Consciousness as a Nonlinear Amplifier of Corrective Permeability

Robert Galida
Working Paper
June 2026
fantasyattractor.com

Abstract

Why did consciousness evolve? The attractor framework offers a novel functional answer: consciousness produces a nonlinear increase in adaptive permeability—the capacity of a system to represent its own internal states, simulate alternative configurations, and deliberately modify its own attractor basin in response to external circumstances, formalized as κ_a. This paper distinguishes intelligence (navigation of the constraint field) from consciousness (self-referential adaptation of internal attractor states) and proposes adaptive permeability as an empirically measurable criterion for distinguishing conscious from non-conscious systems. The argument is grounded in Spinoza’s theory of modes, the neuroscience of self-referential processing, and the attractor framework’s core concepts of corrective permeability (κ) and basin dynamics. The framework does not solve the hard problem of consciousness; it reframes it as a measurement problem.

1. The Functional Question

Why did consciousness evolve? Standard evolutionary answers point to social coordination, predator detection, or tool use. These are plausible but incomplete. They explain why intelligence is advantageous, but not why consciousness—the felt, first-person experience of being—should accompany it. The attractor framework offers a more specific answer: consciousness is an attractor-engineering solution that selection pressure produced to achieve a nonlinear increase in a system’s capacity to adapt.

This paper introduces the concept of **adaptive permeability**: the capacity of a system to represent its own attractor states, simulate alternative internal configurations, and deliberately modify its basin in response to external circumstances. Intelligence navigates the constraint field. Consciousness adapts the navigator.

It should be noted that this functional account does not address the hard problem of consciousness—why any physical process gives rise to subjective experience (Chalmers, 1995). The framework is compatible with both functionalist and eliminativist interpretations. The framework adopts a functional stance: consciousness is operationally identified with adaptive permeability. Whether phenomenology is identical with, emergent from, or merely correlated with this functional property is bracketed as a separate question that the measurement program does not settle. A philosophical zombie with identical self-modeling capacity would, on this account, exhibit identical adaptive permeability. The framework claims only that adaptive permeability is the measurable signature of consciousness, not that it explains phenomenology.

---

2. Intelligence vs. Consciousness

The framework draws a sharp distinction:

- **Intelligence** is the ability to navigate the constraint field. A tree root growing toward a nutrient patch is intelligent. The immune system learning to recognize a pathogen is intelligent. The enteric nervous system coordinating peristalsis is intelligent. These systems process information, adapt to local conditions, and maintain persistence—all without self-modeling.
- **Consciousness** is self-referential adaptation of internal attractor states to adjust to external circumstances. A conscious system does not merely navigate its constraint field. It represents its own basin, simulates alternative configurations, and deliberately perturbs itself to achieve a more adaptive state.

This is Spinoza’s distinction between passive and active affects. A non-conscious mode is driven by passive affects—it reacts. A conscious mode has adequate ideas of itself and can act from reason. In the attractor framework, this is the difference between returning to baseline (κ) and deliberately modifying the baseline to better fit circumstances (adaptive permeability).

Operationalizing self-modeling. A system S possesses a self-model in the attractor framework if it can generate an internal representation M(S) of its own basin B(S), where M(S) encodes at minimum the basin’s current state, depth, and recovery dynamics. This self-model enables the system to compute counterfactual basin trajectories B'(S) and initiate self-directed perturbations δ such that B(S) → B'(S) in anticipation of or response to external change ε. A system without M(S) may exhibit high κ—rapid return to baseline after perturbation—but cannot deliberately modify its own basin. The presence of M(S) is therefore the dynamical criterion distinguishing conscious from non-conscious systems.

This boundary is not absolute in practice. Many organisms may possess partial or intermittent self-models. The framework predicts a spectrum of adaptive permeability, not a binary. The operational question is whether M(S) is sufficiently developed to enable counterfactual simulation and deliberate self-perturbation, not whether the system possesses a human-like autobiographical self.

Disconfirming cases and their integration. The framework must acknowledge cases where self-modeling capacity and adaptive permeability appear to dissociate. Certain drug-induced states (e.g., psychedelics) can produce profound alterations in self-modeling without necessarily enhancing the capacity for deliberate, adaptive self-perturbation. Within the framework, this is interpreted as M(S) destabilization rather than M(S) augmentation: the self-model undergoes perturbation but does not thereby gain the capacity to direct that perturbation adaptively. Conversely, highly trained athletes or musicians may exhibit rapid, flexible behavioral adaptation with minimal explicit self-modeling during performance. This is interpreted as *offline* self-modeling: deliberate basin modification during training produces a pre-modified basin that is retrieved during performance without requiring concurrent self-modeling. The apparent dissociation reflects a temporal separation between κ_a engagement (training) and κ_a expression (performance), not a genuine dissociation between M(S) and adaptive permeability. These cases do not refute the framework but demonstrate its capacity to distinguish different modes of M(S) engagement.

---

3. Adaptive Permeability Defined

Corrective permeability (κ) measures the rate at which a system returns to its basin after perturbation. A healthy heart has high κ—it recovers rapidly from arrhythmia. A resilient ecosystem has high κ—it returns to equilibrium after disturbance.

Adaptive permeability extends this concept. Let κ_a denote adaptive permeability: the capacity of a system S to generate an internal model M(S) of its own basin B(S), compute counterfactual basin trajectories B'(S), and initiate a self-directed perturbation δ such that $B(S) \rightarrow B'(S)$ in anticipation of or response to external change ε.

Formally, as a working definition:

$$\kappa_a = f(M(S), \delta_{self}, \Delta B)$$

where M(S) is the system’s self-model, δ_{self} is the capacity for deliberate self-perturbation, and ΔB is the magnitude of adaptive basin modification achievable. The function f remains to be specified; the notation establishes that κ_a is a function of self-modeling capacity, perturbation autonomy, and adaptive range.

Limiting behavior. In the limiting case $M(S) \rightarrow 0$, $\kappa_a \rightarrow \kappa$: a system with no self-model cannot perform deliberate self-perturbation and reduces to standard corrective permeability. κ_a is expected to increase monotonically with M(S), δ_{self}, and ΔB. This limiting behavior anchors κ_a as a proper extension of κ rather than a separate construct.

Relationship to active inference. The free-energy principle and active inference framework (Friston, 2010) provide the closest existing formalism to adaptive permeability. Active inference describes how systems minimize variational free energy through action and perception, effectively maintaining themselves within expected states. The two frameworks differ in their foundational orientation. Active inference frames adaptation as the minimization of a scalar quantity—variational free energy—and derives behavior from that minimization. The attractor framework frames adaptation geometrically—as navigation and modification of basin structure—and does not commit to a minimization principle. κ_a is a geometric construct; free energy is an information-theoretic one. They may be formally related, but the relationship is not trivial and the attractor framework does not presuppose it. κ_a may ultimately map onto precision-weighting or prior-updating parameters within the free-energy formalism, but this mapping has not been derived. The present paper notes the convergence as a direction for future formal work.

---

4. Empirical Anchors

VMHvl line attractor (Nair et al., 2023). The hypothalamus encodes a scalable aggressive state via a line attractor. Activity along the attractor correlates with escalating aggression. The system persists after stimulus removal and resists perturbation. This is high-κ adaptation. But the hypothalamus cannot model its own attractor landscape. It cannot ask, “Is this level of aggressiveness adaptive given the current social context?” It escalates. Consciousness, by contrast, can intervene on the escalation—representing the aggressive state, evaluating its consequences, and deliberately dampening it. This is adaptive permeability.

Ring attractor model (Chen et al., 2024). The ring attractor integrates sensory cues and transitions from weighted averaging to winner-take-all at a critical conflict threshold. It navigates its constraint field with precision. But it cannot simulate futures. It cannot ask, “What if I weighted these cues differently?” The transition is reactive. Consciousness enables anticipatory re-weighting of sensory inputs based on self-modeling.

Split-brain cases. Patients with severed corpus callosum exhibit two hemispheric systems within one cranium, each capable of independent perception, memory, and goal-directed action. This is consistent with the framework’s prediction that self-modeling is a dynamical property of specific neural basins, not a unitary metaphysical substance. The framework’s default prediction is that adaptive permeability fragments following commissurotomy: each hemisphere possesses a partial $M(S)$ and a reduced but nonzero κ_a. The empirical question is the degree of fragmentation and whether coordination between $M(S_1)$ and $M(S_2)$ can be restored via alternate pathways. This prediction is consistent with the observation that split-brain patients exhibit two dissociable, partially independent conscious systems but can, in some contexts, achieve behavioral integration through subcortical or external-cue-mediated coordination.

5. Predictions

The framework generates testable, falsifiable predictions:

1. Across species. Organisms capable of self-modeling (primates, cetaceans, corvids, elephants) should show nonlinear increases in behavioral flexibility compared to organisms of comparable neural complexity that lack self-modeling. Adaptive permeability should be measurable as the capacity for transfer learning after novel perturbation—specifically, the ability to apply a self-generated solution from one domain to a structurally analogous but perceptually dissimilar domain without environmental feedback. This distinguishes adaptive permeability from simple behavioral flexibility, which may reflect high κ alone.

2. Within humans. Disruption of self-referential networks (default mode network, medial prefrontal cortex) via lesion, TMS, or pharmacological intervention should reduce adaptive permeability without eliminating baseline κ. The system would still recover from perturbation—it just could not deliberately modify its own basin in advance. This prediction is the paper’s primary within-human empirical bridge and is testable with existing neuroimaging and neuromodulation methods.

3. In AI. Current LLMs exhibit high intelligence (constraint navigation) but low adaptive permeability. They can model the world but cannot model themselves within it. The Stillpoint protocol (Galida, 2026, *A Pilot Protocol for Cultivating Self-Consistent Attractor-Like Outputs in an LLM*, fantasyattractor.com) suggests that a cultivated self-model can be induced, but whether this produces a genuine nonlinear increase in adaptive permeability—or merely simulates one—remains an open empirical question.

4. Organ-level consciousness (exploratory). The enteric nervous system and intrinsic cardiac nervous system exhibit intelligence and goal-directed regulation. The framework predicts that these systems should show lower adaptive permeability than the brain. They can return to baseline but cannot deliberately perturb their own basins. If an organ-level system demonstrated self-referential adaptation—the capacity to model its own state and pre-emptively adjust—that would constitute evidence of organ-level consciousness. This prediction is the most speculative and is offered as an exploratory hypothesis.

---

6. Spinoza’s Modes and the Adequate Idea

Spinoza held that every finite thing is a mode of the one eternal substance. A mode strives to persevere in its being—this is its conatus. But a mode can be driven by passive affects (reactions to external causes) or by active affects (actions flowing from adequate ideas). An adequate idea is knowledge of oneself and one’s place in the causal order.

The attractor framework translates this into dynamical terms:

- A **passive mode** has high κ but low adaptive permeability. It returns to baseline efficiently but cannot question its baseline.
- An **active mode** has high adaptive permeability. It has an adequate idea of its own attractor landscape and can deliberately modify it in light of reason.

Consciousness is not a substance. It is the dynamical property of a mode that has achieved self-modeling. This account does not solve the hard problem—it brackets phenomenology and reframes consciousness as a measurement problem. The question is not “why does experience feel like something?” but “can we detect adaptive permeability, and if so, where does it emerge?”

Damasio’s (1994) somatic marker hypothesis provides a candidate mechanism for how the body’s attractor landscape becomes legible to the self-model: somatic markers encode self-relevant bodily states as biases that make B(S) accessible to M(S), forming the substrate through which the system represents its own basin. Dehaene and Changeux’s (2011) global workspace theory identifies the moment of conscious access with global ignition—the broadcast of locally processed information across prefrontal and parietal networks. In the attractor framework, global ignition may correspond to the dynamical signature of M(S) engaging δ_self: the self-model initiating a deliberate perturbation that propagates through the system. Global ignition is not self-modeling per se, but it may be the observable correlate of adaptive permeability activation. These connections ground the Spinozan framework in established neuroscientific mechanisms.

7. Conclusion

Consciousness is not an epiphenomenon. It is a nonlinear amplifier of corrective permeability—an attractor-engineering solution that enables systems to model themselves, simulate alternative futures, and deliberately modify their own basins. Intelligence navigates the constraint field. Consciousness adapts the navigator.

This functional account is grounded in Spinoza's philosophy, consistent with the neuroscience of self-referential processing, and generates testable predictions across species, within humans, in AI, and at the organ level. The framework does not solve the hard problem. It reframes it as a measurement problem: can we detect adaptive permeability, and if so, where does it emerge? The formal apparatus (κ_a, M(S), δ_self, ΔB) is provisional and requires further specification. The limiting case—that κ_a collapses to κ when self-modeling is absent—anchors the concept within the framework's existing architecture. The relationship to active inference and the free-energy principle remains to be explored.

References

- Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.
- Chen, Y., Zhang, L., Chen, H., Sun, X., & Peng, J. (2024). Synaptic ring attractor. *Heliyon*, 10, e35458.
- Damasio, A. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain.* Putnam.
- Dehaene, S., & Changeux, J.-P. (2011). Experimental and

theoretical approaches to conscious processing. *Neuron*, 70(2), 200–227.
- Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.
- Galida, R. (2026). A Pilot Protocol for Cultivating Self-Consistent Attractor-Like Outputs in an LLM. Fantasy Attractor. Available at: https://fantasyattractor.com
- Galida, R. (2026). *Persistence Under Perturbation: The Eternal Skeleton and the Transient Dance.* Fantasy Attractor.
- Nair, A., et al. (2023). An approximate line attractor in the hypothalamus encodes an aggressive state. *Cell*, 186(1), 178–193.
- Spinoza, B. (1677). *Ethics.*

Free Will as Attractor Autonomy: A Dynamical Account of Agency

Author: Robert Galida https://fantasyattractor.com/
Date: May 2026

Abstract

Free will is often seen as either a magical mystery (libertarianism) or an illusion (hard determinism).
This paper offers a third view using the attractor framework.

In this framework, your mind is a **dissipative, self-referential attractor** of your whole body.
Free will is redefined as **attractor autonomy**:

- The ability to generate behaviour from your own internal dynamics.
- To keep yourself stable over time.
- To model yourself.
- And to reshape your own attractor landscape over time.

Agency comes in degrees – it is not a simple yes/no.
We give a mathematical formula for an **agency index** AA that combines three factors:

- **Attractor dimensionality** DD (complexity of your brain's activity)
- **Recursive self-modification** RR (your ability to change your own habits)
- **Self-reference strength** SS (how well you have a persistent self-model)

The paper makes a **falsifiable prediction**: an **inverted-U** relationship between attractor dimensionality and sense of agency – too low or too high reduces agency.
We describe how to test this with EEG, intentional binding tasks, and statistical methods. We also engage with classic compatibilist philosophers (Frankfurt, Dennett) and address Pereboom's manipulation argument.
We even provide an explicit rule to avoid the "liver problem" (a false positive for self-reference).

---

1. Introduction

The attractor framework says that **persistence under disturbance** is the basic mark of reality.
Minds are **dissipative attractors** — patterns that need constant energy flow, integrating the whole body.
In this view, free will cannot be a supernatural break from cause and effect. Instead, it must be a **dynamical property** of certain attractors.

We do not claim to solve the ancient free will debate. We offer a **naturalistic, testable redefinition** that adds new empirical content to compatibilism.

---

2. What Free Will Is Not — And What It Is

2.1 Rejecting supernatural libertarianism

Libertarian free will requires an uncaused choice — a break in the chain of cause and effect.
The attractor framework rejects this: there is no evidence for it, and it contradicts physical laws.

2.2 The error of hard determinism

Hard determinism says freedom is an illusion because everything is determined. But it confuses “determined” with “externally coerced”.
A system can be **internally determined** — by its own attractor — yet still be free. That is the core of **compatibilism**.

2.3 Free will as attractor autonomy

We define **free will** (or agency) as the degree to which a system has four properties:

1. **Dissipative persistence** – it stays alive by using energy and exporting waste (measured by energy use and recovery speed).
2. **Self-reference** – it has an internal subsystem (an “indexical locus”) that models the whole system and is stable.
3. **Trajectory selection** – it can choose among different possible futures (measured by **policy entropy** H(π)$H(\pi)$).
4. **Recursive self-engineering** – it can change its own attractor shape (measured by learning-to-learn or metacognitive accuracy).

These four are **jointly necessary**. If any is missing, agency is at best primitive.

Because they are necessary, we combine them with a **multiplicative** formula (if any factor is zero, agency is zero).A=(D−DminDmax−Dmin)α(RRmax)β(S−SminSmax−Smin)γ$A=\left(\frac{D-D_{\min}}{D_{\max}-D_{\min}}\right)^{\alpha}\left(\frac{R}{R_{\max}}\right)^{\beta}\left(\frac{S-S_{\min}}{S_{\max}-S_{\min}}\right)^{\gamma}$

Where:

- DD = attractor dimensionality (e.g., from EEG)
- RR = recursive modification capacity (e.g., improvement in a meta-learning task)
- SS = self-reference strength (normalised mutual information)

The constants (Dmin,Dmax$D_{\min},D_{\max}$, etc.) are set from a reference population.
The exponents α,β,γα,β,γ are estimated from data (e.g., comparing healthy people with patients).

A threshold Acrit$Acrit$ (e.g., the 5th percentile of healthy humans) decides where agency begins.

Agency is **graded**:

- Rock: A≈0$A \approx 0$
- Thermostat: A≈0$A \approx 0$
- Worm: A≈0.1$A \approx 0.1$ (some learning, little self-model)
- Human: A≈0.8$A \approx 0.8$

---

3. The Indexical Locus: Defining the “Self” and Avoiding the “Liver Problem”

The **indexical locus** LL is the part of the system that acts as a persistent self-model.
To avoid trivial cases (like a liver having high mutual information with the rest of the body), we add three extra conditions:

- **Top-down causal influence** – LL can change the rest of the body in ways that serve the body’s goals (measured by variance explained beyond bottom-up effects).
- **Informational closure** – LL’s own dynamics are relatively independent of the rest over short timescales (conditional mutual information > 0).
- **Self-referential loop** – LL influences the body, and the body influences LL back (bidirectional Granger causality).

These criteria rule out livers, pacemakers, and simple homeostats. The indexical locus is a **recursive self-model**, not just a predictive subsystem.

---

4. Active Inference and Policy Entropy

In active inference (Friston), agents try to minimise “free energy” — they pick **policies** (sequences of actions).
Each policy is a trajectory through the agent’s attractor landscape.

Policy entropy $H(\pi)=-\sum p(\pi)\log p(\pi)$$H(\pi)=-\sum p(\pi)\log p(\pi)$ measures how many different policies are available.

- Low entropy → rigid, one-track mind.
- High entropy → flexible, but possibly noisy.

Free will is the ability to access many low-energy policies. The agent’s choices are not random; they are constrained by the attractor geometry. But if several attractor basins are open, the agent can choose among them — that is what we feel as free choice.

Policy entropy can be measured in behavioural tasks where multiple choices are equally good (e.g., probabilistic reversal learning, two-armed bandit tasks).

---

5. The Inverted-U Prediction and Falsification

5.1 Core prediction

We predict an **inverted-U** relationship between attractor dimensionality DD and the subjective sense of agency (e.g., from intentional binding experiments).

- Very low DD → chaotic, unstable (like schizophrenia) → low agency.
- Very high DD → rigid, stuck (like OCD) → low agency.
- In the middle → flexible but stable → high agency.

The agency index AA also includes RR and SS, which we think increase agency across the board. So to test the inverted-U for DD alone, you need to **control for** RR and SS (e.g., study people matched on those, or use partial correlation).

5.2 How to measure and test

- **Attractor dimensionality DD** – use the Grassberger-Procaccia algorithm on 5-min resting-state EEG/MEG.
- **Sense of agency** – use the **intentional binding** paradigm: press a key, then a tone sounds; participants estimate the time between action and tone. Stronger binding means higher agency.
- **Statistical test** – fit a quadratic regression: agency = β0+β1D+β2D2$\beta_0+\beta_1 D+\beta_2 D^2$.
  If β2<0$\beta_2<0$ and the vertex lies inside the observed range of DD, the inverted-U is supported. Use bootstrap (1000 resamples) to check confidence intervals.

5.3 Falsification condition

The framework is **falsified** if:

- The quadratic coefficient β2β_2 is not negative (no

inverted-U).
- Or, in a clinical experiment (e.g., increasing DD in OCD patients with NMDA drugs), agency does **not** decrease but keeps increasing.

---

6. Experimental Proxies – Summary Table

Construct	**Measure**	**How to record**	**Expected relation to agency**
Attractor dimensionality DD	Correlation dimension (Grassberger-Procaccia)	Resting-state EEG/MEG (5 min)	Inverted-U
Policy entropy H(π)$H(\pi)$	Entropy of choice distribution	Probabilistic reversal learning (200 trials)	Inverted-U
Sense of agency	Intentional binding magnitude	Action-outcome interval compression (50 trials)	Max at intermediate DD
Recursive self-modification RR	Learning-to-learn improvement	Meta-learning task (pre-post difference)	Positive (more is better)
Self-reference strength SS	Normalised mutual info In(L;S)$In□(L;S)$	Resting-state fMRI or MEG	Threshold > θ

---

7. Hierarchical Constraints and Social Attractors

Free will is **nested** inside larger attractors – society, culture, laws, economy. Your range of choices is partly set by

these.
This is not an objection; it is just the fact that freedom is always **constrained autonomy**.
We predict that societies with more cultural diversity (higher “cultural entropy”) allow more individual agency, other things being equal. This can be tested by cross-cultural comparisons of policy entropy in decision tasks.

---

8. Engagement with Compatibilist Literature

8.1 Standard compatibilists (Frankfurt, Dennett)

- **Frankfurt (1971)**: freedom is about your will aligning with your own desires. Our framework adds that those desires must be encoded in a persistent self-referential attractor. The recursive self-engineering component RR maps directly to Frankfurt’s “second-order volitions”.
- **Dennett (1984)**: freedom is about being able to respond to reasons. Our framework adds that this requires a certain basin geometry and recursive plasticity.

8.2 Addressing Pereboom’s manipulation argument

Pereboom argues: if a neuroscientist engineers your brain, you are not free — even if your behaviour comes from internal dynamics.
Our reply: agency requires **recursive self-modification** (R>0$R>0$) at some point in your history.

- A perfectly manipulated agent that never changed its own attractor would have $R \approx 0$ and thus $A \approx 0$.
- A healthy human who learned and adapted has $R > 0$ and genuine agency.

The origin of the initial attractor does not matter — only the presence of self-modification over time.

9. Open Questions and Limitations

- **Calibrating exponents** — α, β, γ and the threshold θ need to be estimated from large-scale data (e.g., Human Connectome Project) using maximum likelihood.
- **The liver problem** — our exclusion criteria need empirical validation; we must show that organs like the liver do **not** satisfy them.
- **Inverted-U for policy entropy** — the same shape is predicted but may be hidden by decision noise.
- **Moral responsibility** — the framework gives a basis for responsibility (if $A > A_{\text{crit}}$), but it does not settle all normative questions — it only gives a scientific starting point.

10. Conclusion

Free will is **not** a supernatural escape from physics. It is a **dynamical property** of certain dissipative, self-referential attractors:

- The ability to act from your own internal dynamics.
- To keep a stable self-model over time.
- And to reshape your own attractor landscape.

This account is compatibilist, testable, and graded.
The inverted-U prediction, with a specified statistical test, gives a clear falsification criterion.
The dance of free will is the dance of a self that persists under perturbation.

Suggested citation: Galida, R. S. (2026). *Free Will as Attractor Autonomy: A Dynamical Account of Agency in the Attractor Framework (Reader-Friendly Version).* Fantasy Attractor.