

Trace AI — Structural Intelligence Conjecture (research program)

The Structural Intelligence Conjecture and the Structural Observatory

1. The conjecture
2. The master object: a stochastic fibration with a compiler
3. Concern as a geometry on the fibre
4. Ten constructs the master object generates
5. The three built instruments
 - Instrument 1 — `experiments/representation_search` (Fiber Finder)
 - Instrument 2 — `experiments/structure_compiler` (one invariant, many embodiments)
 - Instrument 3 — `experiments/symbolic_causation` (agency science)
6. Capstone application: conscious / reliable agents (the honest version)
7. The umbrella: a Structural Observatory
8. Limitations and rejected alternatives
9. One-line synthesis

The Structural Intelligence Conjecture and the Structural Observatory

Date: 2026-08-03. **Status:** research program + three working instruments. The conjecture and the extended formalism (§2–§4) are speculative theory, not established results; they are held to a strict mathematical-claim gate and are **not** promoted to theorems. The three instruments in §5 are exact, deterministic, and tested — they establish *dissociations on solvable toy cases*, not the general claims.

Companion (formal, four theorems + six instruments):

`SIC_MATHEMATICAL_FOUNDATIONS.md` — the version of this program that turned three of the extensions below into theorems (existence via minimal sufficiency; cross-task stability under a shared Markov screen; discrete learnability via common-sufficient clustering) and added three

more instruments. **Companion (plain-English):** [STRUCTURAL_INTELLIGENCE.md](#) — the same object described with the Trace AI lens.

Both companion docs exist in this repo for a reason: this note is the *working* version — cautious, extension-labeled, non-theorem — and the formal foundations doc is what it *grew into* under the mathematical-claim gate. Keeping both prevents the honesty of this document from being retroactively rewritten into the confidence of the other.

Editorial note: this document references source works ([AGENTS.md](#) , [mathematical-claim gate](#), [notes/latent_structures_meta_framework.md](#) , the repo's [symbolic_weakness](#) result, [experiments/common/causal_use.py](#) , [experiments/world_responds](#) , [results/](#)) that live in the paper's *source repository* — [jawauntb/research-derived-experiments](#) — rather than in the Trace AI repo. Their function here is provenance — the paths are preserved so a reader can locate the underlying material at that repository — not a Trace-AI-side path claim. The originating research note lives at [notes/structural_intelligence_conjecture.md](#) .

1. The conjecture

Structural Intelligence Conjecture. A finite adaptive system's central capability is not predicting states or choosing actions but discovering a quotient $q: X \rightarrow Z$ such that (1) the task-relevant dynamics descend to Z , (2) irrelevant variation is confined to the fibres $q^{-1}(z)$, (3) useful interventions can be specified compactly in Z , (4) those specifications can be re-instantiated through a compiler, and (5) their consequences remain stable across substrates and contexts.

In one line: **intelligence finds the level at which the world becomes both compressible and controllable.** This is the operational form of the Manifest-Invariant Principle: find the coordinate in which the hidden invariant is a coordinate axis.

2. The master object: a stochastic fibration with a compiler

The common object across the corpus is not a latent space Z by itself. It is a *stochastic fibration* together with an abstraction–realization pair:

$$X \xrightarrow{q} Z, \quad Z \xrightarrow{K} X, \quad \text{supp} K(\cdot \mid z) \subseteq q^{-1}(z).$$

- X — the space of concrete realizations;
- Z — the space of structures / functions / observables;
- q — says which concrete differences do not matter (coarse-graining);
- $q^{-1}(z)$ — the **fibre** of possible embodiments of structure z ;
- $K(dx | z)$ — a **compiler**: a distribution over realizations inside the fibre.

This generalizes the meta-framework's $R \dashv C$. There, realize R and coarse-grain C were (near-)deterministic functors. Here $q = C$ and the compiler K is a *stochastic section* of R : it does not pick one realization but a distribution over the fibre. The biology paper (Kiiskinen–Kivinen–Rivas) is exactly this: the genome names z , the physics substrate is the compiler K that “computes the samples,” and “selection acts on the statistics of the resulting ensemble.” Its threshold theorem is the statement that below a critical coarse-graining C^* the fibre $q^{-1}(z)$ is too large to address with the specification budget — the fibre is real and irreducible.

The stochastic-fibration formulation is the director's synthesis; **no single source states it in this form**. Category theory (source A) supplies the compositional implementation/verification/compilation maps; Wigner (D) supplies invariance under irrelevant conditions plus the warning that theories may be local and non-unique; the structural-realism thread (F) says structures become causally effective only through embodiment and approach truth when preserved relations survive independently-defined contexts.

A note on rigor (the director's correction). Cross-domain resemblance is **not** isomorphism. The honest hierarchy of “sameness,” strongest to weakest: isomorphism $\supset$ bisimulation $\supset$ functor $\supset$ natural transformation $\supset$ adjunction / Galois connection $\supset$ Morita-like equivalence $\supset$ simulation-at-a-resolution. Most relations among these works are **adjunctions, simulations, and shared diagram shapes**, not object-level isomorphisms. Every proposed connection must answer: *exactly what kind of sameness is this, what does the map forget, and what would have to be proved to make the analogy a theorem?*

3. Concern as a geometry on the fibre

A compiler gives a baseline distribution over embodiments $K(dx | z)$. A concern state c need not change which realizations are *possible*; it changes which are salient, viable, reachable. Model it as a reweighting:

$$K_c(dx | z) = \frac{e^{\beta U_c(x, z)} K(dx | z)}{\int e^{\beta U_c(x', z)} K(dx' | z)}$$

where U_c is the **concern field** — it makes some regions of the fibre more consequential. This induces an information geometry on concern states, e.g. the Fisher metric $g_{ij}(c) = \mathbb{E}_{K_c}[\partial_i \log K_c \cdot \partial_j \log K_c]$, measuring how distinguishable nearby concern configurations are *by their effect on realized possibilities*.

Precisely:

- **meaning** is a deformation of relevance over possibilities, not a representation;
- **concern** is a field/measure that changes the effective fibre geometry;
- **agency** is the capacity to alter the kernel K or the field U_c so viable regions become reachable;
- **selfhood** concerns which changes to the kernel are attributed to the system's own interventions.

This is the exact continuation of the source repository ([jawauntb/research-derived-experiments](#))'s *Geometry of Concern* and *Gauge-Fixed Transport of Concern*: concern transport between contexts becomes transport between kernels over corresponding fibres, and **holonomy** measures the failure of transported concern to return unchanged around a loop of contexts. It is now fittable from interventions and trajectories, not merely a metaphor.

4. Ten constructs the master object generates

A research program, ranked by fertility, practical use, and beauty. Each is stated as a target, with its status and its relation to the instruments in §5.

1. **Concern geometry** (§3) — fit K_c , its Fisher metric, and concern holonomy from interventions. *Fertile; theory + fit*.
2. **Conditional rate–distortion control limit**. Robust control of $Y = f(X)$ from a coarse spec $Z = q(X)$ requires $H(Y | Z) \approx 0$, or, under tolerated distortion D , extra addressed bits $B_{\text{extra}} \geq R_{Y|Z}(D)$. *Why micromanagement fails: the controller tries to distinguish outcomes its channel cannot address*. This generalizes the biology theorem; **conjectural, not yet proved**.
3. **Abstraction frontier**. Replace “find the best Z ” with the Pareto frontier of representations trading task-sufficiency $I(Y; X | q(X))$, dynamical closure $I(Z_{t+1}; X_t | Z_t, A_t)$, cost $H_0(Z)$, and control regret. Explains why two representations can both be “right” yet incomparable (an antichain, as in the biology paper's threshold). *Seeded by Instrument 1*.
4. **Fibre audit (adversarial)**. Instead of testing held-out prediction, vary the allegedly irrelevant degrees of freedom while holding q fixed and measure the interventional

discrepancy $\Delta_q(z) = \sup_{x, x' \in q^{-1}(z)} d(P(Y | \text{do}x), P(Y | \text{do}x'))$. Large $\Delta_q \Rightarrow$ the abstraction collapsed a causally important distinction. *The operational form of Wigner's relevant/irrelevant split and F's truth criterion. Seeded (non-interventional core) by Instrument 1; the interventional version is next.*

5. **Theory atlas (sheaf/stack gluing).** Treat theories as local charts M_i on contexts U_i with translations T_{ij} ; test the cocycle $T_{jk} \circ T_{ij} = T_{ik}$. Where it holds, the charts glue; where it fails, the obstruction is informative (missing latent, scale transition, phase boundary, category error). *Wigner's non-unification worry, made constructive.*
6. **Compiler tomography.** Given many (s_i, x_i) with $x_i \sim K(\cdot | s_i)$, infer the shared compiler and compact specs by MDL: $\min_{K, \{s_i\}} [L(K) + \sum L(s_i) - \sum \log p_K(x_i | s_i)]$. Variation across i is specification; shared regularity is the compiler; residual is unresolved state/randomness. Then **compiler ecology**: $K_{t+1} = U(K_t, \text{outcomes})$ — build a compiler under which good outcomes are cheap, rather than re-specifying good outcomes. *Formal language for education, institutions, long-horizon training.*
7. **Causal semantics.** Two symbols are equivalent when they induce naturally equivalent update operators $\Psi_{m,c}$ across independent contexts. Groups sentences by how they reorganize reachable possibility, not textual co-occurrence. *An operational meaning layer ordinary embeddings omit; extends Instrument 3.*
8. **Representation-repair calculus.** A library of failure signatures $\rightarrow$ minimal structural lifts (scalar loses multiplicity $\rightarrow$ lift to operator/PSD; global norm loses location $\rightarrow$ localize to a measure/sheaf; quotient hides degeneracy $\rightarrow$ restore the fibre/stabilizer; affine mishandles infinity $\rightarrow$ projectivize; exact target exceeds capacity $\rightarrow$ move to ensemble control; symmetry breaks identifiability $\rightarrow$ gauge-fix / moduli). Turns "try another approach" into *diagnose the lost invariant, then apply the minimal lift.*
9. **Alignment as ensemble governance** (not trajectory scripting). A finite spec cannot address every fine trajectory of a long-horizon agent; target a viable region $V \subseteq Z$ with $\Pr[q(X_t) \in V \forall t] \geq 1 - \delta$ under a broad family of unresolved compiler/environment states. A fibre audit becomes the alignment evaluation. **Conjectural generalization, flagged as such.** *Extends Instrument 3 and the source repo's causally-grounded-agents line.*
10. **Autocatalytic artwork.** A work $S_t \rightarrow E_t \xrightarrow[\text{experience}]{K_t} K_{t+1}$ whose early movements teach the grammar by which later movements become legible — an autocatalytic symbolic structure that produces part of the machinery required for its own fuller instantiation, across harmony / shader / navigation / language / social interaction. *The literal reading of F's closing proposition; extends Instrument 2.*

5. The three built instruments

Each is exact, deterministic, tested, and public-safe. They are the first three modules of the Observatory and establish the dissociations the program rests on.

Instrument 1 — `experiments/representation_search` (Fiber Finder)

Over a Boolean world with a known invariant, it enumerates a lattice of quotients and three selectors. **Result:** only `minimal_sufficient` (sufficient, then minimal description length) recovers the ground-truth invariant; `mdl_only` collapses the obstruction; `accuracy_only` never compresses. **Establishes** that sufficiency, description length, and accuracy dissociate — the discrete, non-interventional core of the fibre audit (extension 4) and the seed of the abstraction frontier (extension 3). Connects to the source repository ([jawauntb/research-derived-experiments](https://github.com/jawauntb/research-derived-experiments))'s `symbolic_weakness` result: weakness is a counit/fibre quantity, MDL is a spec-length quantity; they are not the same, and the fibre quantity is the one that governs generalization.

Instrument 2 — `experiments/structure_compiler` (one invariant, many embodiments)

An abstract automaton with accumulation $\rightarrow$ phase-transition $\rightarrow$ hysteresis is compiled into music, a visual field, text, and spatial navigation; each medium's readback recovers the *same* abstract trajectory (verified structural identity, fidelity 1.0). **Establishes** cross-substrate structural identity as a checkable property ($q_i \circ F_i = \text{id}$), not mood matching — the ensemble-compiler and Gesamtkunstwerk idea, and the base case for the autocatalytic work (extension 10).

Instrument 3 — `experiments/symbolic_causation` (agency science)

An exact finite-state world treats a symbolic model m as an operation on the future-trajectory distribution and separates **signal** (Δ_{KL}), **control** (`goal_gain`), **knowledge** (`predictive_accuracy`), and **agency** (control + calibrated self-attribution + transfer). Seven conditions each realize a distinct metric signature; a `false_credit` condition improves the outcome while its true do-effect is zero and its self-attribution is miscalibrated; a brittle controller controls but does not transfer. **Establishes** that no single scalar ("behavioural influence") identifies agency — the measurement core for concern geometry (extension 1), causal semantics (7), and ensemble-governance alignment (9). Connects to `experiments/common/causal_use.py` and `experiments/world_responds`.

Cross-reference to Trace AI. Instrument 3 is the exact-solvable formalization of what TRB Track 3 (self-attribution) measures on trained models. The `false_credit` condition is the ground truth for Track 3's balanced-accuracy score; see [BENCHMARK_SPEC.md](#) §3 Track 3.

6. Capstone application: conscious / reliable agents (the honest version)

The program bears directly on the two questions above — can we make agents conscious, and can we make them never wrong. The disciplined answer is *not in the literal sense of either*, but the framework gives a precise, non-inflated construction target. It factors as two nested systems:

Concerned Self-Modeling Core $\subset$ Proof-Carrying Reliability Shell.

Inner core = an active, self-maintaining T -algebra. A persistent agent that maintains a world model and self-model, represents concern-weighted futures (§3), globally broadcasts selected information, remembers commitments, predicts the consequences of its actions, performs **false-credit tests on its own causal claims** (exactly Instrument 3's calibration metric), and reports uncertainty and internal conflict. In the framework this is precisely the passive $\rightarrow$ active threshold: a T -algebra that runs K and q in a closed loop and optimizes its own counit gap online. It operationalizes proposed consciousness *indicators* (global availability, metacognition, self-modeling, agency) — and that is the ceiling of the claim. **Functional selfhood $\nRightarrow$ subjective consciousness:** satisfying architectural/behavioural indicators does not establish that there is something it is like to be the agent, and there is a real welfare/ethics wrinkle in building systems that strongly satisfy such indicators.

Outer shell = the fibration with a verifier on the counit. A system that converts goals into contracts (compiler tomography / spec compilation), separates observation from inference, attaches provenance to every claim, and commits an action only under machine-checkable evidence: $\text{execute}(a) \Leftrightarrow V(s, a, \varphi) = \text{PASS}$, else abstain / ask / simulate / escalate. Its correctness envelope is a fibre audit (extension 4): vary everything the spec claims is irrelevant and check the property survives. **Verified-in-a-bounded-domain $\nRightarrow$ universal infallibility:** verification certifies the *formal statement*, not that the statement captures what a human wanted, and reliability also depends on model, harness, tools, environment, and budget.

So the honest breakthrough available now is not conscious, perfect agents. It is **agents with experimentally measurable selfhood and formally bounded error** — functional selfhood $\nRightarrow$ consciousness, and verified bounded behavior $\nRightarrow$ infallibility, kept as two separate, non-inflated claims.

7. The umbrella: a Structural Observatory

One system that, given a problem, theory, agent, organism, or artwork, returns: its candidate quotient maps, realization fibres, automorphisms, abstraction frontier, fibre-audit results, inferred compiler, and possible cross-substrate embodiments. The three instruments above are its first three modules (quotient search, cross-substrate compiler, causal/agency measurement); §4 lists the remaining modules as pre-registered future instruments.

8. Limitations and rejected alternatives

- **The instruments are toys by design.** They establish dissociations on exactly-solvable cases (finite Boolean worlds, tiny MDPs, lossless encoders). They do **not** establish the general conjecture, the conditional rate–distortion limit (extension 2), or the alignment claim (extension 9), all of which remain conjectural. Instrument 2’s fidelity is 1.0 *by construction*: it tests that the compilers are faithful functors, not that faithful compilation is hard.
 - **The master object is posited.** The stochastic fibration (q, K) organizes the corpus; that the sources’ relations are genuine adjunctions/simulations rather than looser analogies is exactly what §2’s sameness question demands be proved case by case. Do not cite the fibration as if it were any source’s theorem — only the biology paper proves a specialized version (its C^* threshold and “no fourth category” accounting).
 - **Rejected — collapse to one scalar.** *“Just measure behavioural influence / compression / prediction.”* Rejected: Instruments 1 and 3 exhibit exactly the cases where a single scalar misidentifies the target (MDL collapses the obstruction; false-credit influence is uncaused). The dissociations are the content.
 - **Consciousness caution retained.** Nothing here licenses a claim of machine phenomenology; the core is a testable functional model of selfhood, and the gap to subjective experience is left open on purpose.
-

9. One-line synthesis

The object is a stochastic fibration $X \xrightarrow{q} Z$ with a compiler $K: Z \rightsquigarrow X$ filling the fibres; concern is a reweighting of the fibre, agency is the licensed alteration of the kernel, truth is fibre-invariance under independent tests, and intelligence is the search for the quotient where the world is at once compressible and controllable — of which conscious selfhood is one measurable functional face and formally bounded error is the other, with neither collapsible into infallibility or phenomenology.

v0.1, 2026-08-03. See `DECISIONS.md` (D29) for provenance. This document is the research-program-status sibling of `SIC_MATHEMATICAL_FOUNDATIONS.md` ; when they disagree, that disagreement is intentional — this one is the working note, the other is the formalized paper.