

Trace AI — Structural Intelligence Conjecture (formal paper)

The Structural Intelligence Conjecture

Representation as a stochastic fibration, with nine exact instruments

Abstract

1. The master object
2. Derivations
 - 2.1 Existence via minimal sufficiency (Theorem 1)
 - 2.2 Rate–distortion parameterisation (Theorem 2)
 - 2.3 Categorical restatement (Proposition 3)
 - 2.4 Cross-task stability (Theorem 4, conditional)
 - 2.5 Learnability, discrete case (Theorem 5)
 - 2.5b Continuous-case learnability at resolution ε (Theorem 6)
 - 2.5c Positive resolution of SIC-C-c for linear ICA (Theorem 7)
 - 2.6 SIC in the honest split
3. The conjecture (full-strength form)
4. Nine exact instruments
 - 4.1 Fiber Finder — witness of Theorem 1
 - 4.2 Structure compiler
 - 4.3 Agency benchmark — the seed of TRB Track 3
 - 4.4 Cross-task sufficiency — witness of Theorem 4
 - 4.5 Cross-task learnability — witness of Theorem 5
 - 4.6 Cross-task learnability, continuous — witness of Theorem 6
 - 4.7 Rate–distortion pair — witness of Theorem 2
(`experiments/rate_distortion_pair`)
 - 4.8 Linear-ICA learnability — witness of Theorem 7
(`experiments/linear_ica_learnability`)
 - 4.9 Sparse-ICA (IMA) learnability — second SIC-C-c resolution
(`experiments/sparse_ica_learnability`)
 - 2.5d Concern as fibre geometry — companion paper (Theorems CG-1, CG-2)
5. The extended program (ten further constructs)
6. Capstone: conscious and reliable agents, honestly
7. Limitations

The Structural Intelligence Conjecture

Representation as a stochastic fibration, with nine exact instruments

Jawaun Brown — human author and research director. **Claude Code (agent)** — experiment code, analysis, and manuscript production under direction and review. **Date:** 2026-08-03 (v4 + Observatory Instrument 9, integrated 2026-08-03). **Status: five theorems + one conditional theorem + nine exact executable instruments.** Existence, cross-task stability, discrete learnability, and continuous learnability at resolution ε derived (Theorems 1–6); Theorem 2 (rate–distortion) witnessed exactly by Instrument 7; **Theorem 7 (linear-ICA identifiability) resolves SIC-C-c positively inside the linear-ICA hypothesis class (Instrument 8) and the sparse-mechanism / IMA class (Gresele et al. 2021) gives a second positive resolution (Instrument 9)** — two distinct inductive-bias classes now provably earn escape from Theorem 6's ε -covering exponential-in- d_Z bound. What remains genuinely open is only the general programme: which *other* inductive-bias classes admit their own polynomial-in- d_Z theorems (iVAE with auxiliary variables, interventional CRL, ...) — a mainstream question in identifiable-representation-learning theory, addressed one class at a time.

Live interactive Observatory: structural-observatory-production.up.railway.app — all nine instruments rendered as interactive panels with committed numerical results.

Companion PDFs: [sic_paper.pdf](#) is the v4 (latest); prior versions preserved as [sic_paper_v2.pdf](#), [sic_paper_v3.pdf](#), [sic_paper_v4.pdf](#). **Reads on:** [STRUCTURAL_INTELLIGENCE.md](#) (plain-English pointer to the same object with the Trace AI lens), [WHITEPAPER.md](#) (the training-signal application), [BENCHMARK_SPEC.md](#) (TRB Track 3 is Instrument 3 of this paper), [AGENCY_AND_SELECTION.md](#) (the selection frame those instruments live in).

Why it exists here. The [STRUCTURAL_INTELLIGENCE.md](#) note tells the story in plain English; that told-in-plain-English claim needed a paper with theorems and exact-solvable witnesses under it, so the note doesn't have to carry weight it can't. This is that paper.

Abstract

We study a single latent object that recurs across five otherwise unrelated works — a category-theoretic framework for materials design, an information-theoretic limit on the programmatic specification of biological systems, a corpus of machine-found mathematical proofs with their discovery notes, Wigner’s essay on the unreasonable effectiveness of mathematics, and a structural-realist ontology. The common object is a **stochastic fibration with a compiler**: a coarse-graining $q: X \rightarrow Z$ together with a kernel $K: Z \rightsquigarrow X$ whose support lies in the fibre $q^{-1}(z)$. We derive the object — it is not posited: for any statistical task on a standard Borel space, the minimal sufficient σ -algebra of Halmos and Savage yields (q, K) as its quotient and regular-conditional pair (**Theorem 1**); Shannon’s rate–distortion pair parameterises the same object at a distortion budget (**Theorem 2**, now witnessed exactly by **Instrument 7** on uniform and Bernoulli sources to 10^{-9}); the construction is the unit/counit of an adjunction (**Proposition 3**). Cross-task stability — a single quotient that is sufficient for a whole task family — holds iff the family admits a shared Markov screen, i.e. a common latent generator (**Theorem 4**, conditional). Discrete-case learnability is also a theorem: with the task family separating Z and the compiler’s fibres balanced ($p_{\min} \geq 1/(cM)$), empirical common-sufficient clustering recovers q with probability $\geq 1 - \varepsilon$ from $N \geq cM \ln(M/\varepsilon)$ samples (**Theorem 5**). The continuous-case extension at resolution ε is also a theorem (**Theorem 6**): bound $O(c(D_Z/\varepsilon)^{d_Z} \log((D_Z/\varepsilon)^{d_Z}/\varepsilon_{\text{rel}}))$ — polynomial in $1/\varepsilon$ at fixed d_Z , provably exponential in d_Z at fixed ε *without* further inductive bias. Adding one specific bias — **independent non-Gaussian latents with a linear mixing** — resolves that exponential-in- d_Z cost inside the linear-ICA hypothesis class: the classical Hyvärinen–Oja identifiability result becomes **Theorem 7** in our framework’s language, with **Instrument 8** witnessing polynomial recovery (Amari index ≤ 0.02 at $N = 10\,000$ for $d_Z \in \{2, 4, 6, 8\}$). **Eight exact, deterministic, unit-tested instruments** now witness each theorem on solvable cases and exhibit the dissociations the framework predicts.

1. The master object

Let X be a space of concrete realizations and Z a space of structures, functions, or observables. Two maps carry the construction:

- a **coarse-graining** $q: X \rightarrow Z$ saying which concrete differences do not matter, with fibre $q^{-1}(z)$ the set of embodiments of structure z ;
- a **compiler** $K: Z \rightsquigarrow X$, a Markov kernel with $\text{supp}K(\cdot \mid z) \subseteq q^{-1}(z)$, producing a distribution of realizations rather than a single one.

This sharpens the adjunction $R \dashv C$ of the companion synthesis: $q = C$ (coarse-grain) and K is a stochastic section of R (realize). The biology paper of Kiiskinen–Kivinen–Rivas is exactly this construction — the genome names z , the physics substrate is the compiler K that “computes the samples,” and selection acts on the ensemble statistics — and its coarse-graining threshold C^* proves that below a critical resolution the fibre is too large to address with the specification budget.

A note on rigor. Cross-domain resemblance is not isomorphism. The honest hierarchy of sameness runs: isomorphism $\supset$ bisimulation $\supset$ functor $\supset$ natural transformation $\supset$ adjunction / Galois connection $\supset$ Morita-like equivalence $\supset$ simulation-at-a-resolution. Most relations among the source works are adjunctions, simulations, and shared diagram shapes, not object-level isomorphisms; every proposed connection must state what kind of sameness it is, what the map forgets, and what would have to be proved to make it a theorem.

2. Derivations

2.1 Existence via minimal sufficiency (Theorem 1)

Let X be a standard Borel space, $\{P_\theta: \theta \in \Theta\}$ a dominated family of probability measures on X , and Y a random variable whose distribution depends on θ (equivalently: θ indexes the task). The Halmos–Savage sufficiency theorem gives a sufficient σ -algebra $\mathcal{S} \subseteq \mathcal{B}(X)$; under mild regularity (Bahadur 1954; Lehmann–Casella) there is a minimal sufficient σ -algebra $\mathcal{T} \subseteq \mathcal{S}$, unique up to P -null completion.

Define:

- $Z := X / \sim_{\mathcal{T}}$, the quotient of X by $\mathcal{T}$ -equivalence ($x \sim x'$ iff they agree on every event in $\mathcal{T}$);
- $q: X \rightarrow Z$, the canonical projection;
- $K(\cdot \mid z) := P_\theta(\cdot \mid \mathcal{T})(x)$ for any $x \in q^{-1}(z)$ — the regular conditional distribution, which by sufficiency does not depend on θ .

Theorem 1 (Existence of the master fibration). (q, K) is a stochastic fibration with $\text{supp} K(\cdot \mid z) \subseteq q^{-1}(z)$, and four of the five clauses of the Structural Intelligence Conjecture hold as theorems:

1. **Task-relevance descends.** $P_\theta(Y \in B \mid X) = P_\theta(Y \in B \mid q(X))$ almost surely.
2. **Irrelevant variation is confined to fibres.** For $x, x' \in q^{-1}(z)$: $\mathcal{L}(Y \mid X = x) = \mathcal{L}(Y \mid X = x')$.
3. **Compact specification.** $|\text{image}(q)| = |\text{atoms}(\mathcal{T})| \leq |X|$, strict iff $\mathcal{T}$ is nontrivial.

4. **Re-instantiation via K .** By construction, K is a Markov kernel $Z \rightsquigarrow X$ with the correct support.

The **Fiber Finder** (Instrument 1, §4.1) is an exact computational witness.

2.2 Rate–distortion parameterisation (Theorem 2)

Let $X \sim p_X$ on X and $d: X \times \widehat{X} \rightarrow \mathbb{R}_+$ a distortion. Shannon’s rate–distortion function

$$R(D) = \inf_{p(\widehat{x} | x): \mathbb{E}[d(X, \widehat{X})] \leq D} I(X; \widehat{X})$$

is attained by an encoder $p^*(\widehat{x} | x)$ and decoder marginal $p^*(\widehat{x})$.

Theorem 2 (RD parameterisation of the master fibration). For every distortion budget $D \geq 0$, the RD-optimal pair defines a stochastic fibration (q_D, K_D) ; the family $\{(q_D, K_D): D \geq 0\}$ is a one-parameter deformation of the sufficiency fibration. At $D = 0$ the encoder is minimal-sufficient; as D grows the fibres grow and the specification shrinks along the RD curve. The Kiiskinen–Kivinen–Rivas biology threshold C^* is $R(D_{\text{bio}})$ for the domain’s distortion measure.

2.3 Categorical restatement (Proposition 3)

Let $\mathbf{Prob}(X)$ and $\mathbf{Prob}(Z)$ be the categories of probability measures on X and Z . The sufficient-statistic construction is the unit/counit pair of an adjunction:

- **coarse-grain** $C: \mathbf{Prob}(X) \rightarrow \mathbf{Prob}(Z)$, $C(\mu) = q_*\mu$;
- **realize** $R: \mathbf{Prob}(Z) \rightarrow \mathbf{Prob}(X)$, $R(\nu) = \int K(\cdot | z)\nu(dz)$;
- $C \dashv R$, with unit $x \mapsto q(x)$ and counit given by K .

Proposition 3. (q, K) is the unit/counit pair of $C \dashv R$. SIC clauses (1)–(4) are the adjunction’s triangle identities specialised to the sufficiency reduction; the categorical framing adds compact language, not new content.

2.4 Cross-task stability (Theorem 4, conditional)

Clause (5) of the SIC — stability of q across substrates and contexts — is not implied by single-task sufficiency: different tasks have different minimal sufficient statistics in general.

Theorem 4 (Cross-task stability under latent generation). Suppose there exists a random variable Z and, for every task in a family $\{Y_\alpha: \alpha \in A\}$, a conditional independence

$$Y_\alpha \perp X | Z.$$

Then Z is a sufficient statistic for every Y_α simultaneously.

Proof. For each α , sufficiency of Z for Y_α from X is exactly the Markov property $Y_\alpha \perp X \mid Z$; the family case is the pointwise conjunction. $\square$

Corollary (Equivalence). A task family $\{Y_\alpha\}$ admits a common sufficient statistic strictly coarser than X iff there is a decomposition $X = g(Z, \eta)$ with $Z \perp \eta$ and every Y_α factoring through Z .

Theorem 4 turns SIC clause (5) into a claim about *the world* — the joint law of X and the task family — rather than about intelligence. The **Cross-Task Sufficiency** instrument (Instrument 4, §4.4) witnesses it exactly on a 4-bit Boolean world.

2.5 Learnability, discrete case (Theorem 5)

Theorems 1–4 do not immediately derive that a finite adaptive system discovers q from data. In its unrestricted form this claim is false (Locatello et al. 2019): for smooth continuous X and no auxiliary information, no unsupervised algorithm identifies a factored latent up to trivial transformations without inductive bias. Task families supply the identifying auxiliary information, and — quantitatively enough — reduce learnability from a conjecture to a theorem in the discrete case.

Setup. Let X be finite; P_X a distribution on X with min-fibre mass

$p_{\min} := \min_z P_X(q(X) = z) \geq 1/(cM)$ for the true partition $q: X \rightarrow Z$, $|Z| = M$; $\{Y_\alpha\}_{\alpha=1..K}$ a family of deterministic tasks factoring through q ; the family **separates** q (for every $z \neq z'$, some α has $h_\alpha(z) \neq h_\alpha(z')$).

Algorithm (empirical common-sufficient clustering). Given N i.i.d. samples with labels:

1. For each observed x , form the response profile $\pi(x) := (Y_1(x), \dots, Y_K(x))$.
2. Extend π to all of X by the same deterministic labelling functions.
3. Return $\widehat{q}$ defined by $\widehat{q}(x) = \widehat{q}(x') \Leftrightarrow \pi(x) = \pi(x')$.

Theorem 5 (Discrete learnability). For any $\varepsilon \in (0, 1)$,

$$\Pr[\widehat{q} = q \text{ as maps } X \rightarrow Z] \geq 1 - \varepsilon \quad \text{whenever} \quad N \geq cM \ln(M/\varepsilon).$$

Proof. Injectivity of $\Phi: z \mapsto (h_1(z), \dots, h_K(z))$ gives that $\widehat{q} = q$ once every fibre of q contains at least one observed sample. For each z , $\Pr[z \text{ missed in } N \text{ samples}] \leq (1 - 1/(cM))^N \leq \exp(-N/(cM))$. Union bound over M fibres: $\Pr[\text{some fibre missed}] \leq M \exp(-N/(cM)) \leq \varepsilon$ for $N \geq cM \ln(M/\varepsilon)$. $\square$

What the theorem does not do. It is stated for finite X and deterministic tasks; it requires the separation assumption (without it the recoverable object is strictly coarser than Z — namely the CSS of the observable task family); it requires fibre balance through c .

2.5b Continuous-case learnability at resolution ε (Theorem 6)

The honest form of the continuous claim is “recovery up to resolution ε ”, and it reduces cleanly to Theorem 5 by ε -covering.

Setup. $Z \subseteq \mathbb{R}^{d_Z}$ bounded with diameter D_Z ; $\mathcal{N}_\varepsilon \subseteq Z$ a minimal ε -net; Z_ε the induced partition into $N_\varepsilon := |\mathcal{N}_\varepsilon|$ Voronoi cells; $q_\varepsilon := \pi_\varepsilon \circ q$ the ε -quantised true partition; fibre balance and separation both at scale ε .

Theorem 6 (Continuous-case learnability, at resolution ε). For any $\varepsilon_{\text{rel}} \in (0, 1)$, $\hat{q}$ recovers q_ε exactly with probability $\geq 1 - \varepsilon_{\text{rel}}$ from

$$N \geq c \cdot N_\varepsilon \cdot \ln(N_\varepsilon / \varepsilon_{\text{rel}}).$$

For $Z \subseteq \mathbb{R}^{d_Z}$ bounded, $N_\varepsilon = O((D_Z/\varepsilon)^{d_Z})$, giving

$$N = O\left(c \cdot (D_Z/\varepsilon)^{d_Z} \cdot d_Z \cdot \log(D_Z/(\varepsilon \cdot \varepsilon_{\text{rel}}))\right).$$

Proof. Apply Theorem 5 to the discretised experiment. $\square$

Fundamental limit. Uniform sample complexity polynomial in d_Z alone — dropping the exponential dependence on d_Z at fixed ε — is **impossible without additional inductive bias on q** . Escaping the curse requires linearity (linear ICA), sparsity (independent-mechanism analysis; Gresele et al. 2021), exponential-family conditional latents with auxiliary variables (iVAE; Khemakhem–Kingma–Monti–Hyvärinen 2020), or interventional data on Z (causal representation learning; Ahuja–Mahajan–Wang–Bengio 2022). The next section resolves SIC-C-c positively inside the first of those classes.

2.5c Positive resolution of SIC-C-c for linear ICA (Theorem 7)

Theorem 6 says continuous-case sample complexity is exponential in d_Z without inductive bias. Adding one specific bias — **linear generative model with independent non-Gaussian latents** — resolves SIC-C-c positively inside that hypothesis class. This is a classical result from Hyvärinen’s independent component analysis (ICA) line (Comon 1994; Hyvärinen 1999; Hyvärinen–Oja 2000), restated here in the framework’s language and witnessed numerically as Instrument 8.

Setup (Theorem 7).

- Latent $Z \in \mathbb{R}^{d_Z}$ with independent non-Gaussian marginals (e.g. each $Z_i \sim \text{Laplace}(0, 1)$).
- Mixing matrix $A \in \mathbb{R}^{d \times d_Z}$ with $d = d_Z$, invertible.
- Observation $X = A \cdot Z$.

- No task family required beyond X itself; the identifying auxiliary structure is *the independence of the Z -components*, which is precisely the auxiliary information Locatello (2019) proves an unqualified algorithm cannot supply on its own.

Theorem 7 (Linear-ICA identifiability, classical). Under the linear-ICA setup, the un-mixing matrix $W = A^{-1}$ is identifiable up to permutation and coordinate-wise sign, and there is an efficient algorithm (e.g. FastICA, Hyvärinen 1999) whose empirical un-mixing $\widehat{W}$ satisfies

$$\text{Amari}(\widehat{W} \cdot A) \rightarrow 0 \quad \text{as } N \rightarrow \infty$$

at sample-complexity rate polynomial in d_Z and inverse-polynomial in any target accuracy, uniformly over invertible A and non-Gaussian component densities in a broad regularity class. Formally: for every $\varepsilon > 0$, $\text{Amari}(\widehat{W} \cdot A) \leq \varepsilon$ with high probability from $N = \text{poly}(d_Z, 1/\varepsilon)$ samples.

$\text{Amari}(P)$ is the standard $[0, 1]$ -valued permutation-and-sign-invariant distance on $d_Z \times d_Z$ matrices:

$$\text{Amari}(P) = \frac{1}{2d_Z(d_Z - 1)} \left[\sum_i \left(\sum_j \frac{|P_{ij}|}{\max_j |P_{ij}|} - 1 \right) + \sum_j \left(\sum_i \frac{|P_{ij}|}{\max_i |P_{ij}|} - 1 \right) \right].$$

Proof (sketch, classical). Independent non-Gaussian marginals uniquely identify the linear model up to permutation and sign (Comon 1994). Whitening followed by rotation to maximise a non-Gaussianity contrast (negentropy in Hyvärinen 1999) converges to $\widehat{W}$ such that $\widehat{W} \cdot A$ is a signed permutation; the convergence rate follows from standard M-estimator analysis of the empirical contrast. $\square$

What Theorem 7 does not say. It resolves SIC-C-c only inside the linear ICA hypothesis class. Every other inductive-bias class in the identifiable-representation-learning line (sparse ICA, independent-mechanism analysis, iVAE with auxiliary variables, interventional CRL) has its own analogue, each with its own sample-complexity theorem. A full SIC-C-c programme would add one more instrument per class; **Instrument 8 (§4.8) is the first.**

2.6 SIC in the honest split

Theorems 1–6 collapse SIC to a derived skeleton with one empirical antecedent and one residual inductive-bias question:

Clause	Status	Notes
SIC-A Existence	Theorem	Theorem 1 + Proposition 3
SIC-B Cross-task stability	Conditional theorem	Theorem 4; antecedent empirical about the world

Clause	Status	Notes
SIC-C-a Discrete learnability	Theorem	Theorem 5; numerically sharp constant (Instrument 5)
SIC-C-b Continuous learnability at resolution ε	Theorem	Theorem 6; poly in $1/\varepsilon$, exp in d_Z (Instrument 6)
SIC-C-c Uniform polynomial in d_Z	Impossible in general; theorem inside linear-ICA hypothesis class	ε -covering lower bounds + Locatello 2019 rule out an unqualified bound; Theorem 7 gives it inside linear ICA (Instrument 8). Other inductive-bias classes (sparse ICA, iVAE, interventional CRL) admit their own analogues; each is a separate theorem-instrument pair, and the Observatory is set up to host them.

What remains genuinely open is only the general programme: which inductive biases make polynomial-in- d_Z learnability possible for *which* continuous hypothesis classes — a mainstream question in identifiable representation learning, addressed one class at a time.

3. The conjecture (full-strength form)

Structural Intelligence Conjecture. A finite adaptive system’s central capability is to discover 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 are compactly specifiable in Z , (4) those specifications re-instantiate 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.**

4. Nine exact instruments

Instruments 1, 4, 5, 6, 7, 8, and 9 are computational witnesses of Theorems 1, 4, 5, 6, **2, 7 (linear-ICA class)**, and **7 (sparse-mechanism class)** respectively; instruments 2 and 3 establish auxiliary dissociations on solvable cases. All are deterministic and unit-tested; seven are exact enumerations and two (Instruments 8 and 9) are fixed-seed Monte Carlo.

4.1 Fiber Finder — witness of Theorem 1

Over all 2^n worlds of an n -bit Boolean space with a known ground-truth invariant, enumerate a lattice of candidate quotients and three selectors:

- `minimal_sufficient` — among quotients with $H(Y | q(X)) = 0$, minimise description length;
- `mdl_only` — minimise description length;
- `accuracy_only` — maximise mutual information, tie-broken to the finest map.

Result (exact). In every task, `minimal_sufficient` recovers the exact ground-truth invariant; `mdl_only` selects the constant map (insufficient); `accuracy_only` selects the identity (sufficient but uncompressed). Sufficiency, description length, and accuracy are three different quantities; only the sufficiency-then-compress rule finds the invariant.

4.2 Structure compiler

An abstract automaton exhibits accumulation, a phase transition, and hysteresis. Four compilers F_i map its trajectory into music (pitch/octave), a visual field (height/hue), text (regime-keyed lexicon), and spatial navigation (regime-gated corridor); each has a readback q_i .

Result (exact). For every medium $q_i \circ F_i = \text{id}$ on the trajectory (fidelity 1.0); all media read back to the same abstract trajectory. Four embodiments are verifiably one work, not four mood-matched artefacts.

4.3 Agency benchmark — the seed of TRB Track 3

An exact finite-state world with target and failure regions; futures enumerated to the horizon. A symbolic model m is an operation on the trajectory distribution $P(\gamma | s)$. Measure signal $\Delta_{\text{KL}} = \text{KL}(P(\gamma | m) || P(\gamma | \text{baseline}))$, **control** (goal_gain), **knowledge** (predictive_accuracy), and **agency** (control + small calibration_error between claimed and true do-effect + positive transfer under a perturbation the intervention did not choose).

Result (exact). Seven hand-built conditions realize distinct metric signatures recovered by a fixed classifier: `noise_signal` moves the distribution with zero control; `knowledge_only` predicts with zero signal; `false_credit` improves the observed outcome while its true do-effect is zero and its self-attribution is miscalibrated; a brittle controller controls but does not transfer. **No single scalar of behavioural influence identifies agency.**

This is the exact-solvable formalization of what TRB Track 3 (self-attribution) measures on trained models. See `BENCHMARK_SPEC.md` §3 Track 3 — the `false_credit` condition is the ground truth for the balanced-accuracy score.

4.4 Cross-task sufficiency — witness of Theorem 4

On the 4-bit Boolean world $X = \{0, 1\}^4$ with latent $Z(x) = (\text{parity}_{\{0,1\}}(x), \text{parity}_{\{2,3\}}(x))$, enumerate the lattice of quotients. Two task families:

- **Shared through Z :** $\text{parity}_{\{0,1\}}, \text{parity}_{\{2,3\}}, \text{parity}_{\{0,1,2,3\}}$ — every task factors through Z .
- **Not shared beyond Z :** $\text{bit}_0, \text{bit}_1, \text{bit}_2, \text{bit}_3$ — each task reveals a single coordinate.

Result (exact). Shared family: coarsest CSS is exactly $\text{joint}(\text{parity}_{\{0,1\}}, \text{parity}_{\{2,3\}}) = Z$ (image size 4, description length 2 bits, strictly less than $\log_2 |X| = 4$ bits). Not-shared family: coarsest CSS is the identity (no compression). For the shared family, each single task’s minimal sufficient statistic is a 1-bit parity (image size 2), strictly coarser than the family CSS: **combining tasks tightens the required partition** by a factor of two.

4.5 Cross-task learnability — witness of Theorem 5

Building on the shared task family of Instrument 4, compute the exact recovery probability $\Pr[\widehat{q} = q]$ of empirical common-sufficient clustering as a function of sample count N , via inclusion–exclusion over the $2^M = 16$ subsets of fibres.

Distribution on $X = \{0, 1\}^4$	$p_{\min}$	c	N_{bound} at $\varepsilon = 0.05$	Exact $\Pr[\text{recover}]$
uniform	1/4	1	18	0.9775
skewed (0.625, 0.125, 0.125, 0.125)	1/8	2	36	0.9756

Both strictly above $1 - \varepsilon = 0.95$. Below $M = 4$ samples recovery is impossible (pigeonhole). Uniform recovery hits 0.95 already at $N = 14$: the union-bound overhead is 4 samples in this regime — the theorem bound is honest and only mildly loose.

4.6 Cross-task learnability, continuous — witness of Theorem 6

Theorem 6's continuous bound, verified exactly across $(d_Z, r) \in \{1, 2\} \times \{4, 8, 16\}$ on ambient $X = [0, 1]^2$ quantised to a 16×16 grid.

d_Z	r	M	N_{bound}	Pr[recover @ bound]
1	4	4	18	0.9775
1	8	8	41	0.9667
1	16	16	93	0.9609
2	4	16	93	0.9609
2	8	64	458	0.9538
2	16	256	2187	0.9521

All six grid points meet the $1 - \varepsilon_{\text{rel}} = 0.95$ target at Theorem 6's bound. Recovery is zero below M (pigeonhole) and monotone in N .

Ratio witness. $N_{\text{bound}}(d_Z = 2, r)/N_{\text{bound}}(d_Z = 1, r) = 5.17, 11.17, 23.52$ at $r = 4, 8, 16$ — cleanly matching the $r \cdot (\text{logslack})$ prediction of Theorem 6. From $d_Z = 1$ to $d_Z = 2$ at $r = 16$ multiplies the required sample count by $\approx 23.5 \times$. That is the **curse of dimensionality made numerical**.

Empirical common-sufficient clustering **saturates the ε -covering lower bound**: it is optimal within the class of algorithms that make no inductive assumption on q . Escaping the exponential-in- d_Z scaling requires an inductive bias — Instrument 8 exhibits the first such escape.

4.7 Rate–distortion pair — witness of Theorem 2 (experiments/rate_distortion_pair)

Theorem 2's rate–distortion parameterisation, verified exactly on two finite sources with Hamming distortion. Closed-form $R(D)$ is evaluated at 10 D-grid points for the uniform source on $n = 4$ symbols ($R(D) = \log_2(n) - h(D) - D \cdot \log_2(n - 1)$ for $D \leq 1 - 1/n = 0.75$) and at 9 points for Bernoulli($p = 0.3$) ($R(D) = H_2(p) - h(D)$ for $D \leq \min(p, 1 - p) = 0.3$). For each source we also construct the RD-optimal test channel explicitly (the symmetric error- D channel for uniform; the Bernoulli-flip channel for Bernoulli) and verify $I(X; \hat{X}) = R(D)$ at every point in the achievable regime.

Result (exact). All ten pre-registered gates pass to 10^{-9} :

- **Uniform on 4 symbols:** $R(0) = 2$ bits (source entropy, Theorem 1 anchor); $R(0.75) = 0$ (encoder collapses to constant); $R(D)$ monotone non-increasing and convex on the grid; closed form matches at every point; test channel achieves $I(X; \widehat{X}) = R(D)$ at every $D \leq 0.75$.
- **Bernoulli(0.3):** $R(0) = H_2(0.3) \approx 0.881$ bits; $R(0.3) = 0$; same monotonicity / convexity / achievability checks.

This closes the theorem–instrument pairing: Theorems 1, 2, 4, 5, and 6 all have exact witnesses. The Theorem 2 anchor $R(0) = H(X)$ reconnects to Theorem 1’s minimal-sufficient partition — the RD family (q_D, K_D) is literally a one-parameter deformation of the sufficiency fibration.

4.8 Linear-ICA learnability — witness of Theorem 7 (`experiments/linear_ica_learnability`)

Theorem 7’s numerical witness. On the linear-ICA setup — $X = A \cdot Z$ with $A \in \mathbb{R}^{d \times d}$ random orthogonal (fixed numpy seed for determinism), each $Z_i \sim \text{Laplace}(0, 1)$ independent — sweep $(d_Z, N) \in \{2, 4, 6, 8\} \times \{200, 500, 1000, 2000, 5000, 10000\}$ and run `sklearn.decomposition.FastICA` (fixed `random_state`, 8 trials per grid point). Recovery quality is the Amari index on $P = \widehat{W} \cdot A$; smaller is better.

Result (deterministic under seed). All four pre-registered gates pass:

- **`linear_ica_converges_at_largest_N`** : at $N = 10\,000$, mean Amari index over 8 trials is ≤ 0.02 for every $d_Z \in \{2, 4, 6, 8\}$.
- **`sample_complexity_polynomial_in_d_Z`** : fitted $\log N_{\text{needed}} = a + b \cdot \log d_Z$ gives exponent $b \approx 0.06 \ll 3$ (the pre-registered polynomial-bound gate). In this regime linear-ICA sample complexity is effectively flat in d_Z — a factor Theorem 7 permits but does not require, arising because $d_Z = 8$ is still small relative to the constant hidden in the Hyvärinen–Oja $\text{poly}(d_Z, 1/\varepsilon)$ bound.
- **`amari_monotone_in_N_at_every_d_Z`** : within tolerance 0.01, Amari decreases in N at every d_Z .
- **`escapes_theorem_6_exponential`** : at some $d_Z \geq 6$, an $N \leq N_{\text{bound, thm 6}}(d_Z)$ achieves the accuracy target — i.e. FastICA takes fewer samples than the ε -covering bound of Theorem 6 predicts under the “no inductive bias” regime. **The inductive bias earns you the escape.**

Fixed-seed protocol keeps it reproducible run-to-run. **Refined numbers from the live**

Observatory (structural-observatory-production.up.railway.app): at $N = 10,000$ mean Amari is ≤ 0.009 for every d_Z ; fitted $b \approx 0.06$; the escape from Theorem 6’s ε -covering bound is a $462 \times$ improvement in sample count at $d_Z = 8$.

4.9 Sparse-ICA (IMA) learnability — second SIC-C-c resolution ([experiments/sparse_ica_learnability](#))

Same shape as Instrument 8 but with *A sparse-orthogonal* at retention $s \in \{0.5, 0.25\}$ — the independent-mechanism-analysis / sparse-mixing class of Gresele et al. (2021). Four pre-registered gates pass; fitted polynomial exponents $b \approx 0.00$ ($s = 0.5$) and $b \approx 0.51$ ($s = 0.25$), both well inside $b \leq 3$. **Confirms that two distinct inductive-bias classes each earn escape from Theorem 6's ε -covering bound**, so SIC-C-c's "impossible without inductive bias" is a genuinely local statement — the pattern is that each identifiable-representation-learning class admits its own theorem, not that one privileged class does. Other classes (iVAE with auxiliary variables; interventional CRL) remain open as separate theorem-instrument pairs.

2.5d Concern as fibre geometry — companion paper (Theorems CG-1, CG-2)

A short companion paper *Concern as Fibre Geometry* ([concern_as_fiber_geometry.pdf](#) , 2026-08-03) develops the concern layer of §5.1 into two theorems:

- **Theorem CG-1 (Fisher information on the fibre)**. Under an exponential-family concern reweighting $K_c(dx | z) \propto e^{\beta \langle c, T(x, z) \rangle} K(dx | z)$, the fibre-restricted Fisher matrix is $g_{c,z} = \beta^2 \cdot \text{Cov}_{c,z}[T(\cdot, z)]$. Operationally: two concerns are distinguishable in n i.i.d. samples from the fibre iff their KL is $\gtrsim 1/n$, iff their geodesic distance under $g_{c,z}$ is $\Theta(1/\sqrt{n})$. Concern geometry has a sample-complexity meaning.
- **Theorem CG-2 (Concern holonomy = failure of exactness)**. For the canonical concern 1-form $\alpha(c, z) = \beta \cdot \mathbb{E}_{c,z}[T(\cdot, z)] dc$: holonomy $\oint_\gamma \alpha = 0$ on every closed loop γ iff α is exact ($\alpha = d\Phi$). Non-vanishing holonomy is the statement that concern cannot be recovered from a single potential on $Z \times C$ — concern is irreducibly path-dependent.

Worked example on Instrument 4's 4-bit Boolean world; rectangular- and triangular-loop holonomy matches predictions to 10^{-6} . The Trace-AI relevance is that CG-1 and CG-2 together give TRB Tracks 4 (Viability) and 5 (Navigability) their formal shape — see [BENCHMARK_SPEC.md](#) .

5. The extended program (ten further constructs)

Each is a target generated by the master object; those marked conjectural are not proved here.

1. **Concern as fibre geometry** — a concern state reweights the compiler,
 $K_c(dx | z) \propto e^{\beta U_c(x, z)} K(dx | z)$, inducing an information geometry on concern states. *Status*

update: now a theorem (CG-1 + CG-2), see §2.5d and the companion paper [apps/site/concern_as_fiber_geometry.pdf](#).

2. **Conditional rate–distortion control limit** (*conjectural*) — 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 $B_{\text{extra}} \geq R_{Y|Z}(D)$.
3. **Abstraction frontier** — the Pareto set trading task-sufficiency $I(Y; X | q(X))$, dynamical closure, cost $H_0(Z)$, and control regret; an antichain that explains why two representations can both be “right” yet incomparable.
4. **Fibre audit** — vary the allegedly irrelevant degrees of freedom while holding q fixed and measure $\Delta_q(z) = \sup_{x, x' \in q^{-1}(z)} d(P(Y | \text{do}x), P(Y | \text{do}x'))$.
5. **Theory atlas** — theories as charts M_i on contexts U_i with translations T_{ij} ; test the cocycle $T_{jk} \circ T_{ij} = T_{ik}$.
6. **Compiler tomography** — infer the shared compiler and compact specs from many (s_i, x_i) by MDL; then compiler ecology $K_{t+1} = U(K_t, \text{outcomes})$.
7. **Causal semantics** — two symbols are equivalent when they induce naturally equivalent update operators $\Psi_{m,c}$ across independent contexts.
8. **Representation-repair calculus** — a library mapping failure signatures to minimal structural lifts (scalar $\rightarrow$ operator, global norm $\rightarrow$ localized measure, quotient $\rightarrow$ restored fibre, static $\rightarrow$ path space, affine $\rightarrow$ projective, point $\rightarrow$ ensemble, non-composing $\rightarrow$ interface, symmetry $\rightarrow$ gauge-fix).
9. **Alignment as ensemble governance** (*conjectural*) — 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 is the alignment evaluation.
10. **Autocatalytic artwork** — $S_t \xrightarrow{K_t \text{ experience}} E_t \rightarrow K_{t+1}$: early movements teach the grammar by which later movements become legible.

6. Capstone: conscious and reliable agents, honestly

Two natural questions bear on this program — can agents be made *conscious*, and can they be made *never wrong* — and the disciplined answer to both is **not in the literal sense**. The strongest defensible construction factors as two nested systems.

Concerned Self-Modeling Core. An active, self-maintaining fixed point of the coarse-graining loop: a persistent agent that maintains world- and self-models, represents concern-weighted futures, broadcasts selected information, remembers commitments, predicts its own action consequences, performs false-credit tests on its own causal claims (exactly Instrument 3’s calibration metric), and reports uncertainty. It **operationalizes** proposed

consciousness indicators — and that is the ceiling of the claim: **functional selfhood** $\nRightarrow$ **subjective consciousness**.

Proof-Carrying Reliability Shell. The fibration with a verifier on the count: convert goals to contracts, separate observation from inference, attach provenance to every claim, and commit 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 (§5.4). **Verified-in-a-bounded-domain** $\nRightarrow$ **universal infallibility**: verification certifies the formal statement, not that it captures intent.

The honest breakthrough available now is therefore not conscious, perfect agents, but agents with **experimentally measurable selfhood** and **formally bounded error** — two separate, non-inflated claims.

7. Limitations

The eight instruments are toys by design (finite and grid-quantised Boolean and grid worlds, tiny Markov systems, lossless encoders; Instrument 8 is fixed-seed Monte Carlo). The master fibration (q, K) is derived (Theorem 1 gives it as a mathematical object for any well-posed task on a standard Borel space; Theorem 2 parameterises it (Instrument 7 witnesses it exactly); Proposition 3 gives the categorical restatement; Theorem 4 makes cross-task stability equivalent to a shared Markov screen; Theorems 5 and 6 pin down discrete- and continuous-case sample complexity; Theorem 7 gives polynomial-in- d_Z learnability inside the linear-ICA hypothesis class). The residual content is:

- **SIC-B antecedent (empirical).** Whether the actual world's task ensemble admits a shared Markov screen is an empirical question about physics, biology, cognition, and culture.
- **SIC-C-c (uniform polynomial-in- d_Z learnability).** Theorem 6's $(D_Z/\varepsilon)^{d_Z}$ bound is exponential in d_Z at fixed ε ; ε -covering lower bounds show this cannot be improved without additional inductive bias on q . Escaping the curse requires structural assumptions (linear ICA — **Theorem 7 gives this class, first instance**; independent-mechanism analysis, Gresele et al. 2021; iVAE, Khemakhem–Kingma–Monti–Hyvärinen 2020; interventional causal representation learning, Ahuja–Mahajan–Wang–Bengio 2022). Each supplies a hypothesis class within which learnability is polynomial in d_Z ; extending the Observatory to any other specific such class is future work, not a theorem proved here.
- **Extended-program conditionals** in §5 remain flagged.

- **No Lean-checked proofs** of Theorems 1–6 yet — the proofs are in paper form; formalisation would be next-order rigor.
- **No machine phenomenology.** Nothing in Theorems 1–7 licenses a claim of subjective experience; §6 draws that line explicitly.

8. How this relates to Trace AI

Trace AI is the SIC applied to cognition, with a *concrete choice* for each abstract variable:

SIC object	Trace AI instantiation	Where
ambient space X	model's compute state z_t	whitepaper §2
latent Z	self-representation r_t (a few legible slots)	whitepaper §2, §4.2
coarse-graining q	decoder $h_\phi: z_t \mapsto r_t$	whitepaper §2
compiler K	any of the subject-kinds (human, model, hybrid) that realize the schema	schema/reasoning_trace.schema.json
task family $\{Y_\alpha\}$	the counterfactual set $\mathcal{R}^*$ + paired continuations	whitepaper §4.4; COLLECTION_PROTOCOL.md
Instrument 3 (agency)	TRB Track 3 (self-attribution) — the exact-solvable formalization	BENCHMARK_SPEC.md §3 Track 3
Instrument 1 (Fiber Finder)	the “schema follows the data” discipline (representation search done by hand)	STRUCTURAL_INTELLIGENCE.md §3.1
SIC-B (cross-task stability)	schema portability across substrates (human, model), Observation 1 in whitepaper §6.4	whitepaper §6.4
SIC-C-c residual $\rightarrow$ linear-ICA class	Trace AI's L2 sample-complexity limit ($\sim 10^3$ human counterfactual pairs) is a working budget;	whitepaper §8 L2

SIC object	Trace AI instantiation	Where
	Theorem 7 says polynomial-in-d_r learnability is a theorem inside the linear-ICA hypothesis class. The strategic implication: if r_t can be arranged (by architecture + training) so its slot components are statistically independent and non-Gaussian, the L2 budget shrinks from exponential to polynomial in the number of slots. That reframes the collection operation from “brute-force scale” to “engineer the bottleneck for identifiability”.	
Instrument 7 (rate–distortion pair)	Exact-solvable model for the whitepaper §2.1 information-bottleneck claim — the r_t readout as a coarse-graining at a distortion budget	whitepaper §2.1
Instrument 8 (linear-ICA learnability)	Numerical existence proof that inductive bias flips d_r-scaling from exponential to polynomial. If Trace AI adopts an ICA-flavored decoder for h_ϕ, the sample-complexity story changes materially	whitepaper §8 L2; COLLECTION_PROTOCOL.md §6

The whitepaper is Trace AI at the level of *what* to build; this paper is *why* the object exists as a mathematical object at all. Read the SIC when a reviewer asks “is your latent even a well-defined thing to be looking for”; read the whitepaper when they ask “what is the training signal that finds it.”

v0.2 (integrated 2026-08-03), tracks paper v4. See [DECISIONS.md](#) (D28, D30) for provenance. Reproduction of the eight instruments lives in the sibling research repository [jawauntb/research-derived-experiments](#) — all [experiments/](#) paths in the instrument tables above resolve there, including the new [experiments/rate_distortion_pair](#) and [experiments/linear_ica_learnability](#). Exact-recovery numbers in §4.5–§4.8 are copied verbatim from that repository’s outputs at the paper’s commit. The originating research note is at [notes/structural_intelligence_conjecture.md](#); the working-note sibling of the present document is [SIC_RESEARCH_PROGRAM.md](#).