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$p_1, p_2, \ldots, p_T$ over time.

The *Flatland inference problem* is: what can the observer infer about hth_t (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 ptp_t is compatible with many hth_t. No amount of computation can uniquely recover hth_t from one prompt — this is an information-theoretic fact.

2.3 Why Temporal Sequences Help

When the observer receives p1,p2,…,pT$p_1, p_2, \ldots, p_T$, 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 → 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 → ∞** (infinitely deep basin): The system never updates. Disconfirming evidence is ignored (fantasy attractor). κ → 0.
- **κ → 0** (low permeability): The system resists revision even when evidence strongly contradicts its candidate. It may eventually update, but too slowly for practical inference.
- **κ → ∞** (infinite permeability): Instantaneous, complete

revision — in practice this collapses to B → 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 → 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 d*d* 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/