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** $H(\pi)$).
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−Dmin⁡Dmax⁡−Dmin⁡)α(RRmax⁡)β(S−Smin⁡Smax⁡−Smin⁡)γ*A*=(*D*max⁡−*D*min⁡*D*−*D*min⁡⁡)*α*(*R*max⁡*R*⁡)*β*(*S*max⁡−*S*min⁡*S*−*S*min⁡⁡)*γ*

Where:

- D*D* = attractor dimensionality (e.g., from EEG)
- R*R* = recursive modification capacity (e.g., improvement in a meta-learning task)
- S*S* = 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,γ*α*,*β*,*γ* are estimated from data (e.g., comparing healthy people with patients).
A threshold Acrit*A*crit⁡ (e.g., the 5th percentile of healthy humans) decides where agency begins.

Agency is **graded**:

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

---

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

The **indexical locus** L*L* 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 $H(\pi) = -\sum p(\pi) \log p(\pi)$ 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 D*D* and the subjective sense of agency (e.g., from intentional binding experiments).

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

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

5.2 How to measure and test

- **Attractor dimensionality D*D*** – 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

Construct	Measure	How to record	Expected relation to agency
Sense of agency	Intentional binding magnitude	Action-outcome interval compression (50 trials)	Max at intermediate D*D*
Recursive self-modification R*R*	Learning-to-learn improvement	Meta-learning task (pre-post difference)	Positive (more is better)
Self-reference strength S*S*	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 R*R* 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≈0*R*≈0 and thus A≈0*A*≈0.
- A healthy human who learned and adapted has R>0*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*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>Acrit$A > A_{crit}$), 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.

Attractor Dynamics in Belief Formation, Correction, and Mental Health: A Research Programme

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

Abstract

This paper applies the attractor framework (persistence under disturbance) to **belief systems** and **mental health**.

We introduce three measurable concepts:

- **Attractor depth** – how rigid or unstable a belief is.
- **Error half-life** – how long it takes for a false belief to fade after correction.
- **Coupling strength to error signals** – how open a belief is to reality checks.

We contrast two disorders:

- **OCD** (obsessive-compulsive disorder) may involve *overly deep* (rigid) attractors.
- **Schizophrenia** may involve *too shallow* (unstable) attractors – with appropriate caution.

We propose experiments to measure error half-life, detect early warning signs of belief shifts (while managing false alarms), and find the optimal pace for correction ("critical damping").

We also outline:

- **N=1 attractor engineering** (self-experimentation)
- **Wearable early-warning systems** for relapse prevention (discussing lag time and false positives)
- **Cross-coupling** as a measure of resilience (distinguishing healthy from brittle coupling)

This paper is a **research roadmap**, not a finished theory.

---

1. Introduction

In the attractor framework, your mind is a **dissipative attractor of your whole body** – a pattern that needs energy, can be disturbed, and can adapt (Galida, 2026, *Persistence Under Perturbation*).
Beliefs are smaller attractors inside that landscape. Their stability determines how easily you update when faced with contradictory evidence.

This paper turns attractor concepts into testable ideas about how beliefs form, stick, and change – and how to help them change. It is a roadmap, not the final word.

2. Attractor Depth and Mental Disorders

Neurocomputational models suggest a contrast between OCD and schizophrenia, but we must be careful.

Disorder	Attractor Property	Behavioural Sign	Example Task
OCD	Too deep (rigid)	Stuck, hard to switch	Reversal learning (changing rules)
Schizophrenia	Too shallow (unstable)	Jumpy, over-sensitive to noise	Delayed match-to-sample with distractions

Evidence:

- Unmedicated OCD patients make many perseverative errors on reversal-learning tasks; this correlates with symptom severity (Remijnse et al., 2006).
- Reduced NMDA/GABA function in schizophrenia makes attractor networks unstable, leading to cognitive slips and delusions (Rolls, 2021).

Caveats:

- Mental disorders are complex, with multiple attractors. We are talking about symptom clusters, not whole-disorder diagnoses.
- Disorders like anxiety, depression, and personality disorders lie in the middle – their attractors are **domain-specific** (e.g., depression has deep negative-belief basins but shallow positive ones).

Prediction: Attractor depth could be measured from behaviour (switching rates, reaction time variability) by fitting a two-state hidden Markov model to reversal-learning data — a hypothesis for future work.

---

3. Error Half-Life: A New Measure of Belief Rigidity

Error half-life T1/2$T_{1/2}$ is the time it takes for a false belief's confidence to drop by half after you present corrective evidence.

How to measure it

1. Give people a false belief (e.g., a made-up fact).
2. Give them correct information (text, video) every day for a while.
3. Ask them to rate their belief confidence (0–100) at intervals.
4. Assume a simple **exponential decay** model C(t)=C0e−t/τ$C(t) = C_0 e^{-t/\tau}$ as a starting point (real decay could be sigmoidal or power-law).
5. Then T1/2=τln⁡2$T_{1/2} = \tau \ln 2$.

What we expect in different conditions

- **Delusional disorders** → very long half-life (deep attractor).
- **Depression** → long half-life for negative self-beliefs, but normal for positive ones (asymmetric updating).
- **Anxiety** → short half-life, but possible overshoot (shallow basin → oscillation).

Therapeutic application

The goal is to **shorten error half-life**. Methods like **spaced repetition** and **active recall** (quizzing) could help – they strengthen corrective memory traces, similar to memory reconsolidation.

Relationship to attractor depth

Attractor depth is a **static** measure (inertia). Error half-life is a **dynamic** measure (recovery speed). They are related but not the same: depth gives initial resistance, half-life gives the time course. We need both.

---

4. Critical Slowing Down Before Belief Shifts

Before a sudden change of belief (e.g., leaving a cult, political conversion, therapy breakthrough), you may see **early warning signals** – rising variance, higher autocorrelation, slower recovery from small disturbances. This is called **critical slowing down** (Scheffer et al., 2009).

How to detect it

- Collect daily belief ratings, mood scores, or social media sentiment.
- Compute rolling variance and autocorrelation with a moving window.
- If they exceed a baseline threshold, a shift may be coming.

False positive problem

Rising variance can be caused by other things (seasonal mood, life events). To reduce false alarms:

- Use control periods (compare with a stable trait belief).
- Combine multiple signals (HRV, sleep, activity) with self-report.
- Use a conservative threshold (e.g., 3 standard deviations above baseline).

This is a research tool, not a clinical diagnostic yet.

Prediction: You can detect these signals in diaries before a person deconverts, changes politics, or relapses into depression. A well-timed prompt might help, but false positives must be managed.

---

5. Optimal Correction Dosing (Critical Damping)

From control theory, there is an **optimal pace** for delivering corrections: not too slow (oscillates), not too fast (overshoot/backfire). This is called **critical damping**.

N=1 protocol

- Vary the gap between corrections (massed vs. spaced).
- Track belief confidence over time.
- Measure how quickly and smoothly it changes.

Hypothesis: Spaced correction (e.g., daily micro-doses) works

better than one big confrontation – a well-known finding in memory research (Ebbinghaus, spaced repetition). The twist is applying it to **beliefs**, which are more emotional and identity-linked. The mechanism may be similar, but emotional valence may change the optimal schedule.

---

6. Fantasy vs. Shared Reality Attractors – Operational Metrics

Metric	Low Corrective Permeability (Fantasy)	High Corrective Permeability (Shared Reality)
Coupling to error signals	Low (few fact-checks, no update)	High (active correction)
Basin depth	Deep (needs large evidence)	Shallow (small anomalies work)
Error-correction latency	Long (days/weeks)	Short (hours/days)
Information diversity tolerated	Low (echo chamber)	High (multiple sources)

Double-bind computational model

In conspiracy cultures, contradictory evidence gets reinterpreted as confirmation ("cover-up"). We can model this as an **asymmetric Bayesian update**:P(belief∣contrary evidence)≥P(belief∣supporting evidence)*P*(belief∣contrary evidence)≥*P*(belief∣supporting evidence)

Example: Start with belief probability 0.9. A contrary piece of evidence that would normally lower it to 0.3 is instead interpreted as evidence of suppression, so the new probability

stays at 0.85. The belief drifts only slowly.

Breaking the loop: Indirect interventions work better than direct refutation:

- Point out internal inconsistencies.
- Seed doubt through trusted messengers.
- Use graduated reality-testing.

---

7. Wearable Early Warning of Attractor Shifts

Protocol: Use consumer wearables (HRV, skin conductance, actigraphy, sleep) plus daily self-reports (mood, belief rigidity). Compute rolling variance and autocorrelation in real time.

Evidence: Drops in nocturnal HRV preceded a depressive relapse in a case study (Tonge et al., 2024).
Prediction: Rising variance/autocorrelation in HRV, plus mood volatility, can predict an imminent crisis.

Latency and false alarms

- Useful lead time is **days**, not hours. HRV changes can appear 1–2 weeks before relapse.
- False positives are a concern. Use a **two-stage alert**: first detect statistical anomaly, then confirm with a brief self-report (EMA).
- Specificity needs to be established in longitudinal N=1 studies.

Intervention: When thresholds are crossed, trigger a

micro-intervention (mindfulness, therapist call) – a closed-loop prevention system.

---

8. N=1 Attractor Engineering – Minimal Perturbation Protocol

Goal: Find the smallest intervention that shifts a maladaptive attractor (phobia, obsessive thought) without causing oscillation or backfire.

Procedure:

1. Define the target (e.g., fear rating 0–10).
2. Start with very low-intensity perturbations (e.g., brief exposure, mild counter-evidence).
3. Measure change after each step.
4. When a threshold shift is detected (say, 30% reduction – a provisional starting point; adjust based on baseline variability), record the dose.
5. Back off slightly and check stability.

Principle: Never collapse an attractor faster than reality can correct. Use fine steps (5–10% increments) and frequent monitoring. This is **precision self-regulation**. Generalisability from N=1 to populations is an open question (see Section 12).

---

9. Cross-Coupling as a Resilience

Metric

Hypothesis: High cross-domain coupling (e.g., HRV ↔ mood ↔ sleep) indicates **adaptive resilience** – the system is coordinated and self-correcting. Low coupling or unidirectional cascades indicate **brittle coupling** (a disturbance in one area spreads uncontrollably).

Measurement: Collect simultaneous time series (HRV, sleep, activity, mood). Compute cross-correlation or Granger causality.

- **Adaptive** = bidirectional, with negative feedback (e.g., poor sleep → lower HRV → mood drop → social support → sleep improves).
- **Brittle** = unidirectional, amplifying (e.g., sleep loss → stress → more sleep loss).

Prediction: Good recovery from stress shows strong bidirectional influences. Low coupling or unidirectional cascades will precede breakdowns.

Intervention: Improve adaptive coupling with synchrony exercises (e.g., daily breathing with light exposure, yoga, social rhythm therapy). Testable in an N=1 self-tracking experiment.

---

10. Philosophical Extensions (Brief)

- **Are attractors real?** Yes, as structural patterns (process metaphysics). They have causal power – like the path of a river.

- **Free will as attractor autonomy** – acting according to your own attractor is compatibilist freedom. Our framework adds that freedom is about basin width and flexibility, not a binary.
- **Cosmic attractor** – speculative. The universe might have a global attractor (e.g., heat death), but it's untestable now.
- **Darwinian problem of evil** – animal suffering is a strong challenge to theism; the "deep harmonies" hypothesis is hard to falsify.

---

11. Open Questions and Next Steps

- Can error half-life be measured reliably from smartphone-based belief tracking? What decay model fits best?
- What is the dose-response curve for corrective interventions? Linear, exponential, or threshold? How does it vary with attractor depth?
- Can wearables detect early warning signs before a psychiatric relapse? What are the false-positive rates and lead times?
- Does adaptive cross-coupling improve after synchrony-based therapies?
- How are error half-life and attractor depth related? Same thing at different timescales, or different constructs?
- How can N=1 findings be aggregated into population-level knowledge? One approach: meta-analysis of single-subject time series using hierarchical Bayesian models.

12. Conclusion

This research programme puts attractor dynamics to work on beliefs and mental health.

We have proposed **testable metrics** (attractor depth, error half-life, coupling strength) and **experimental protocols** for N=1 self-engineering and early warning.

The framework provides a naturalistic language for understanding why some beliefs resist correction and how to intervene optimally.

We acknowledge our limitations – the exponential decay assumption, false positives in early warning, and the generalisability of N=1 results – and treat them as open questions for future work.

This extends the attractor trilogy into **actionable health and epistemology**.

Suggested citation: Galida, R. S. (2026). *Attractor Dynamics in Belief Formation, Correction, and Mental Health: A Research Programme (Reader-Friendly Version).* Fantasy Attractor.