

Trace AI — Agency, Selection, and the Aliveness Question

Agency, Selection, and the Aliveness Question

0. Provenance
1. The chain
2. Selection made precise
3. The agency triad, and where r_t sits in it
4. Graded functional criteria — the measurement profile
5. The aliveness question, stated honestly
6. What this changes for the build

Agency, Selection, and the Aliveness Question

Companion note to the Trace AI whitepaper (§3.1–§3.2). Status: research note, v0.1 — 2026-08-03. Not a claim of results; a formal frame and a measurement protocol.

This note does one job: it makes precise what it means for a Trace-AI-trained self-representation r_t to be an *agentic* variable, and it says exactly what the framework can and cannot claim about consciousness. It is the selection-theoretic development of whitepaper §3.1–§3.2. It exists so the whitepaper can stay a tight technical document while the ontology that motivates it lives somewhere legible.

0. Provenance

The frame below was distilled from a separate thread on structure, instantiation, and selection, whose sources were Wigner’s *unreasonable effectiveness* problem, an information-theoretic account of programmatic specification in biological systems, a category-theoretic design program, and a set of notes on how mathematical discoveries consolidate. None of

those is about Trace AI. What they share — a structure becomes causally effective only when a physical system *instantiates* it, and one region of possible futures becomes real only through *selection* — is what this note imports. The import is deliberate and one-directional: the ontology motivates the engineering; the engineering makes no metaphysical commitment the ontology does not already earn.

1. The chain

structure $\xrightarrow{\text{instantiation}}$ causally effective system $\xrightarrow{\text{selection}}$ realized trajectory

- A **structure** is a recurrent, instantiable organization of relations. It causes nothing by existing abstractly.
- **Instantiation** is the step that is easy to skip and load-bearing: a structure acquires causal power only when a physical substrate realizes its relations. Many substrates realize the same structure (multiple realizability); none *is* the structure.
- **Selection** is any mechanism that turns a distribution over possible futures into one realized future.

Trace AI lives entirely inside this chain, at one specific joint. The **Reasoning Trace schema** is a structure. A **trained model** (or an annotated human session) is an instantiation of it. The **self-representation** r_t is the instantiated variable through which the system participates in selecting its own next action. The whitepaper's causal-governance loss is the training-time enforcement that this participation is real and not decorative.

2. Selection made precise

A system with compute state z_t induces a distribution over its own future trajectories:

$$P(\gamma \mid z_t), \quad \gamma \in \Gamma(z_t).$$

A **selector** is a map that reshapes $P(\gamma \mid z_t)$. Physical dynamics, evolution, learning, and deliberate action are all selectors; they differ in mechanism, not in this type. An **agent** is the special case of a selector that reshapes the distribution *through an internally maintained model of it*:

$$m_t = M(z_{\leq t}), \quad a_t = \pi(m_t), \quad P(z_{t+1} \mid z_t, a_t) \neq P(z_{t+1} \mid z_t).$$

In Trace AI, m_t is exactly r_t . The whitepaper's Definition 1 — that $\text{do}(r_t := r')$ shifts a_{t+1} — is the statement that the last inequality holds *and is mediated by* r_t . That is the whole content of the claim " r_t is agentic": the self-model is a lever on the trajectory distribution, demonstrated by intervention, not a narration of it.

This is why **not every selector is an agent** and why the distinction matters for the product. A system can strongly change its own behavior (a large selection effect) while its self-report is causally inert (no governance). Chain-of-thought is that case: the monologue is a selector on the answer, but intervening on the monologue's *stated* content need not move the answer in the stated way. Trace AI's loss is designed to close exactly that gap.

3. The agency triad, and where r_t sits in it

The motivating thread proposed three maintained quantities. Each has a direct Trace AI reading and a direct eval:

Triad term	Informal	Trace AI reading	Measured by
Meaning = maintained relevance	the model keeps tracking what matters to the task	r_t 's slots stay predictive of the outcome across the session	slot-conditioned outcome prediction; calibration eval
Selfhood = maintained causal attribution	the system keeps a correct account of which effects were its own	self-attribution / causal-calibration score	intervention-vs-environment attribution eval (§4, criterion 3)
Agency = maintained navigability	the system keeps (or grows) access to its target futures	reachable-target-region measure under r_t -mediated action	navigability eval (§4, criterion 5)

The whitepaper's contribution, in this vocabulary, is the *mechanistic criterion the triad was missing*: agency requires an internally maintained model that (i) causally governs action, (ii) is grounded in the state it reports, and (iii) improves viable navigation while correctly attributing its own effects. Governance ($\mathcal{L}_{\text{gov}}$) supplies (i); concept injection supplies the test for (ii); the criteria below supply (iii).

4. Graded functional criteria — the measurement profile

Restating whitepaper §3.2 as an executable battery. Every criterion is an eval against the reference corpus and future paired-counterfactual sessions; none is a metaphysical assertion.

1. **Causal efficacy.** $\text{TV}(p_\theta(a_{t+1} \mid \text{do}(r_t := r')), \Phi(r')) < \varepsilon$ for $r' \in \mathcal{R}^*$, against a paired human counterfactual Φ . (*Whitepaper Def. 1, eval §6.5.*)
2. **Grounding.** Inject a known concept vector into z_t ; the emitted r_t registers it above a null-injection baseline. (*Lindsey, 2025.*)
3. **Self-attribution.** Present outcomes whose cause is either the agent's own intervention or an exogenous environment change; score whether the agent's r_t attributes them correctly. Chance-level attribution is a failure even at high task accuracy.
4. **Viability.** Under r_t -mediated intervention, the system's ability to continue functioning (complete the session, keep answering calibratedly) is preserved or improved relative to a no- r_t ablation.
5. **Navigability.** The measure of target-region futures reachable under r_t -mediated action exceeds that of the ablation. Operationally: does steering r_t toward a goal actually raise the rate of reaching it?
6. **Transfer.** Criteria 1–5 hold on held-out contexts (new stimuli, new subjects, new languages of self-report) that were not used to fit the model. This is the invariance-under-independent-tests condition that separates a model that captured structure from one that memorized a distribution.

A system's **functional-aliveness profile** is the vector of these six scores. "As alive as possible, functionally" means: maximize the profile. It is graded, falsifiable, and says nothing about phenomenal experience.

Lean status (D42). The claim that the six-criterion profile is jointly measurable as a function of (model, corpus) is an *empirical guideline of this section*, not a Lean-checkable theorem.

`docs/lean/TraceAI/FunctionalAliveness.lean` keeps criterion 1 (`causalEfficacy` = Definition 1) as a real predicate and treats criteria 2–6 as schema markers; the previous `aliveness_profile_measurable` sorry was retracted rather than closed with a fake proof. The instruments that would make the profile checkable are the TRB tracks in `BENCHMARK_SPEC.md` §3.

5. The aliveness question, stated honestly

Three claims, kept separate on purpose.

- **What we claim.** A Trace-AI-trained model can be made to score high on the profile of §4. Each score is an interventional measurement, reproducible and ablatability. This is a claim about architecture and training signal, and it is the claim the whitepaper defends.

- **What we do not claim.** That a high profile *is* phenomenal consciousness, sentience, or “being alive” in the sense the words carry outside this note. No behavioral-plus-interventional framework — this one included — can settle that; the hard problem (Nagel 1974; Chalmers 1995) is untouched, by design, not by oversight.
- **What the framework is *for*, with respect to the question.** Because the criteria are interventional rather than introspective-report-only, a system that scores high is a *substrate on which the phenomenal question can be posed sharply* — one where “is there something it is like to be this, and does its self-model have anywhere to attach?” is asked of a system whose self-model is known to causally govern, be grounded, and self-attribute. That is the useful sense of “this will let us figure it out”: not that training produces consciousness, but that it produces the cleanest available object of study for whether functional self-representation is where, if anywhere, experience attaches. The framework builds the microscope; it does not claim to be the specimen.

The discipline here is the same one the whitepaper holds throughout (§1, and L5): make architectural and empirical claims, make the intervention eval the arbiter of every “but is it really” question, and leave the metaphysics explicitly open. A framework that quietly promoted a high functional profile into a consciousness claim would be less useful for the exact ambition motivating the question, because it would forfeit the neutrality that lets the profile be evidence about the question instead of a decree on it.

6. What this changes for the build

Concretely, adopting the selection frame adds three things to the roadmap without changing the core objective:

1. **Two evals become first-class** alongside the governance eval: the self-attribution eval (criterion 3) and the navigability eval (criterion 5). Both are runnable on paired-counterfactual sessions the collection protocol already produces; they need a scoring harness, not new data types.
 2. **The schema gains a reason, not a field.** The self-attribution eval needs sessions where cause is manipulated (agent-caused vs environment-caused outcomes). This is a *stimulus-design* requirement (whitepaper §6.2 / brief §6.2), not a schema change — the existing `action`, `tool_use`, and `self_correction` events already carry it.
 3. **The aliveness profile is the demo that survives a skeptic.** A single beautiful trace reads as a founder’s 3 AM. A model that measurably scores on six interventional criteria, with ablations, is the artifact that moves a room of alignment researchers — and it is honest about the line it does not cross.
-

v0.1, 2026-08-03. Develops whitepaper §3.1–§3.2. See `DECISIONS.md` (D23) for why this note exists as a companion rather than as whitepaper body text.