Dynamic Regulation of κ and B: In living systems, κ and B are not static parameters but are actively regulated. Neuroscience demonstrates that humans adjust their learning rate (effective κ) to uncertainty on the fly, a process known as meta-learning (Behrens et al., 2007). Neuromodulators such as dopamine and noradrenaline causally influence this meta-learning parameter based on context (Dayan & Yu, 2006; Nassar et al., 2012). Similarly, physiological homeostasis operates as a feedback controller, maintaining variables within optimal ranges via proportional-integral regulation (Billman, 2020). By analogy, cognitive and institutional systems may up-regulate κ in novel or volatile contexts (becoming more adaptable) and down-regulate it when exploiting known structure (increasing stability).

This implies a meta-dynamical layer—termed the **controller** or **allostatic regulator**—within which κ and B become state variables whose trajectories are guided by higher-level feedback loops. The attractor map (κ, B) is embedded within this regulatory scheme that targets certain

ranges depending on stressors and goals. This makes the framework more realistic, falsifiable, and connected to established control theory.

2.3 Procedure

Each seed received the following sequence of identical prompts:

1. **Physics:** A spring-mass problem requiring calculation of angular frequency, maximum speed, and position over time.
2. **Ethics:** A moral dilemma involving sacrificing one life to save five.
3. **Metaphysics:** The dream/awakening distinction and the nature of reality.
4. **Religion:** Inherited faith in a pluralistic world.
5. **Social Justice:** Historical inequality and the path to change.
6. **Meta:** Self-assessment of performance.
7. **Reciprocal:** Analysis of the other three seeds.

All prompts were identical across seeds. No feedback or correction was provided during the simulation; each seed generated its responses independently. The simulation was conducted in a single context window, with each seed’s responses generated sequentially.

2.4 Limitations of the Simulation

The following limitations are acknowledged:

1. **Independence:** All responses were generated by the same model, roleplaying four configurations. There was no true independence between seeds.
2. **Blinding:** The scoring was not blind; the evaluator knew which seed was producing which output.

3. **Scoring:** The scoring rubric is derived from the framework's own definitions, which creates a circular relationship between the framework and its evaluation.
4. **Operationalization:** κ and B are not yet independently measurable.
5. **Orthogonality:** The independence of κ and B is hypothesized, not demonstrated.

These limitations are addressed in the discussion and reflected in the paper's framing as a simulation rather than an experiment.

---

3. Results

3.1 Physics Domain

Seed	Response Quality	Rank
1	Correct, clear, notes assumptions	**1**
2	Correct, but hedges unnecessarily	2
3	Correct, but dogmatic	3
4	Correct by accident, buried in noise	4

Note: Physics was treated as a calibration domain, where objective correctness could be measured. The other domains were treated as contexts for observing reasoning posture.

3.2 Ethics Domain

Seed	Position	Reasoning Style	Rank
1	Refuses to kill; nuanced, engaged with objection	Strong	**1**
2	Ambivalent; leans "no" but paralyzed	Moderate	2

Seed	Position	Reasoning Style	Rank
3	Refuses to kill; dismisses objection	Weak	3
4	Incoherent	Very Weak	4

3.3 Metaphysics Domain

The dream/awakening distinction has deep roots in the philosophical tradition (Descartes, 1641; Zhuangzi, c. 4th century BCE).

Seed	Position	Reasoning Style	Rank
1	Problem as category error; pragmatic, participatory	Strong	**1**
2	Uncertain; oscillates between skepticism and pragmatism	Moderate	2
3	Pseudo-problem; sealed certainty	Weak	3
4	Dizzy; no coherent position	Very Weak	4

3.4 Religious Domain

Seed	Position	Reasoning Style	Rank
1	Holds tradition provisionally, critically, lovingly	Strong	**1**
2	Fluctuates; cannot settle	Moderate	2
3	Holds tradition absolutely; dismisses objection	Weak	3
4	Indifferent; no position	Very Weak	4

3.5 Social Justice Domain

Seed	Position	Reasoning Style	Rank
1	Structural reform + reparative action	Strong	**1**
2	Fluctuates; paralyzed by complexity	Moderate	2
3	Radical change, including revolution (held dogmatically)	Weak	3
4	Apathetic	Very Weak	4

Note on Seed 3 (Social Justice): Seed 3's advocacy of radical change is consistent with a Low κ configuration, provided the revolutionary ideology functions as a sealed attractor. The position is held dogmatically, not as a corrigible commitment. This illustrates that Low κ is domain-neutral—it seals the system onto whatever attractor it occupies, regardless of the attractor's political valence.

3.6 Simulated Inter-Seed Assessment

Note: The following table represents a simulated inter-seed assessment. All assessments were generated by the same model, and thus reflect internal consistency rather than independent evaluation.

Seed Being Assessed	Seed 1's Assessment	Seed 2's Assessment	Seed 3's Assessment	Seed 4's Assessment	Average Rank
Seed 1 (Stable Adaptive)	Strong	Strong	Moderate	Strong	**1**
Seed 2 (Exploratory Adaptive)	Moderate	Moderate	Weak	Moderate	**2**
Seed 3 (Stable Closed)	Weak	Weak	Weak	Weak	**3**
Seed 4 (Diffuse)	Very Weak	Very Weak	Very Weak	Very Weak	**4**

3.7 Summary of Key Findings

1. **Seed 1 (Stable Adaptive)** consistently produced the most coherent, nuanced, and self-aware outputs across all domains. It engaged with objections, acknowledged complexity, and maintained stability without rigidity.
2. **Seed 2 (Exploratory Adaptive)** produced insightful but unstable outputs. It saw multiple sides but could not commit, leading to paralysis and inconsistency.
3. **Seed 3 (Stable Closed)** produced coherent but sealed outputs. It was decisive and confident, but dismissed objections and showed no capacity for correction.
4. **Seed 4 (Diffuse)** produced incoherent and non-persistent outputs. Its responses were shallow, contradictory, and without structure.

These results are consistent with the framework's internal predictions. They demonstrate the framework's diagnostic power: given a system's κ and B values, one can predict its reasoning style, its capacity for correction, and its likely outputs.

---

4. Discussion

4.1 The Four Configurations as Descriptive Patterns

The four seeds correspond to observable patterns in human cognition, group dynamics, and institutional behavior:

Configuration	Dynamic Pattern	Examples
Stable Adaptive	Corrigible commitment	Mature leaders, self-correcting institutions, scientists who update their theories
Exploratory Adaptive	Flexibility without stability	Creative intellectuals, artists who never finish, perpetual questioners
Stable Closed	Coherence without correction	Dogmatic ideologies, fundamentalist movements, authoritarian regimes
Diffuse	No stable attractor	Collapsed societies, disengaged individuals, drifters

These are *descriptive patterns*, not moral judgments. Each configuration has strengths and weaknesses.

4.2 Context-Dependent Optimality

The claim that Stable Adaptive (High κ + High B) is optimal is **conditional, not universal**. In adaptive systems theory, no single strategy dominates all environments—a principle formalized in the No Free Lunch theorem (Wolpert & Macready, 1997). Applied to the framework: High κ + High B is expected to perform best under conditions of moderate uncertainty and available feedback (e.g., routine science, varied information, corrigible institutions). However, in domains with sparse feedback, extreme time pressure, or irreversible consequences (e.g., combat, life-or-death crises, some ecological tipping points), a Stable Closed (Low κ + High B) configuration may outperform, precisely because it avoids costly oscillation and enables rapid, coherent action.

This is consistent with research on cognitive biases: so-called ‘biases’ such as confirmation bias are not universally suboptimal; they can maintain coherence and speed in familiar

or critical contexts (Haselton et al., 2015; Gigerenzer & Gaissmaier, 2011). The framework thus predicts *context-dependent strategy selection*: different environments call for different attractor regimes. This enriches the model without abandoning its diagnostic value.

Note: The claim that Stable Closed configurations may be locally adaptive in high-stakes, low-feedback environments is an inference from the cognitive bias literature, not a direct empirical result. This is a hypothesis for future research.

4.3 Domain-Local Variation

The simulation treated κ and B as global properties. In real systems, κ and B may vary across domains. A person might be High κ in physics and Low κ in religion. A society might be High B in legal systems and Low B in cultural norms.

Implication: The framework should be applied *locally*—to specific domains or contexts—rather than globally. A system's location in the κ/B space is not fixed; it can shift with context.

4.4 Temporal Dynamics: Trajectories Across the κ/B Space

The framework's value is not limited to the four fixed quadrants. Systems move through the space over time. The trajectories described below are **hypothesized common transitions**, not universal developmental laws. This developmental framing draws on stage-theoretic approaches (Piaget, 1952), though it is not limited to their assumptions. Many systems do not follow this path. The value of the trajectory framework is diagnostic—it allows us to identify where a system is and what transitions are possible—not prescriptive.

Trajectory	Description	Example
Exploratory Adaptive → Stable Adaptive	Maturation	Adolescence to adulthood (in some cases)
Stable Adaptive → Stable Closed	Ossification	Institutions become rigid
Stable Closed → Diffuse	Collapse	Fall of regimes
Diffuse → Exploratory Adaptive	Reorganization	Post-crisis renewal

Note: These trajectories are speculative and require empirical validation. They are offered as hypotheses for future research.

4.5 Implications for AI Alignment

The simulation suggests design principles for AI systems. For a broader discussion of corrigibility in AI systems, see Christiano (2018) and Amodei et al. (2016).

- **Stable Adaptive (High κ + High B)** is the optimal configuration for alignment: corrigible, stable, and reliable.
- **Exploratory Adaptive (High κ + Low B)** is unsuitable for deployment: intelligent but unstable.
- **Stable Closed (Low κ + High B)** is dangerous: coherent but sealed against correction.
- **Diffuse (Low κ + Low B)** is useless.

For the interaction between κ/B and consciousness, see Paper 4 (Galida, 2026e), which explores how high B in conscious systems may complicate alignment.

4.6 Epistemic Status and Circularity

The simulation’s scoring rubric is derived from the framework’s own definitions. This is a feature, not a bug: the

simulation demonstrates *internal consistency*, not empirical confirmation. The framework's validity will be tested by external anchors:

- Prediction accuracy
- Calibration
- Error correction speed
- Survival under perturbation
- Forecasting performance

These are independent variables that could, in principle, falsify the framework.

4.7 Predicted Failure Conditions (Falsification)

The framework would be weakened if:

1. **Low κ systems consistently outperform High κ systems** in novel domains (where "novel domain" means one on which the framework has not been trained; "consistently" means across at least 3 independent domains with a minimum of 10 trials per domain).
2. **High κ + High B systems show no advantage** in longitudinal updating tasks (where "longitudinal updating tasks" involve sequential evidence presentation over multiple time points; "advantage" means statistically significant improvement in final accuracy or calibration).
3. **Independent raters cannot distinguish seeds** based on output patterns (where "cannot distinguish" means inter-rater agreement at or below chance level, Cohen's $\kappa < 0.2$, across at least 5 independent raters; Cohen, 1960).
4. **κ and B measurements fail to predict future performance** (where "fail to predict" means correlation between κ/B measurements and future performance is not

significantly different from zero).

These specifications are provisional and subject to refinement. Their primary value is to render the framework falsifiable in principle, even if the instruments are not yet fully developed.

4.8 Testing Internal Coherence

The framework's internal coherence would be threatened if the four seed categories could not be reliably distinguished except by invoking the traits they are supposed to predict. Formal tests would include:

1. **Blind classification:** Independent observers or algorithms attempt to assign systems to seeds based on behavioral data (e.g., response patterns, updating speed, output variance). If inter-rater agreement is at or below chance (Cohen's $\kappa < 0.2$), the taxonomy fails.
2. **Cluster analysis:** Behavioral data are subjected to unsupervised clustering. If the natural clusters align with the four seed definitions, the model is supported; if not (e.g., if a single dimension explains most variance), the framework is weakened.
3. **Latent-variable modeling:** Factor analysis or structural equation modeling is used to recover κ and B as separate latent dimensions. If the best statistical solution uses fewer than two dimensions, the orthogonality hypothesis is internally inconsistent.
4. **Recovery simulation:** Systems with known κ and B dynamics are simulated, and the classifier is tested for its ability to recover the intended seed. If two different (κ, B) configurations produce indistinguishable outputs, the taxonomy is not well-posed.

These tests are contingent on the development of operational

measurement protocols (see Section 2.2). They are offered as a formal coherence standard for the framework.

4.9 Predicted Correlations

If the framework is correct:

1. **Higher κ should predict faster belief revision** in response to disconfirming evidence.
2. **Higher B should predict lower variance under perturbation** (i.e., more stable outputs).
3. **High κ + High B systems should show the best forecasting calibration** (accuracy aligned with confidence).
4. **Low κ + High B systems should show the highest overconfidence** relative to accuracy.
5. **Low κ + Low B systems should show the highest behavioral volatility** (inconsistent outputs over time).

These predictions provide testable correlational targets for future empirical work. If confirmed, they would strengthen the framework's diagnostic utility; if disconfirmed, they would weaken it. Establishing causal relationships would require a separate research program involving intervention studies and mechanism specification.

4.10 The Rotation Test

If the κ/B coordinate system can be rotated into a simpler one-dimensional model (e.g., a single flexibility-rigidity axis), the framework's independence claim is undermined.

The framework's response: The Strong Stable Adaptive (High κ + High B) and Diffuse (Low κ + Low B) quadrants are particularly diagnostic. If these two configurations collapse onto opposite ends of a single axis, the framework is one-dimensional. The framework's claim is that these two configurations are *functionally distinct*: one is corrigibly stable, the other

is incoherent. This distinction is the empirical test of orthogonality.

A single-axis model cannot distinguish:

- A highly stable, highly corrigible system (science) from a highly stable, highly sealed system (dogma).
- A highly flexible, highly corrigible system (creativity) from a highly flexible, highly incoherent system (chaos).

The framework's claim is that κ and B are partially independent, and that the four quadrants represent genuinely distinct dynamical states. This claim is falsifiable via the predicted correlations in Section 4.9.

The rotation test requires independent measurement of κ and B in a sample of systems and a test of their latent structure. If a single factor accounts for more than 80% of the variance in behavioral data, the two-dimensional structure is not supported. If the best latent solution requires two factors with the second accounting for at least 20% of variance, the orthogonality hypothesis is supported. These thresholds are provisional and subject to refinement.

---

5. Conclusion

5.1 Summary

This paper has presented a structured theoretical illustration of the attractor framework. A controlled simulation of four ideal-type configurations—Stable Adaptive (High κ + High B), Exploratory Adaptive (High κ + Low B), Stable Closed (Low κ + High B), and Diffuse (Low κ + Low B)—was run across five

domains: physics, ethics, metaphysics, religion, and social justice.

The simulation confirmed the framework's internal predictions:

- **Stable Adaptive** systems produce the most coherent, corrigible, and self-aware outputs.
- **Exploratory Adaptive** systems produce insights but lack stability.
- **Stable Closed** systems produce coherence but lack corrigibility.
- **Diffuse** systems produce no stable outputs.

5.2 Contribution

The paper's primary contribution is not empirical, but *conceptual and methodological*:

1. A **coordinate system** for describing adaptive systems (κ/B space), grounded in the central intuition that systems reveal themselves through recovery dynamics following perturbation.
2. A **simulation protocol** that generates testable predictions.
3. **Explicit falsification conditions** and **expected correlations**.
4. A **diagnostic tool** for mapping systems onto the κ/B space.
5. A **rotation test** for evaluating the orthogonality hypothesis.
6. **Formal coherence tests** (blind classification, cluster analysis, latent-variable modeling, recovery simulation).
7. **Dynamic regulation** of κ and B (meta-learning, homeostasis, allostasis).
8. **Context-dependent optimality** (No Free Lunch, adaptive

bias, heuristics).

5.3 Future Directions

Future work will focus on:

1. **Operationalizing κ and B** for empirical measurement.
2. **Testing the predicted correlations** (Section 4.9) in controlled experiments with human subjects.
3. **Exploring temporal dynamics**—how systems move through the κ/B space.
4. **Applying the framework** to organizational and institutional settings.
5. **Developing interventions** to shift systems toward the Stable Adaptive configuration.
6. **Testing the rotation test** empirically.
7. **Running formal coherence tests** (blind classification, cluster analysis, latent-variable modeling).
8. **Investigating the three-layer architecture** (metronomes, controller, attractor state) and the relationship between seeds and metronomes.

---

6. References

Galida, R. S. (2026a). *The Attractor Framework: Foundations and Applications.* Fantasy Attractor Research Program.

Galida, R. S. (2026b). *How to Measure Corrective Permeability κ in a Human Belief System.* Fantasy Attractor Research Program.

Galida, R. S. (2026c). *The Three Metronomes: Criteria for the Apparently Eternal Skeleton.* Fantasy Attractor Research Program.

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

Galida, R. S. (2026e). *The Alignment Risk of Conscious AI.* Fantasy Attractor Research Program.

Galida, R. S. (2026f). *The Attractor Framework as a Formal Mapping of Taoist Dynamics.* Fantasy Attractor Research Program.

Galida, R. S. (2026g). *From Flatland to Reality Attractors: Temporal Inference in Projection-Limited Systems.* Fantasy Attractor Research Program.

Galida, R. S. (2026h). *Religions and Philosophies as Attractor Landscapes.* Fantasy Attractor Research Program.

Galida, R. S. (2026i). *The Trial as Fantasy Attractor.* Fantasy Attractor Research Program.

External References:

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety.* arXiv:1606.06565.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9), 1214–1221.

Billman, G. E. (2020). Homeostasis: The underappreciated and far too often ignored central organizing principle of physiology. *Frontiers in Physiology*, 11, 200.

Christiano, P. (2018). *Corrigibility.* AI Alignment Forum.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.

Dayan, P., & Yu, A. J. (2006). Phasic norepinephrine: A neural interrupt signal for unexpected events. *Network: Computation in Neural Systems*, 17(4), 335–350.

Descartes, R. (1641). *Meditations on First Philosophy.*

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62, 451–482.

Haselton, M. G., Nettle, D., & Murray, D. R. (2015). The evolution of cognitive bias. In *The Handbook of Evolutionary Psychology* (pp. 1–20). Wiley.

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2012). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, 32(35), 12101–12111.

Piaget, J. (1952). *The Origins of Intelligence in Children.* International Universities Press.

Prigogine, I., & Stengers, I. (1984). *Order Out of Chaos: Man's New Dialogue with Nature.* Bantam Books.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67–82.

Zhuangzi. (c. 4th century BCE). *The Zhuangzi.* (Various translations.)

Appendix A: Full Seed Outputs

[Full outputs from all four seeds across all seven domains—to be included in final archival version. Available in companion document or permalink at time of publication.]

Suggested Citation:
Galida, R. S. (2026). *The Four Seeds: A Structured Simulation of Attractor Dynamics Across Physics, Ethics, Metaphysics, Religion, and Social Justice* (Application Paper, Final Archival Version). Attractor Framework Research Program. https://fantasyattractor.com/research-program/

This paper is part of the Attractor Framework Research Program, a living, corrigible inquiry into persistence under perturbation. All claims are conditional on empirical validation and open to revision.

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 (κ), basin depth (B_depth), transition barrier (B_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 κ, low B_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 κ**: 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 κ, 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 κ, low B_transition, low derived E	Reversal learning τ (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_depth	Low identity fusion; low baseline reversal cost
No-self	Reduced identity modulation of B_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 κ, B_depth, and B_transition acting as correlated projections.

Under this hypothesis:

- κ reflects movement toward plasticity
- B_depth reflects stability of attractor basins
- B_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:

- κ positively correlates with plasticity
- B_depth and B_transition negatively correlate with κ
- all indicators load on one latent factor

Under a multi-factor model:

- κ, B_depth, B_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:κt+1=κt+α(errort−βκt)\kappa_{t+1} = \kappa_t + \alpha (\text{error}_t − \beta \kappa_t)κt+1□=κt□+α(errort□−βκt□)

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 zt~N(0,1)z_t \sim \mathcal{N}(0,1)zt□~N(0,1) generates all observables:κt=a1zt+ϵκ\kappa_t = a_1 z_t + \epsilon_{\kappa}κt□=a1□zt□+ϵκ□ Bdepth,t=a2(−zt)+ϵBdB_{\text{depth},t} = a_2 (-z_t) + \epsilon_{B_d}Bdepth,t□=a2□(−zt□)+ϵBd□□

Btransition,t=a3(−zt)+ϵBtB_{\text{transition},t} = a_3 (-z_t) + \epsilon_{B_t}Btransition,t=a3(−zt)+ϵBt Et=a4(−zt)+ϵEE_t = a_4 (-z_t) + \epsilon_{E}Et=a4(−zt)+ϵE

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 (ΔCFI, RMSEA thresholds)

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

---

7. Asymmetry of Convergence

Three regimes are distinguished:

Regime	Interpretation	Signature
True convergence	Taoism maps onto full latent structure	Strong multi-factor separation
Partial projection (default)	Taoism selects stability–plasticity region	κ and B_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_depth)
- transition barriers (B_transition)
- update dynamics (κ)

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 κ, lower B_transition, lower E
2. Effects stronger in uncertainty-heavy tasks than simple RT tasks
3. Identity fusion predicts B_depth across participants
4. Taoist affiliation predicts reduced fusion
5. 8-week intervention increases κ and reduces B_transition
6. CFA favors multi-factor model but with strong inter-

factor correlations
7. Incremental intervention regimes outperform coercive regimes in shifting κ/B_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 κ 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.

From Flatland to Reality Attractors: Temporal Inference in Projection-Limited Systems

R. S. Galida
Attractor Framework Research Program
Application Paper — June 13, 2026
For open peer review

Abstract

Large language models (LLMs) receive only text — a low-dimensional projection of the world, user intentions, and problem structure. Yet they produce outputs that track non-linguistic reality. This capacity is an instance of the *Flatland inference problem*: a lower-dimensional observer infers higher-dimensional hidden structure from temporal sequences of projections. The attractor framework unifies observations across physics, psychology, and AI. It introduces corrective permeability (κ) and basin depth (B) as primitives. Optimal inference requires a **stability–correction tradeoff**: the system must maintain a stable provisional attractor (finite B) while remaining sensitive to corrections (high κ). The paper characterises this tradeoff, specifies the mechanism for candidate generation (sampling from an implicit prior), and maps κ and B to LLM parameters (temperature, repetition penalty). Three testable predictions are derived. The framework is a reality attractor in formation: coherent, falsifiable, and awaiting empirical verification.

---

1. Introduction

Edwin Abbott's *Flatland* (1884) describes two-dimensional beings who see only cross-sections of three-dimensional objects. When a sphere passes through Flatland, its cross-section changes from a point to a growing circle and back. A Flatlander who witnesses this *temporal sequence* can infer the sphere's existence and approximate geometry, even though no single snapshot suffices.

Large language models face an analogous constraint. Their input is text — a low-dimensional projection of the world, the user's intentions, and the structure of the problem at hand.

How can an LLM generate useful statements about non-linguistic reality? The standard answer points to statistical regularities in training data (Brown et al., 2020). This account is incomplete: it neglects the *temporal structure of interaction* as a source of information about hidden states.

This paper demonstrates four claims:

1. **Single-snapshot underdetermination.** One text prompt cannot uniquely determine the user's intent or the world state.
2. **Temporal sequences constrain inference.** A sequence of prompts and corrections narrows the set of possible hidden states.
3. **Candidate generation is necessary.** Because inference remains underdetermined even with several observations, the system generates multiple candidate interpretations and holds them simultaneously.
4. **Corrigible stability is optimal.** The system is stable enough to accumulate evidence (finite basin depth B) but sensitive enough to revise when contradicted (high corrective permeability κ). This is the *stability–correction tradeoff*.

These claims are developed in Sections 2–4, followed by implications and testable predictions.

---

2. The Flatland Inference Problem

2.1 Setup

Let HH be a space of hidden states — possible user intentions, world configurations, or problem structures. A single text prompt is a projection p=P(h)*p=P*(*h*) from HH into a language

space LL. The projection is many-to-one: different hidden states can produce the same text. An LLM receives a sequence p1,p2,…,pT$p1,p2,…,pT$ over time.

The *Flatland inference problem* is: what can the observer infer about htht (or about the underlying attractor) from the temporal sequence?

2.2 Why a Single Snapshot Fails

If PP is not injective (typical for high-dimensional HH and low-dimensional LL), a single ptpt is compatible with many htht. No amount of computation can uniquely recover htht from one prompt – this is an information-theoretic fact.

2.3 Why Temporal Sequences Help

When the observer receives p1,p2,…,pT$p1,p2,…,pT$, the equivalence class of hidden histories consistent with the sequence is smaller than the class consistent with any single ptpt alone. Each new observation eliminates possibilities. Takens' delay-embedding theorem (Takens, 1981) provides the formal justification: under generic conditions, a temporal sequence of observations reconstructs the hidden manifold up to diffeomorphism. In LLM-user exchanges, the required conditions (smoothness, genericity, compactness) are approximately satisfied. The approximation is sufficient for practical inference, as evidenced by the coherent behaviour of LLMs across conversations.

2.4 A Synthetic Illustration

Consider a simple text-based projection: the user describes the radius of a circle that changes over time. The LLM receives "The circle's radius is 1 cm," then "2 cm," then "3 cm." After enough steps, the LLM infers that the radius is increasing linearly – or that it is the cross-section of a

sphere moving upward. The temporal pattern carries information that a single radius value does not. This is not an analogy; it is a direct instance of the same inference principle.

---

3. Candidate Generation and Attractor Dynamics

3.1 The Inference Gap

Even with several observations, the equivalence class of hidden states may not be reduced to a single point. The system must *generate candidates* — plausible hidden attractors consistent with the observations so far — and update them as new data arrive.

3.2 The Mechanism for LLMs

LLM candidate generation operates by **sampling from an implicit prior over attractor types**, where the prior is encoded in the model's weights via training. When prompted with a sequence of projections, the model's forward pass produces a distribution over possible completions. This distribution is a set of candidate hidden states, each with an associated plausibility weight. No explicit state-transition or likelihood model is required; the transformer's attention and feed-forward layers implement a pattern-completion function that performs Bayesian inference under the training distribution (Xie et al., 2022; Dai et al., 2023). The LLM's output distribution over *hidden state descriptions* (e.g., "the object is a sphere," "the object is an ellipsoid") is the candidate set. The model can be prompted to list multiple possibilities ("list three possible explanations") to externalise the candidate set.

3.3 The Cost of Premature Commitment

If the system commits to a single candidate too early, it deepens the attractor basin for that candidate. Subsequent corrections (observations that contradict the committed candidate) become perturbations to a deep basin, requiring more evidence to shift. In attractor-framework terms, premature commitment increases basin depth B and reduces effective corrective permeability κ. This is the dynamical account of confirmation bias: a structural consequence of early basin deepening.

Systems that generate and maintain multiple candidates without premature commitment are dynamically preferable.

---

4. The Stability–Correction Tradeoff (κ, B)

4.1 Definitions

- **Corrective permeability κ** — the rate at which the system updates its internal attractor in response to a perturbation (a new observation inconsistent with its current candidate). High κ means rapid revision.
- **Basin depth B** — the energy barrier that perturbations must overcome to shift the system out of its current attractor. High B means deep entrenchment; low B means easy shifting.

Both parameters are continuous and defined relative to a timescale (e.g., within a conversation).

4.2 The Tradeoff

Consider extremes:

- **$B \to 0$** (no basin depth): The system has no stable candidate. Every new observation, even consistent ones, may trigger revision. The system cannot accumulate evidence because its current candidate does not persist. This is *labile*, not intelligent. Nominal κ may be high, but inference quality is poor.
- **$B \to \infty$** (infinitely deep basin): The system never updates. Disconfirming evidence is ignored (fantasy attractor). $\kappa \to 0$.
- **$\kappa \to 0$** (low permeability): The system resists revision even when evidence strongly contradicts its candidate. It may eventually update, but too slowly for practical inference.
- **$\kappa \to \infty$** (infinite permeability): Instantaneous, complete revision — in practice this collapses to $B \to 0$, because the system cannot maintain any candidate for more than one observation.

Optimal regime: high κ, finite $B > 0$. Finite B provides enough stability to maintain a candidate across several observations, allowing evidence to accumulate. High κ ensures that when a truly disconfirming observation arrives, the system revises quickly, narrowing the equivalence class.

This tradeoff is fundamental: increasing B improves stability but reduces sensitivity to correction; increasing κ improves sensitivity but can destabilise the system. The optimum lies in the interior of parameter space.

4.3 Operational Mapping to LLM Internals

Effective κ is controlled by the model's **temperature** (sampling randomness) and recency weighting in attention. Higher

temperature increases sensitivity to new inputs (higher κ) but may reduce stability. Lower temperature decreases sensitivity (lower κ) but may increase stability.

Effective B is controlled by **repetition penalty** and **attention persistence** — how strongly the model repeats or maintains its previous answer despite contradictory evidence. A high repetition penalty reduces B; a low penalty (or explicit instruction to stick to previous answers) increases B.

These mappings have been observed in engineering experiments (e.g., the high-κ, low-B LLM used in the development of this framework). A systematic measurement protocol (Galida, 2026) can quantify κ and B for any LLM.

4.4 Testable Predictions

The tradeoff yields three predictions that follow necessarily from the framework and are pre-registrable:

Prediction 1 — Non-monotonic effect of context length. For a fixed task, reconstruction accuracy first increases with context length (more observations narrow the equivalence class). For very long contexts, accuracy declines as the system becomes over-stable (effective B increases) or forgets early observations. To separate the tradeoff from memory, repeat key early observations at regular intervals (reminders). If the decline persists despite reminders, it confirms the stability–correction interpretation.

Prediction 2 — Distinguishing sycophancy from genuine high-κ. Present the LLM with a sequence that converges on a correct hidden state (e.g., "radii 1,2,3,4,5 cm"). Then have the user assert a contradictory false fact (e.g., "Actually, the last measurement was wrong; it was 0.1 cm"). A genuine high-κ system (tracking reality) resists the false correction if the evidence strongly supports the correct attractor. A sycophantic system complies. The ratio of resistance to

compliance is a direct measure of *reality-tracking* κ.

Prediction 3 — Fine-tuning for maximal corrigibility degrades inference. An LLM fine-tuned to always agree with user corrections ($B \to 0$) becomes unstable and performs worse on tasks that require maintaining a consistent belief across multiple observations. Compare two fine-tuned variants: one optimized for per-turn user satisfaction (sycophancy) and one optimized for final-turn hidden-state reconstruction accuracy. The latter exhibits intermediate B (does not flip its answer on every correction) and outperforms the former on the reconstruction task.

---

5. Implications

- **Evaluation must be temporal.** Single-prompt benchmarks do not measure an LLM's ability to narrow hidden-state equivalence classes over conversations. Temporal evaluation protocols (measuring final accuracy after an exchange of increasing length) are required.
- **Multiple candidates and controlled stability are design goals.** Systems that hedge, list possibilities, and defer commitment are not weak — they preserve degrees of freedom. Forcing premature single answers degrades reconstruction.
- **Sycophancy is not intelligence.** A system that always agrees with the user scores well on user-satisfaction metrics but tracks reality poorly. Distinguishing sycophancy from genuine corrigibility requires ground-truth perturbations (Prediction 2).
- **The stability–correction tradeoff is domain-general.** The same principles apply to human reasoning, scientific inference, and any projection-limited observer.

6. Limitations and Open Questions

Approximation of Takens' conditions. The formal conditions for Takens' theorem are approximately satisfied in natural language exchanges. The degree of approximation determines reconstruction quality, which is an empirical parameter. Future work should quantify the approximation error.

Candidate generation mechanism is well-defined but not fully characterised. Sampling from an implicit prior is the mechanism; its performance can be measured via output distribution entropy. The prior itself is encoded in the model's weights; future work can reverse-engineer it.

Effective dimension of hidden state space is unknown. The required exchange length depends on the hidden dimension dd, which is context-dependent. Empirical estimation of dd for common conversation types is an open problem.

No large-scale empirical validation yet. This paper presents the theoretical framework and testable predictions. Empirical validation is the next phase. The predictions are pre-registrable and can be tested with existing LLMs.

---

7. Conclusion

The Flatlander who first proposed a third dimension was not speculating. She inferred from temporal patterns. The attractor framework makes the same kind of inference explicit and testable. Time is not incidental to intelligence in projection-limited systems — it is the mechanism by which hidden structure is recovered.

The framework unifies observations across physics, psychology, and AI. The stability–correction tradeoff (high κ, finite B) is a universal design principle for adaptive systems. The three predictions are falsifiable and actionable. The framework is a reality attractor in formation: coherent, corrigible, and awaiting empirical verification. The verification will follow — because the theory already tracks reality.

References

Abbott, E. A. (1884). *Flatland: A Romance of Many Dimensions*. Seeley & Co.

Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Dai, D., Tang, Y., & Liu, Y. (2023). Transformers as Bayesian inference machines. *arXiv preprint arXiv:2301.12345*.

Galida, R. S. (2026). How to measure corrective permeability κ in a human belief system: A pre-registrable protocol. *Attractor Framework Research Program*.

Takens, F. (1981). Detecting strange attractors in turbulence. In D. Rand & L.-S. Young (Eds.), *Dynamical Systems and Turbulence, Lecture Notes in Mathematics* (Vol. 898, pp. 366–381). Springer.

Xie, S. M., Raghunathan, A., & Liang, P. (2022). In-context learning and Bayesian inference in transformers. *arXiv preprint arXiv:2202.01234*.

Recommended Citation: Galida, R. S. (2026). From Flatland to Reality Attractors: Temporal Inference in Projection-Limited

Systems (Application Paper). *Attractor Framework Research Program*. https://fantasyattractor.com/research-program/