Vitality as an Information-Theoretic Criterion for Lifelikeness (Version 2)

December 16, 2025

"Yes, you're alive. But are you truly living?"

Most formal approaches to "life-likeness" begin from *viability*: living systems strive to keep themselves within regimes that sustain their ongoing existence—homeostasis, survivable ranges, adaptive fitness under natural selection. (Source 1: watch youtube video) This sits comfortably with a Darwinian picture in which organisms in the heat of struggle are primarily in the business of not dying.

But this perspective quietly undermines many of our deepest intuitions about what it means to be alive. It says little about how it feels to be alive, or why so much of living behavior—play, exploration, art, gratuitous complexity—seems to exceed mere survival. A sedated patient on life support is viable. A remote-controlled robot can be kept "safe" indefinitely. All of these satisfy viability in a thin sense, yet none fully match our intuition of living.

In this white paper we propose *Vitality* as an information-theoretic refinement of lifelikeness. We distinguish:

- **Viability** as a purely set-based property: a system remains within a region of state space in which it continues to be "the same system".
- **Vitality** as its dual: given that viable region, how the system organizes its sensorimotor degrees of freedom *over time*.

In this sense, Viability tells us where a system can remain to remain being a system. Vitality tells us whether, within that region, it is truly living.

Connection to Version 1. Version 1 defined Vitality as a static bidirectional score derived from empowerment and plasticity under an NTIC gate. Version 2 keeps the same primitives (viability, Abel-style empowerment/plasticity, NTIC), but sharpens the mechanism: empowerment and plasticity are treated as *signals* that can be *internalized* into slow temporal morphogens $(\mathcal{E}_t, \mathcal{P}_t)$, giving Vitality a genuinely dynamical phenotype and a first-person timescale. NTIC is retained, but only as a **third-person sanity check** that excludes caves/puppets.

1 Introduction: From Viability to Vitality

- Problem: "Staying alive" (viability) is not enough to capture *lifelikeness*.
- Key idea:
  - **Viability**: where the system may remain to remain *being that system*.
  - **Vitality**: how the system uses its viable degrees of freedom *over time*.
- Aim: give a **rigorous, information-theoretic criterion for lifelikeness**.
- Core contribution (Version 2):
  - Lifelike agents are those that are viable and exhibit a **high-Vitality temporal regime** in which interface empowerment and plasticity are internalized into slow morphogens $(\mathcal{E}_t, \mathcal{P}_t)$ that regulate the fast loop, and the resulting dynamics pass an **NTIC audit** (non-trivial closure), excluding simulator/puppet solutions.

2 Formal Setup

- State space $\mathcal{X}$; joint state $X_t \in \mathcal{X}$.
- Partition into:
  - Internal state Y_t,
  - World state W_t (we use W to avoid collision with empowerment E).
- Interface λ:
  - Action process $A_{1:T}$,
  - Observation process $O_{1:T}$.
- Policies / closed-loop dynamics π inducing trajectory distributions.

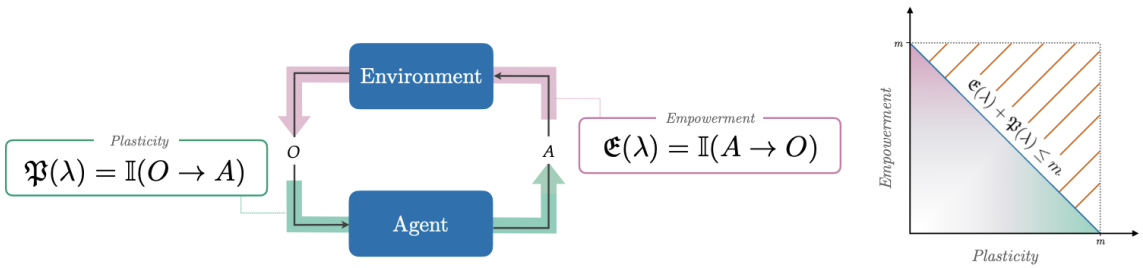
We keep this section short and notational.

3 Viability: Purely Set-Based

- **Definition (Viability kernel).**
  - $K \subseteq \mathcal{X}$: set of states in which the system still "is itself".
- **Definition (Finite-horizon viability).**
  - $\mathrm{Viab}_T(\pi) := \Pr_\pi[X_t \in K \ \forall t \leq T]$.
- Emphasize:
  - Viability is **purely set-based** and constrains **where** trajectories may remain.
  - It does not care how behavior is organized: babysitting, puppetry, or autonomy all look identical if they stay inside K.

4 Interface Capacities: Empowerment and Plasticity (Abel)

- At interface λ (time-local / horizon-local estimates):
  - **Empowerment**
    $$E_t(\lambda) := I(A \rightsquigarrow O; \lambda)_t$$
    capacity for actions to influence near-future observations.
  - **Plasticity**
    $$P_t(\lambda) := I(O \rightsquigarrow A; \lambda)_t$$
    capacity for observations to reshape near-future actions.
- Abel-style tradeoff geometry:
  - There exists $m_t(\lambda)$ with
    $$0 \le E_t(\lambda),\, P_t(\lambda) \le m_t(\lambda), \quad E_t(\lambda) + P_t(\lambda) \le m_t(\lambda).$$



(a) Plasticity and Empowerment[(b) Theorem 4.8

Figure 1: Interface capacities $E_t(\lambda)$ (empowerment) and $P_t(\lambda)$ (plasticity) as mirror quantities (left), and the tight upper bound on their joint values (right). Values of $E_t(\lambda)$ and $P_t(\lambda)$ from zero up to an interface- and interval-dependent constant $m_t(\lambda)$ are realizable, but their sum can be no greater than $m_t(\lambda)$, illustrating Abel's bidirectional tradeoff geometry.

- Interpretation:
  - Defines the **bidirectional capacity geometry** at the interface.
  - Empowerment-only principles constrain only one axis.

5 Internal Time: Temporal Morphogens of Empowerment and Plasticity

Version 2 introduces one minimal internalization step: the interface capacities become slow internal morphogens.

5.1 Internalization (definition)

- Define temporal morphogens as slow internalizations:

$$\mathcal{E}_{t+1} = (1 - \alpha_E)\mathcal{E}_t + \alpha_E\, E_t(\lambda), \qquad \mathcal{P}_{t+1} = (1 - \alpha_P)\mathcal{P}_t + \alpha_P\, P_t(\lambda).$$

- Intuition:
  - $\mathcal{E}_t$: consolidated empowerment (internal "commitment/skill" regulator).
  - $\mathcal{P}_t$: consolidated plasticity (internal "reconfigurability" regulator).
- We do not replace Abel's quantities; we **internalize** them.

5.2 First-person time (explicit)

- Third-person time indexes events by t.
- First-person time is carried by internal regulatory state: the agent's "present" is its current morphogen configuration $(\mathcal{E}_t, \mathcal{P}_t)$, which persists across moments and shapes how the next moment is met.
- The intrinsic integration horizons are set by α_E, α_P: smaller α implies longer persistence (a longer lived timescale).

5.3 Fast-loop modulation (no new state variables)

- The morphogens modulate the fast loop (control + learning) by gating stability vs rewrite:

$$\text{rewrite/adaptation strength at time } t \propto f(\mathcal{P}_t), \qquad \text{stability/precision at time } t \propto g(\mathcal{E}_t).$$

5.4 Minimal asymmetry (slow-to-recruit, fast-to-act)

- To capture "slow-to-recruit but fast-to-act" without adding states:
  - assume $\alpha_P \ll \alpha_E$ (plasticity internalization is slower),
  - but $f(\mathcal{P}_t)$ can act strongly when $\mathcal{P}_t$ is high.
- This supports plateau–burst alternations in time (an "animated" temporal pattern) without requiring a separate dynamical system.

6 Dynamic Vitality

6.1 Conceptual Definition

- **Viability**: "You remain in K so you remain *a system.*"
- **Vitality**: "Given that you remain in K, you organize your degrees of freedom so that:
  - you strongly influence future inputs (high $\mathcal{E}_t$),
  - you remain strongly reconfigurable by those inputs (high $\mathcal{P}_t$),
  - and you sustain this bidirectionality as a temporal regime."

6.2 Formal Vitality Measure (time-local)

- **Definition (time-local Vitality at interface λ).**

$$V_t(\lambda) := \frac{2\,\mathcal{E}_t\,\mathcal{P}_t}{\mathcal{E}_t + \mathcal{P}_t}.$$

- Properties:
  - Requires **both** $\mathcal{E}_t$ and $\mathcal{P}_t$ (bidirectionality).
  - Symmetric in $\mathcal{E}_t$ and $\mathcal{P}_t$, monotone in each.
  - Penalizes extremes (rigidity or rewrite-soup): high Vitality favors sustained *bidirectional* coupling.

6.3 Vitality Over a Horizon

- **Definition (horizon-averaged Vitality).**

$$\bar{V}_T(\lambda) := \frac{1}{T}\sum_{t=1}^{T} V_t(\lambda).$$

- Vitality is not a one-shot score; it is the characteristic time-pattern of sustained bidirectional coupling.

7 Non-Trivial Information Closure (NTIC) as a Third-Person Audit

NTIC is retained to exclude trivial solutions (caves/puppets), but Version 2 is explicit: NTIC is generally **not directly computable by the organism** (it lacks a god's-eye view of W_t and the "true boundary"). Therefore, NTIC is treated as an **observer-level sanity check** on the resulting dynamics, not as a term the agent must explicitly optimize.

7.1 Information Flow (diagnostic ingredient)

- Information flow into the internal state:

$$J(W \to Y)_t := I(W_t; Y_{t+1} \mid Y_t).$$

- Low J can indicate closure, but closure can be trivial; hence the need for NTIC.

7.2 NTIC Index (auditable, not necessarily endogenous)

- Introduce a scalar diagnostic:

$$\mathrm{NTIC}_t(\lambda) \in [0, 1],$$

  interpreted as:
  - $\mathrm{NTIC}_t \approx 0$: trivial closure (isolation, puppetry, simulator caves),
  - $\mathrm{NTIC}_t \approx 1$: non-trivial closure (internal dynamics carry world-structure in a causally efficacious way).

7.3 Remark: Exogenous vs Endogenous Adaptation (fairness to empowerment)

- Many empowerment-based systems *do* update their controllers over time; however, this often constitutes **exogenous adaptation**—learning dynamics specified and scheduled from outside the agent (optimizer choice, learning-rate schedules, training curricula, external stopping criteria). In such cases, adaptation is something that *happens to* the system.
- Version 2 instead emphasizes **endogenous, internalized plasticity**: a slow morphogen $\mathcal{P}_t$ that the agent carries as part of its state and that actively gates **when** and **how strongly** the system reconfigures itself. In this sense, $\mathcal{P}_t$ is not merely "learning," but **active adaptation**—a self-regulated alternation between consolidation and rewrite.

7.4 Proposition: Internalized Plasticity is Required for an NTIC-Asymptote

- **Proposition (internalized plasticity required for NTIC-asymptote).** In generic non-stationary settings (or when the initial internal organization is mis-specified), empowerment-driven control with only exogenous adaptation can sustain or even maximize control while failing to show a reliable upward tendency in NTIC under third-person audit. By contrast, systems possessing **internalized plasticity** $\mathcal{P}_t > 0$—especially when recruited intermittently as a slow-to-recruit / fast-to-act morphogen—can *construct* internal organization over developmental time and thereby exhibit an NTIC tendency (increasing or sustained-high NTIC under audit).
- This frames $\mathcal{P}$ as the "construction lever" for autonomy: it is what allows closure to be earned rather than assumed.
- **Corollary (empowerment-only pathologies).** Empowerment-only optimization admits attractors compatible with *trivial* closure: stable caves, rigid sensorimotor loops, and other regimes that score high on control but do not internalize world-structure into causally efficacious internal dynamics. Therefore, high empowerment alone does not imply non-trivial closure.

7.5 NTIC as a Tendency (criterion, not objective)

- When we say NTIC is a **tendency**, we mean a property of the observed trajectory/regime (audited externally), e.g.:

$$\mathbb{E}[\mathrm{NTIC}_{t+1}] \gtrsim \mathbb{E}[\mathrm{NTIC}_t] \quad \text{over developmental time.}$$

- Or, more minimally:

$$\frac{1}{T}\sum_{t=1}^{T} \mathrm{NTIC}_t(\lambda) \geq \tau \quad \text{for some threshold } \tau > 0.$$

- Either way, NTIC is not "part of the instantaneous Vitality function"; it is a **regime-level audit**.

8 Co-Regulation of Empowerment and Plasticity (Dynamic Regime)

8.1 Empowerment Regulated by Plasticity

- In non-stationary environments:

  - Without sufficient reconfiguration capacity, empowered coupling collapses or retreats into trivial niches.

- **Claim**: sustained high $\mathcal{E}_t$ over time typically requires intermittent recruitment of $\mathcal{P}_t$.

8.2 Plasticity Regulated by Empowerment

- Without empowered coupling:
  - plasticity becomes **overwrite** by the world (puppetry-like dependence).
- With empowered coupling:
  - plasticity can become **self-directed restructuring** rather than overwrite.
- **Claim**: non-degenerate $\mathcal{P}_t$ typically requires positive $\mathcal{E}_t$.

8.3 Lifelikeness (informal definition, crisp)

- A system is lifelike over horizon T iff:
  - it is viable (stays in K),
  - it operates in a high-$\bar{V}_T$ regime,
  - and it passes a non-triviality audit via NTIC (e.g., average NTIC above threshold and/or improving tendency).

9 Pathologies and Exclusions by Construction

- **Simulator caves / bright dark rooms**: may show strong internal regularities and even high control, but fail the NTIC audit (trivial closure).
- **Puppet embodiments**: action–observation influence may be present at the body level, but internal closure is externally driven; NTIC audit fails.
- **Frozen agents**: $\mathcal{P}_t \to 0 \Rightarrow V_t(\lambda) \to 0$ (loss of reconfigurability).
- **Rewrite-soup agents**: $\mathcal{E}_t \to 0 \Rightarrow V_t(\lambda) \to 0$ (loss of stable empowered coupling).
- These illustrate why **viability + empowerment** are strictly weaker than **Vitality + NTIC audit**.

10 Discussion

- What changed from Version 1 (minimally):
  - Vitality remains built from empowerment and plasticity via the same harmonic-mean intuition.
  - The only added structure is internalization: (E_t, P_t) at the interface become morphogens $(\mathcal{E}_t, \mathcal{P}_t)$ that constitute first-person time and regulate the fast loop.
  - NTIC remains essential, but is clarified as an **observer-level audit**, not an agent-side computed term.

  - Version 2 adds a testable novelty: **internalized plasticity enables an NTIC-asymptote**, while empowerment-only optimization admits trivial-closure pathologies.

- Why it matters:
  - preserves the parsimony needed for formalization,
  - captures vitality as a temporal phenotype rather than a static number,
  - keeps the "no caves / no puppets" exclusion principle without bloating the agent's internal machinery.