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.