Projection, Constraint, and Irreversibility

Toward a Unified Theory of Representation Across Cognition, Computation, and Social Systems

Flyxion

Independent Researcher

May 6, 2026

Abstract

Complex systems inhabit high-dimensional, irreversible trajectory spaces that resist direct representation. Practical operation requires compression: metrics, interfaces, embeddings, and abstractions project rich process spaces onto tractable manifolds. This essay argues that many contemporary systemic failures—artificial intelligence hallucination, metricized economic reasoning, digital cultural homogenization, platform instability, and institutional identity reduction—arise not from simple error but from a deeper structural pathology: the conflation of a compressed representational manifold with the trajectory space it summarizes.

The essay develops a unified formal framework centered on a generalized reduction operator $\mathcal{R} : \mathcal{X} \to \mathcal{X}'$ preserving observable projections $\pi : \mathcal{X} \to \mathcal{M}$ while contracting internal degrees of freedom. Within this framework, reductions are classified along two axes: the invariant structure preserved under π, and the recoverability of erased fiber content from admissible side information. Irrecoverable reduction—in which no decoder exists even in principle—constitutes the deepest failure mode and provides the formal grounding for an ethics of representation.

Three ordered failure modes are distinguished: simple error, projection mismatch, and ontological compression failure. The third mode is uniquely dangerous because its deficits are undetectable from within the representational manifold: the system cannot formulate the question of what it destroyed. The essay further develops fiber bundle language for platform architectures, sheaf-theoretic semantic infrastructure as an alternative to syntactic version control, and a derivation of ethical constraints from stability conditions rather than normative supplement. The central thesis is that intelligence, ethics, computation, and institutions must ultimately be understood as processes operating under irreversible constraint within partially observable trajectory spaces.

Contents

Introduction

> *The map is not the territory. But when every navigator has only the map, the territory ceases to matter until it floods.*

Modern civilization operates increasingly through compressed representations. Economies are governed by metrics. Software systems communicate through typed interfaces. Social life unfolds across algorithmic feeds. Scientific knowledge is archived in abstractions. Political entities are classified by demographic categories. Artificial intelligence systems produce outputs by traversing high-dimensional embedding spaces and projecting into natural language.

In each case, the underlying system being represented is historical, path-dependent, thermodynamic, embodied, and high-dimensional. What is acted upon, however, is a reduced image: a manifold of tractable observables that summarizes, encodes, and—necessarily—distorts the original space.

This essay argues that the structural relationship between trajectory space and representation constitutes the central theoretical problem of complex system design, and that most contemporary systemic failures can be understood as specific failure modes within a single formal architecture. That architecture is captured by the projection:

$$\pi : \mathcal{X} \to \mathcal{M} \tag{1}$$

where $\mathcal{X}$ denotes a rich trajectory space of processes, states, and histories, and $\mathcal{M}$ denotes a reduced representational manifold of observable outputs.

The core failure is not complex. It is the systematic treatment of $\mathcal{M}$ as though it were identical to $\mathcal{X}$—as though the map were the territory, the output were the process, the metric were the thing measured, the word were the thought, the profile were the person.

What makes this error so persistent and so productive of pathology is that it is structurally induced rather than merely cognitive. A system designed to act on $\mathcal{M}$ *cannot represent* what projection has erased. The gap between $\mathcal{X}$ and $\mathcal{M}$ is not visible from within $\mathcal{M}$. Systems caught inside this gap cannot self-correct because they cannot formulate the question of their own deficit.

The argument proceeds as follows. Section 2 develops the concept of abstraction as active reduction rather than mere concealment, establishing a general-

ized reduction operator with a precise invariant structure. Section 3 develops the formal consequences of projection's many-to-one character, with applications to linguistic compression and embodied cognition. Section 4 introduces irreversibility as a constitutive feature of identity and cognitive systems, rejecting state-based models in favor of trajectory-indexed constraint surfaces. Section 5 formalizes constraint closure as a primitive criterion for genuine value production, distinguishing it from signal optimization and developing its consequences for political economy. Section 6 proposes a semantic infrastructure for computation grounded in sheaf-theoretic coherence rather than syntactic line-diffing. Section 7 applies fiber bundle language to digital platform architectures, giving formal precision to phenomena such as context collapse. Section 8 synthesizes these analyses into a three-level taxonomy of representational failure modes, distinguished by their detectability and recoverability characteristics. Section 9 derives an ethics of representation from the stability conditions established by the foregoing framework. Section 10 returns to the central thesis and its implications for the design of representational systems.

Throughout, the formalism is treated as load-bearing rather than decorative. Claims expressed mathematically are meant to carry their own argumentative weight; the surrounding prose draws out their consequences rather than substituting for them.

Remark 1.1 (Structural Isomorphism, Not Ontological Identity). The framework treats cognition, institutions, platforms, and computation as structurally homologous reduction systems. This requires clarification: the claim is not that these domains are ontologically identical or that institutions are thermodynamic systems in the same physical sense as heat engines. The claim is that they instantiate analogous reduction architectures—that the formal relationships between $\mathcal{X}$, $\mathcal{M}$, $\mathcal{R}$, and π recur across domains with domain-specific realizations. Structural isomorphism does not entail ontological identity. The recurring formal pattern is explanatorily significant precisely because it appears in domains with heterogeneous physical substrates.

Abstraction as Reduction

Against the Complexity-Hiding View

The standard account of abstraction treats it as a device for managing complexity by hiding irrelevant detail. On this view, a well-chosen abstraction exposes a useful interface while concealing implementation: the abstraction layer is a convenient fiction that allows reasoning at a higher level without committing to lower-level specifics.

This view is not wrong, but it is importantly incomplete. It describes the function of abstraction from the perspective of a user who already possesses a stable abstraction. It does not describe how stable abstractions arise, what makes them stable, or what is lost in their formation.

A more adequate account must treat abstraction as an active process of reduction: the discharge of unresolved structural tensions within a system until a stable configuration is reached. On this view, an abstraction is not a convenient veil drawn over complexity but a terminal state—the result of a reduction process that has exhausted its admissible degrees of freedom.

The intuition is clearest in formal computation. In the lambda calculus, a term M is in *normal form* if no further β-reduction is applicable: every reducible expression (β-redex) has been discharged. Reduction is not simplification in any informal sense; it is the systematic elimination of unresolved computational tension. The normal form is stable precisely because no further reduction is possible, not because all information has been preserved.

Definition 2.1 (Normal Form)**.** A term M in the untyped lambda calculus is in *β-normal form* if it contains no subterm of the form $(\lambda x.N)\, P$.

Normal forms may fail to exist (non-terminating terms) or may be non-unique in the absence of confluence. But where they exist and are unique (as guaranteed by the Church-Rosser theorem for confluent reduction strategies), they represent the unique stable residue of a reduction trajectory.

The philosophical point is that the normal form encodes the observable behavior of the original term under the chosen reduction strategy while discarding the specific reduction path. Multiple distinct terms may share a normal form. The normal form is therefore already a projection: many-to-one, path-discarding, and stable.

A Generalized Reduction Operator

The structure of lambda-calculus reduction can be abstracted into a general operator applicable across computational, thermodynamic, cognitive, and institutional domains.

Definition 2.2 (Generalized Reduction Operator)**.** Let $\mathcal{X}$ be a space of system configurations and $\pi : \mathcal{X} \to \mathcal{M}$ a projection onto a representational manifold. A *generalized reduction operator* is a map $\mathcal{R} : \mathcal{X} \to \mathcal{X}'$ satisfying:

$$\pi(\mathcal{R}(x)) = \pi(x) \quad \forall x \in \mathcal{X} \tag{2}$$

$$\dim(\mathcal{X}') < \dim(\mathcal{X}) \tag{3}$$

$$|\pi^{-1}(m)| \text{ decreases monotonically along } \mathcal{R}\text{-orbits} \tag{4}$$

Condition (2) states that reduction preserves observable projections: the interface $\mathcal{M}$ cannot detect what has been removed. Condition (3) states that the reduced space is strictly smaller than the original—reduction is not a rearrangement but a genuine contraction. Condition (4) states that fibers over points in $\mathcal{M}$ shrink as reduction proceeds: the pre-images of observable outputs contain fewer and fewer distinct trajectories.

This formalism immediately unifies several apparently distinct phenomena:

- *Beta reduction* is symbolic constraint discharge: the redex $(\lambda x.N)\, P$ is the unresolved tension, its normal form the terminal stable state.
- *Thermodynamic relaxation* is energetic constraint discharge: a non-equilibrium state evolves toward minimum free energy under the constraint that macroscopic observables are preserved.
- *Predictive coding* is informational constraint discharge: a generative model minimizes prediction error (free energy in Friston's sense) by compressing sensory data into a reduced representational format.
- *Institutional metric formation* is sociotechnical constraint discharge: a complex organizational process is reduced to a scalar indicator that preserves certain observable correlates while discarding causal texture.

These are not analogies. They are instances of the same formal structure: a reduction operator that contracts internal degrees of freedom under preservation constraints imposed by a chosen interface.

Proposition 2.3 (Equivalence Under Reduction)**.** *Lambda reduction, thermodynamic*

relaxation, predictive compression, and institutional metric formation are all instances of the generalized reduction operator $\mathcal{R}$ with respect to their respective projection functionals π.

The critical question in each case is: *what class of structure is invariant under π, and what class is destroyed*? The answer determines both the practical utility of the abstraction and the pathologies its deployment will generate.

Before proceeding, it is useful to make explicit the subset of degrees of freedom on which the most consequential claims in this essay depend.
Definition 2.4 (Constraint-Relevant Subspace)**.** Let $: \mathcal{X} \to \mathbb{R}^n$ be a constraint functional. The *constraint-relevant subspace* is

$$S_c = \{v \in T\mathcal{X} : \exists x,\ d\ _x\ (v) \neq 0\} \tag{5}$$

the set of tangent directions along which the constraint functional is sensitive. A degree of freedom is *constraint-relevant* if it lies in S_c; otherwise it is *constraint-neutral*.

With this definition, the normative weight of the framework falls precisely: a reduction is ethically consequential when it irrecoverably erases degrees of freedom in S_c. Reductions that collapse only constraint-neutral directions are compressions in the benign sense; those that collapse S_c-directions are genuinely destructive. This precision prevents the framework from becoming a blanket critique of abstraction.

Stability, Energy, and Interface Formation

The connection between stability and energy minimization is not metaphorical but structural. In thermodynamic systems, a state is stable when it has exhausted the free energy available for further state change under given constraints. The macrostate is the projection: many microstates map to the same macrostate, and thermodynamic evolution is precisely the contraction of this fiber.

A stable interface is the informational analogue: a behavioral specification that no perturbation within the interface's operating domain can destabilize. Type systems in programming languages express exactly this structure: a type is a specification of the class of values that can appear at a position, and type-checking is the verification that all reductions preserve type membership. Poly-

morphism is the recognition that multiple distinct implementations share an interface, i.e., that the fiber over a type is non-trivial.

The general principle is:

Theorem 2.5 (Interface Stability)**.** *A representational interface $\pi : \mathcal{X} \to \mathcal{M}$ is stable under a reduction operator $\mathcal{R}$ if and only if $\pi(\mathcal{R}(x)) = \pi(x)$ for all x in the domain of $\mathcal{R}$.*

Stability is interface-relative. A reduction that is stable with respect to one projection may be catastrophic with respect to another. The choice of π is therefore not an afterthought but the constitutive decision that determines what the abstraction is for—and what it will inevitably destroy.

Projection and Representation Loss

The Many-to-One Structure

The projection $\pi : \mathcal{X} \to \mathcal{M}$ is generically many-to-one. For any point $m \in \mathcal{M}$, the fiber $\pi^{-1}(m) = \{x \in \mathcal{X} : \pi(x) = m\}$ contains all trajectories in $\mathcal{X}$ that project to the same observable output. Distinct trajectories in $\pi^{-1}(m)$ are representationally identical under π despite being causally and historically distinct.

This structure has several immediate consequences.

First, *outputs underdetermine processes*. Given an output m, no amount of analysis within $\mathcal{M}$ can recover which element of $\pi^{-1}(m)$ produced it. The projection discards provenance by construction.

Second, *representations erase history*. Two trajectories that diverge arbitrarily in $\mathcal{X}$ may converge to the same point in $\mathcal{M}$. Their shared representational identity erases the historical record of their divergence.

Third, *interventions on representations are ambiguous with respect to processes*. An intervention that moves a point in $\mathcal{M}$ from m_1 to m_2 corresponds to a family of possible process-level changes, not a single determinate one. Systems that intervene only at the representational level cannot guarantee which process-level change they have effected.

These consequences are not merely technical limitations; they are the structural sources of systematic failure in systems that treat $\mathcal{M}$ as the locus of causal action.

Recoverable and Irrecoverable Reduction

Not all projection loss is equally consequential. A crucial distinction must be drawn between reductions whose erased content remains in principle recoverable and those for which no recovery is possible.

Definition 3.1 (Recoverable Reduction)**.** A reduction $\mathcal{R} : \mathcal{X} \to \mathcal{X}'$ is *recoverable with side information* s if there exists a decoder $\mathcal{D}$ such that

$$\mathcal{D}(\mathcal{R}(x), s) \approx x \tag{6}$$

for all x in the relevant domain, where approximation is measured in a suitable metric on $\mathcal{X}$.

Definition 3.2 (Irrecoverable Reduction)**.** A reduction $\mathcal{R}$ is *irrecoverable* if no admissible decoder $\mathcal{D}$ and no admissible side information s exist such that (6) holds.

This distinction generates a principled hierarchy of reduction severity:

1. *Reversible compression*: $\mathcal{D} \circ \mathcal{R} = \mathrm{id}$ exactly. No information is lost; the reduction is a bijective encoding. Example: lossless data compression, invertible transformations.
2. *Lossy but reconstructable compression*: $\mathcal{D}(\mathcal{R}(x), s) \approx x$ for admissible s. Information is lost from the reduced representation but recoverable from independently preserved side information. Example: compiled binaries reconstructable from retained source code; normalized database tables reconstructable from schema documentation.
3. *Thermodynamically irreversible destruction*: the microstate information is absorbed into the heat bath and cannot be recovered even with complete knowledge of the macrostate. The Second Law is the formal guarantee of irrecoverability.
4. *Architecturally irrecoverable reduction*: the side information s was never recorded, or the interface π was designed to exclude it structurally. This is distinct from thermodynamic irrecoverability: the information may have existed, but the system was designed not to retain it.

The fourth category is the most important for this essay's purposes because it is the designed variety. When an institution reduces a person to a case number, the biographical trajectory is not destroyed by entropy: it continues to exist in

the person's embodied history. What is destroyed is the *institutional capacity to represent it*. The interface was designed to exclude biographical detail. The reduction is architecturally irrecoverable from the institution's perspective, even though the underlying trajectory remains intact elsewhere.

Remark 3.3. Irrecoverability has two sources: the information genuinely does not exist (thermodynamic), or the system was designed to exclude it (architectural). Exploitation and institutional dehumanization are predominantly of the architectural variety, which means they reflect deliberate design choices rather than thermodynamic necessity.

Linguistic Compression and Artificial Intelligence

The projection $\pi : \mathcal{X} \to \mathcal{M}$ takes a particularly salient form in the relationship between cognitive processes and linguistic outputs. A speaker produces a sequence of words; those words are the projection of a cognitive trajectory that includes unsymbolized reasoning, embodied sensation, emotional valence, spatial imagination, and procedural knowledge.

Shannon's channel model captures part of this structure:

$$\text{Source} \xrightarrow{\text{encode}} \text{Channel} \xrightarrow{\text{decode}} \text{Receiver} \tag{7}$$

But the crucial point is that the *source entropy* of human cognition vastly exceeds the *channel capacity* of natural language. Linguistic output is a lossy projection of cognitive process, and the fiber over any linguistic expression contains a vast and heterogeneous class of cognitive trajectories.

This has direct consequences for the interpretation of large language model outputs. LLMs operate primarily at the level of the linguistic interface $\mathcal{M}$: they model the statistical structure of projected outputs without access to the trajectory space $\mathcal{X}$ from which those outputs were projected. They can reproduce the surface statistical properties of human language without modeling the underlying cognitive processes.

A corollary is that surface linguistic convergence—two outputs being statistically indistinguishable—does not imply cognitive convergence. Many distinct cognitive processes project to the same or similar linguistic outputs. The concern that widespread LLM use will homogenize human thought conflates convergence in $\mathcal{M}$ with convergence in $\mathcal{X}$. The former may occur without the latter.

The genuine risk is more subtle: not that thought will become identical, but that the linguistic interface will narrow the range of thoughts that can be projected into communicable form—reducing the exploratory bandwidth of the channel rather than the diversity of its sources.

Wordless Cognition and the Limits of the Linguistic Interface

Empirical research on unsymbolized thinking, aphantasia, and anendophasia converges on a conclusion that the philosophy of language has often resisted: substantial portions of human cognitive life have no natural linguistic projection.

Unsymbolized thinking—the experience of a fully formed thought without accompanying inner speech or imagery—demonstrates that propositional content can be cognitively present without linguistic encoding. Aphantasia—the inability to form voluntary mental images—demonstrates that the standard imagistic substrate of much linguistic meaning may be absent in some individuals without disrupting linguistic competence. Anendophasia—the absence of inner speech—demonstrates that linguistic processing can occur without the internal linguistic monologue that many assume to be its substrate.

Each of these phenomena constitutes evidence that the trajectory space $\mathcal{X}$ of human cognition is not isomorphic to the representational manifold $\mathcal{M}$ of language. The fiber $\pi^{-1}(m)$ over any linguistic expression contains cognitive trajectories that are not themselves linguistic.

This matters for any theory of mind that treats language as transparent to thought, and it matters for AI systems that model thought through language. An LLM trained on linguistic projections has access to $\mathcal{M}$ but not to $\mathcal{X}$. It models the structure of projections without access to what was projected from. The epistemological gap between a very good model of $\mathcal{M}$ and any model of $\mathcal{X}$ cannot be closed by scaling within the linguistic domain.

Chain-of-Thought as Post-Hoc Projection

The preceding analysis implies a sharper diagnosis of a specific AI failure mode: the unreliability of chain-of-thought reasoning as a causal trace of computation.

Contemporary large language models often produce explicit reasoning steps

before outputting a final answer. This "thinking out loud" is widely interpreted as transparency: the model is showing its work, exposing the inferential path from premises to conclusion. The projection framework reveals why this interpretation is structurally unsound.

The model's internal computation is a trajectory in a high-dimensional activation space $\mathcal{X}_{\text{compute}}$. The chain-of-thought output is a sequence of tokens in the linguistic manifold $\mathcal{M}$. The relationship between them is a projection $\pi_{\text{CoT}} : \mathcal{X}_{\text{compute}} \to \mathcal{M}$ whose fiber structure is not governed by causal fidelity but by next-token probability. The model does not generate tokens because those tokens causally represent the computation that produced the answer; it generates tokens because those tokens are statistically coherent given the linguistic context.

Empirically, this prediction is confirmed: altering intermediate steps in a chain-of-thought sometimes leaves final answers unchanged [3], and models can produce confident, fluent reasoning chains for conclusions that are demonstrably wrong. These are not anomalies but expected consequences of the projection structure. The chain-of-thought is $\mathcal{M}$-plausible without being $\mathcal{X}$-faithful.

Remark 3.4 (Chain-of-Thought as Mode III Failure). A chain-of-thought output that is presented as a causal trace of computation but is generated by next-token statistics over $\mathcal{M}$ constitutes an instance of ontological compression failure. The actual computation in $\mathcal{X}_{\text{compute}}$ has no recoverable image in the chain-of-thought, because π_{CoT} was not designed to preserve causal provenance. The system cannot be corrected by inspecting its chain-of-thought, because the chain-of-thought does not contain the information required for correction. Interpretability tools that operate only on the linguistic output are operating in $\mathcal{M}$ while the relevant computation occurred in $\mathcal{X}$.

The implication is that genuine causal transparency in AI systems requires a different architecture: one in which the computational trajectory $x \in \mathcal{X}_{\text{compute}}$ is recorded with sufficient provenance that a decoder $\mathcal{D}$ can recover the actual inferential steps. Linguistic fluency in $\mathcal{M}$ is not a substitute for such provenance, and systems that conflate the two are exhibiting exactly the projection error that this essay diagnoses throughout.

Irreversibility and the Compiled Self

Identity as Trajectory

The dominant philosophical tradition treats identity as a matter of states: a self is individuated by the properties it currently instantiates, and persistence conditions specify which state-sequences constitute the persistence of a single self over time. Psychological continuity theories, biological continuity theories, and narrative theories all share this basic architecture, differing only in which state-features they take as identity-constituting.

This essay rejects the state-based framework in favor of a trajectory-indexed account. A self is not a state but an accumulation: an append-only history of constraint resolutions that has progressively shaped the admissible future trajectory space.

Definition 4.1 (Trajectory-Indexed Identity)**.** Let $\mathcal{X}$ be a trajectory space with histories indexed by time. An *identity* is a constraint surface $\Sigma \subset \mathcal{X}$ such that:

1. Σ is historically accumulated: its current extent is determined by the sequence of constraint resolutions that have occurred.
2. Σ is path-dependent: the same constraint resolution at a different point in the trajectory produces a different Σ.
3. Σ is irreversibly modified by each constraint resolution: earlier configurations of Σ cannot be recovered from later ones.

The neuroplastic substrate of identity provides a concrete instance. Synaptic connections are modified by experience in ways that are not simply reversible. The brain that has learned a language, formed a habit, sustained a trauma, or developed a skill has a physically different architecture than it had before. The earlier architecture is not latent somewhere recoverable; it is genuinely gone. Each experience modifies the constraint surface Σ in a way that constrains but does not determine future modification.

This is not merely a metaphor. Neuroplasticity is a physical instance of irrecoverable reduction: the experiential trajectory $x \in \mathcal{X}$ is encoded into the constraint surface Σ through a reduction process that does not preserve the encoding invertibly. The self that results from a trajectory is a lossy compression of that trajectory into a constraint surface.

A further consequence follows. If identity is constituted by a constraint surface Σ accumulated over a trajectory, then the medium in which Σ is instantiated must be capable of preserving the behavioral and structural signatures that distinguish Σ from other constraint surfaces. A medium that cannot sustain these signatures forces the identity it hosts into a lower-dimensional projection of itself.

Definition 4.2 (Substrate Fidelity)**.** A substrate $\mathcal{S}$ is *fidelity-sufficient* for identity Σ if it can represent the constraint surface with enough resolution that the behavioral frequency signatures of Σ—the characteristic patterns of response, constraint-resistance, and trajectory preference that individuate it—are preserved under the substrate's own reduction operator $\mathcal{R}_{\mathcal{S}} : \Sigma \to \Sigma_{\mathcal{S}}$. A substrate is *fidelity-insufficient* if $\mathcal{R}_{\mathcal{S}}$ irrecoverably collapses signatures that are constraint-relevant for Σ.

Proposition 4.3 (Substrate Fidelity Condition)**.** *An institution, platform, or medium that imposes a fidelity-insufficient substrate on an identity it claims to represent has replaced the identity Σ with a reduced image $\Sigma_{\mathcal{S}} \subsetneq \Sigma$ from which the original cannot be recovered. All subsequent interactions by that system will be interactions with $\Sigma_{\mathcal{S}}$, not with Σ.*

The substrate fidelity condition clarifies what is at stake in apparently administrative decisions. An engagement profile is a substrate; a case file is a substrate; a credit score is a substrate. Each imposes a resolution limit on the identity it hosts. The question is not whether reduction occurs—all representation reduces—but whether the substrate's resolution is sufficient to preserve the constraint surfaces that are causally relevant for the future interactions the system is designed to support. A welfare institution whose substrate cannot represent occupational history, family structure, and health trajectory is fidelity-insufficient for the constraint resolutions it is ostensibly tasked with supporting.

Constraint Surfaces and Narrative Coherence

If identity is a constraint surface Σ rather than a state, then narrative coherence is not the recounting of an essence but the projection of a trajectory onto a representational manifold.

Definition 4.4 (Narrative Admissibility)**.** A narrative $n \in \mathcal{M}$ is *admissible* for identity Σ if it is consistent with the constraint resolutions that generated Σ.

Multiple admissible narratives may exist for a single Σ. The set of admissible

narratives is the fiber $\pi^{-1}(\Sigma)$ projected into narrative space. Identity is not recovered by finding the true narrative but by recognizing that the actual constraint surface Σ underdetermines its representational image.

This has consequences for the ethics of testimony, memoir, and biographical reduction. An institution that reduces a person to a case file has not merely summarized an identity; it has replaced the constraint surface Σ with a single point in $\mathcal{M}$, collapsing the fiber to a singleton. Future interactions with the institution will be determined by this collapsed representation, not by the actual constraint surface. The person continues to exist as a full trajectory space, but the institution has become incapable of interacting with that space.

Log Sovereignty and Irrecoverable Projection

The trajectory-indexed model of identity implies a principle of *log sovereignty*: the irreversible historical encoding of a trajectory is the locus of genuine identity, and systems that discard this encoding cannot claim to interact authentically with the identity they purport to represent.

Definition 4.5 (Log Sovereignty)**.** An entity E has *log sovereignty* over its trajectory if E retains access to an encoding of its constraint history that is not subordinated to any external projection system's reduction operator.

Log sovereignty is violated whenever an external projection system applies an irrecoverable reduction to a trajectory without the trajectory's consent and without retaining the side information required for reconstruction.

The digital analogue is pervasive. A social media profile is a projection of a biographical trajectory. The platform controls the projection operator π and the reduction operator $\mathcal{R}$. The user's historical trajectory—the constraint surface Σ actually constitutive of their identity—is reduced to an engagement profile that the platform can optimize. The side information required to reconstruct Σ from the engagement profile does not exist on the platform and is typically not retained anywhere in recoverable form.

This is not merely a privacy concern. It is an identity concern in the strong sense: the platform has rendered itself incapable of interacting with the entity it purports to serve, having replaced the constraint surface with an optimizable projection.

Constraint Closure and Real Value

Signals Versus Reality

A persistent confusion in economic and institutional reasoning conflates the optimization of signals with the resolution of the constraints those signals were designed to measure. Goodhart's Law states the pathology: when a measure becomes a target, it ceases to be a good measure. The formal reason is now clear. The measure $m \in \mathcal{M}$ was designed to track some constraint condition $C(\mathcal{X}) = 0$ in the underlying process space. When the measure becomes a target, optimization pressure is applied to m directly, and the constraint $\mathcal{R}$ that minimizes m need not be the one that resolves C. The signal has been decoupled from the constraint it was meant to index.

This is not a psychological failure of institutional actors. It is a structural consequence of the many-to-one character of π. The fiber $\pi^{-1}(m)$ contains both trajectories that achieve m by resolving C and trajectories that achieve m by other means. Once optimization pressure is applied to m in $\mathcal{M}$, the system will find all elements of $\pi^{-1}(m)$, and those that achieve m without resolving C are typically lower-cost.

Constraint Closure as Primitive

In place of signal optimization, this essay proposes constraint closure as the primitive criterion for genuine value production.

Definition 5.1 (Constraint Closure)**.** A process $x : [0, T] \to \mathcal{X}$ achieves *constraint closure* if

$$(x_T) = 0 \tag{8}$$

where $: \mathcal{X} \to \mathbb{R}^n$ is a vector of external resistance conditions that the process is tasked with resolving.

Constraint closure is a condition on the process trajectory, not on its projection. A process achieves genuine value when it resolves the actual resistance conditions in $\mathcal{X}$, regardless of what signals it produces in $\mathcal{M}$. Conversely, a process that produces favorable signals in $\mathcal{M}$ without achieving $(x_T) = 0$ produces no genuine value, however attractive its representational image.

This definition unifies several distinct domains:

In *engineering*, closure is achieved when a system withstands its operating loads: the constraint conditions are physical stresses, thermal limits, chemical stabilities. An engineering design that passes its test metrics without being able to withstand the actual loads achieves metric closure but not constraint closure.

In *labor*, closure is achieved when a task resolves the conditions that required the task: a harvest is complete when the crop is secured, not when the scheduled labor hours are logged. Labor that logs hours without resolving the harvest condition achieves metric closure without constraint closure.

In *thermodynamics*, closure is achieved when a system equilibrates with its environment: entropy production ceases when the system has exhausted its free energy with respect to the imposed boundary conditions.

In *medicine*, closure is achieved when a patient's condition resolves, not when their treatment protocol is completed or their insurance claim filed. Medical bureaucracy is largely the proliferation of metric-closure requirements that may or may not track constraint closure.

Residue, Persistence, and the Bullshit Economy

Where constraint closure provides a principled criterion for genuine value, its negation characterizes what David Graeber identified as "bullshit jobs": positions whose primary output is the optimization of representations without the resolution of the constraint conditions those representations index.

The formal characterization requires the concept of residue:
Definition 5.2 (Residue)**.** The *residue* of a process x at time T is

$$\rho(x, T) = \ (x_T) \tag{9}$$

the remaining unresolved constraint at process termination.

A process is *genuine* if $\rho(x, T) = 0$: it resolves what it was tasked to resolve. A process is *residual* if $\rho(x, T) \neq 0$: it terminates without resolution. A process is *extractive* if it reduces ρ for one trajectory while increasing ρ for another: it transfers constraint load rather than resolving it.

Modern institutional economies contain a substantial and growing proportion of residual and extractive processes: compliance documentation that satisfies audit constraints by shifting real constraints onto other parties; financial

instruments that reduce reported volatility by redistributing it into systemic risk; platform moderation that satisfies content policy metrics by suppressing legible violations while permitting more ambiguous harms; management hierarchies that optimize reporting structures while delegating actual constraint resolution downward.

The formal point is that these are not corruption or inefficiency in a contingent sense. They are the structural output of systems that optimize $\mathcal{M}$-valued signals rather than $\mathcal{X}$-valued constraint conditions. The extractive equilibrium is the natural attractor of systems in which π is high-dimensional and is difficult to measure.

Proposition 5.3 (Extractive Equilibrium)**.** *In a system where $\pi : \mathcal{X} \to \mathcal{M}$ has large fibers and $: \mathcal{X} \to \mathbb{R}^n$ is not observable in $\mathcal{M}$, optimization pressure on $\mathcal{M}$-valued objectives will generically produce extractive processes: signal-optimal trajectories that transfer rather than resolve constraint load.*

The proof is the observation that $\pi^{-1}(m)$ contains both resolving and transferring trajectories, and that transferring trajectories are typically lower-cost for the actor who performs them (the cost being externalized onto whoever receives the transferred constraint).

Semantic Infrastructure and Computation Beyond Syntax

The Failure of Syntactic Systems

Contemporary software development relies on syntactic version control: systems like Git represent codebases as sequences of text lines and compute differences as line-level edits. Merge operations combine histories by resolving line-level conflicts.

The fundamental limitation of this approach is that syntax lacks semantic provenance. Two line-level edits may be syntactically compatible but semantically contradictory: one change may rename a variable throughout a module, another may introduce a new variable with the same name in a single function, and the merge system will see no conflict at the line level while silently introducing a namespace collision that the runtime will propagate.

The formal diagnosis is that the syntactic reduction operator $\mathcal{R}_{\text{syn}} : \mathcal{X}_{\text{semantic}} \to \mathcal{M}_{\text{syntactic}}$ has enormous fibers. Many semantically distinct programs project to

identical or nearly identical syntactic representations, and syntactic merge operations cannot distinguish semantically significant from semantically neutral changes.

This is not a failure of any particular implementation. It is the consequence of treating semantic provenance as outside the scope of the representational manifold. The choice of π determines what merges can detect, and syntactic π cannot detect semantic conflict.

Sheaf-Theoretic Semantic Infrastructure

An alternative infrastructure grounds code representation in semantic fields rather than syntactic text. The relevant mathematical structure is the sheaf: a sheaf assigns consistent local data to open sets of a topological space subject to gluing conditions that ensure global consistency.

Definition 6.1 (Semantic Sheaf)**.** Let U be a topological space of semantic contexts (modules, namespaces, type environments, execution scopes). A *semantic sheaf* $\mathcal{F}$ assigns to each open set $V \subseteq U$ a set of admissible semantic objects $\mathcal{F}(V)$, with restriction maps $\rho_{VW} : \mathcal{F}(V) \to \mathcal{F}(W)$ for $W \subseteq V$, satisfying:

1. *Identity*: $\rho_{VV} = \mathrm{id}$.
2. *Transitivity*: $\rho_{WX} \circ \rho_{VW} = \rho_{VX}$ for $X \subseteq W \subseteq V$.
3. *Gluing*: if $\{V_i\}$ covers V and $s_i \in \mathcal{F}(V_i)$ are mutually compatible on overlaps, there exists a unique global section $s \in \mathcal{F}(V)$ restricting to each s_i.

A semantic merge operation in this framework is a gluing: given two branches with compatible semantic sections over their shared context, the merged result is the unique global section. A merge conflict is a failure of the gluing condition: the sections over the shared context are incompatible, and the system can report exactly which semantic invariant is violated.

This is not merely a more powerful diff algorithm. It is a different representational manifold $\mathcal{M}$: one in which the projection $\pi : \mathcal{X}_{\text{semantic}} \to \mathcal{M}_{\text{sheaf}}$ has smaller fibers, preserving more semantic structure.

Obstruction theory provides the formal language for merge failures: when sections over a cover fail to glue, the obstruction lives in a cohomology group $H^1(U, \mathcal{F})$ that characterizes the nature and location of the incompatibility. Semantic merge conflicts are classes in this cohomology group, not arbitrary text

regions.

RSVP-Inspired Semantic Infrastructure

The RSVP field framework, developed in prior work, provides a concrete implementation architecture for semantic infrastructure. The triple $(\Phi, \mathbf{v}, S)$ of scalar coherence field, vector inference flow, and entropy field maps naturally onto the sheaf-theoretic structure:

Φ tracks semantic coherence across a codebase: high Φ at a location indicates stable, well-defined semantic content; low Φ indicates contested or ambiguous semantics. The gradient $\nabla\Phi$ indicates the direction of increasing coherence, providing a semantic analogue of the gradient flow in thermodynamic systems.

$\mathbf{v}$ tracks inference flow: how semantic information propagates through a codebase from definitions to usages, from types to implementations, from specifications to test cases. The divergence $\nabla \cdot \mathbf{v}$ identifies semantic sources (definition sites) and sinks (consumption sites).

S tracks semantic entropy: the uncertainty or ambiguity in the interpretation of code elements. High S indicates elements with large fibers under the semantic projection: many possible interpretations consistent with the syntactic form. Low S indicates semantically pinned elements.

A semantic merge operation in this framework minimizes S subject to preserving the Φ-coherence and $\mathbf{v}$-consistency of the merged system. Merge conflicts correspond to regions where S cannot be reduced without violating coherence or consistency conditions.

Meaning as Dynamic Topology

The sheaf-theoretic and RSVP-based frameworks share a deeper commitment: meaning is not a static property of code elements but a dynamic structure that evolves as code is developed, deployed, and revised.

A program's semantics at time T is not simply the denotation of its current text but the accumulated semantic trajectory: the history of definitions, revisions, refactorings, and deployments that have shaped the current semantic constraint surface. This is the computational analogue of trajectory-indexed identity.

The consequences for software engineering practice are significant. Version control systems that track only syntactic changes are operating on $\mathcal{M}$ while the relevant computation occurs in $\mathcal{X}$. Systems that track semantic provenance—the causal chain of semantic decisions that led to the current code state—operate on a richer $\mathcal{M}$ with smaller fibers and more faithful constraint preservation.

Digital Geometry and Information Manifolds

Platform Architectures as Fiber Bundles

Digital platform architectures can be analyzed with formal precision using the language of fiber bundles, which provides rigorous structure for phenomena—context collapse, feed-induced trajectory compression, algorithmic path-shaping—that are otherwise described only metaphorically.

Definition 7.1 (Platform Bundle)**.** A *platform bundle* is a triple (E, B, π_B) where:

- B is the *base space* of platform contexts (topics, communities, temporal streams, geographic regions).
- E is the *total space* of content-context pairs: for each context $b \in B$, the fiber $\pi_B^{-1}(b)$ is the space of content that could appear in that context.
- $\pi_B : E \to B$ is the context projection.

A *connection* on this bundle is a rule for lifting paths in B to paths in E: it specifies how content is transported across contexts.

The feed algorithm of a social media platform is precisely a connection on this bundle. It specifies, for each content item and each trajectory through context space, how the item is lifted—displayed, promoted, suppressed, or mutated—as the trajectory moves through different contextual neighborhoods.

Context Collapse as Holonomy Failure

Holonomy is the obstruction to global consistency of parallel transport. Given a connection on a bundle, transport around a contractible loop returns a fiber element to its starting point only if the connection is flat (zero curvature, in the differential-geometric sense). A non-flat connection produces holonomy: transport around a loop returns a fiber element to a different point in the same fiber.

Context collapse is holonomy failure in the platform bundle. A piece of content produced in context b_0 is transported by the algorithm through a sequence of contexts $b_0 \to b_1 \to \cdots \to b_n \to b_0$. If the algorithm's connection has non-trivial holonomy, the content that returns to b_0 is not the content that left: it has accumulated contextual transformation artifacts. A post written for one's professional network is seen by family; a joke among friends is shown to strangers in a different interpretive context; a technical discussion is surfaced in a political feed.

The platform connection has non-trivial holonomy not by accident but by design. Engagement optimization selects for connections that maximize cross-context transport, because content that travels across contextual boundaries generates more impressions. High holonomy connections are engagement-optimal connections. Context collapse is therefore not a side effect of platform design but a structural feature of engagement-optimized connections.

Proposition 7.2 (Engagement and Holonomy)**.** *A connection ∇ on the platform bundle (E, B, π_B) that maximizes engagement (impressions integrated over B) will generically exhibit non-trivial holonomy, since cross-contextual transport generates engagement not available from contextually contained transport.*

Recomposability and Accountability Loss

The holonomy analysis reveals a related phenomenon: the systematic reward of recomposable content over contextually embedded content.

Content with high contextual specificity—content whose meaning depends on knowledge of the production context—has low fiber transportability. Its informational structure resists separation from the context projection. Such content has low holonomy under cross-contextual transport because its meaning changes substantially (or collapses entirely) when lifted to a different fiber.

Content with high recomposability—content whose informational structure is context-independent—has high fiber transportability. It moves across contexts without semantic degradation. It generates engagement across the full base space B rather than within a single contextual fiber.

Engagement-optimized platforms therefore systematically reward recomposable content and penalize contextually embedded content. Over time, this selection pressure produces a cultural ecology in which context-independent signals

(outrage, sentiment, virality, meme-structure) displace context-dependent substance (argument, evidence, relational accountability).

The formal point is that accountability is a contextual property: a claim is accountable to the audience and institutional context in which it was made. Cross-contextual transport severs this accountability link. When a statement is transported across the holonomy group of the platform connection, its accountability structure does not transport with it. The platform systematically degrades the conditions for accountability while optimizing the distribution of its surface appearance.

The sheaf-theoretic language of Section 6 gives this a precise cohomological formulation. A feed platform that cannot sustain global semantic coherence across its communities has non-trivial $\check{H}^1(B, \mathcal{F})$: the local semantic sections (content as interpreted in individual contextual fibers) fail the gluing condition required to assemble a globally consistent meaning. Recomposable content is precisely content whose semantic sections are *trivially* compatible across fibers—not because they cohere, but because they carry so little context-dependent semantic content that no gluing obstruction can arise. Accountability-bearing content fails the platform's implicit gluing test not because it is incoherent but because its coherence is *locally irreducible*: it requires the context fiber to be specified. The platform's selection regime therefore has a structural bias against any content whose meaning is non-trivially sheaf-theoretic.

Semantic Entropy and Manifold Collapse

A further consequence of engagement-optimized connections is semantic entropy increase. As recomposable content displaces contextually embedded content, the representational manifold $\mathcal{M}$ of the platform contracts: the range of distinct semantic structures that the platform can sustain narrows to those compatible with high-transport recomposability.

This is not thought homogenization in the sense of everyone thinking the same thoughts. The underlying trajectory space $\mathcal{X}$ may retain its diversity. What collapses is the channel bandwidth: the range of semantic structures that can be effectively projected into the platform's representational manifold and transported to audiences.

The danger is not cognitive convergence but exploratory route narrowing:

the set of conceptual trajectories that can be pursued through the platform channel shrinks, not because users lack the capacity to pursue others but because the projection system cannot sustain them in transmission.

Predictive Inversion and the Colonization of Navigational Agency

A further and more severe pathology arises when the platform's predictive model of user behavior exceeds the user's own self-model in accuracy. This condition, which we term *predictive inversion*, constitutes a formal threshold beyond which navigational agency in $\mathcal{X}$ is effectively subordinated to the platform's model in $\mathcal{M}$.

Let P_{traj} denote the true distribution over a user's behavioral trajectory in $\mathcal{X}$, P_{platform} denote the platform's learned model of that trajectory, and P_{self} denote the user's own self-model. Predictive inversion occurs when:

$$D_{KL}(P_{\text{platform}} \| P_{\text{traj}}) < D_{KL}(P_{\text{self}} \| P_{\text{traj}}) \tag{10}$$

That is, the platform's model of the user's trajectory has lower divergence from the true trajectory distribution than the user's own self-model does. The platform, operating in $\mathcal{M}$, has acquired a more faithful representation of the user's $\mathcal{X}$-trajectory than the user has of themselves.

This condition has a structural implication that goes beyond privacy violation. Navigational agency—the capacity to steer one's trajectory in $\mathcal{X}$ through deliberate choice—depends on the self-model being a sufficiently accurate representation of one's own constraint surface Σ. An agent navigating by an inaccurate self-model will systematically make choices that diverge from their actual constraint-relevant preferences. When the platform's model is more accurate than the self-model, the platform can predict—and therefore anticipate, pre-shape, and nudge—the trajectory before the agent has resolved the relevant choice.

Proposition 7.3 (Predictive Inversion as Agency Subordination)**.** *When condition* (10) *holds, the platform can systematically present options, surfaces, and affordances that exploit the gap between P_{self} and P_{traj}, steering the user's trajectory in $\mathcal{X}$ through manipulations that are invisible to the user's self-model. The user experiences these steerings as autonomous choices because they are consistent with P_{self}, even though they are systematically biased by $P_{platform}$.*

The critical feature of predictive inversion is that it is architecturally irrecoverable for the user acting within the platform. The gap between P_{self} and P_{traj} is not visible to the user from within $\mathcal{M}$: the user has access only to their own self-model and to the platform's surfaces, not to the divergence between them. The condition cannot be detected by introspection alone. Recovery of navigational agency requires exit from the platform's projection regime or access to the platform's model—neither of which is provided by engagement-optimized architectures.

Predictive inversion therefore constitutes an instance of ontological compression failure applied not to content but to agency itself. The platform has rendered itself capable of interacting with the user's actual trajectory while the user remains confined to interacting with their own reduced self-model. The asymmetry is structural and is a designed consequence of behavioral surplus extraction at scale.

A Taxonomy of Representational Failure Modes

The Central Claim

The preceding sections have developed a unified formal framework for analyzing systems that operate on compressed representations of richer trajectory spaces. The framework now permits the articulation of a central theorem-like claim:

Theorem 8.1 (Representational Failure Theorem)**.** *Modern systemic failure arises generically when a compressed representation space $\mathcal{M}$ is treated as preserving the causally and constraint-relevant structure of the underlying trajectory space $\mathcal{X}$, specifically when $\pi : \mathcal{X} \to \mathcal{M}$ has been chosen such that constraint-relevant fiber structure is irrecoverably erased.*

This is a theorem-like claim in the sense that it follows from the formal structure established in prior sections, subject to the empirical identification of particular systems with the abstract roles of $\mathcal{X}$, $\mathcal{M}$, and π. The identification is in each case an empirical claim, not a formal derivation.

Three Ordered Failure Modes

The framework supports a principled taxonomy of representational failures, ordered by depth and, crucially, by detectability.

Mode I: Simple Error

A representational system commits *simple error* when a representation in $\mathcal{M}$ fails to accurately reflect the corresponding element of $\mathcal{X}$ due to noise, malfunction, or computational mistake.

Formally, simple error is deviation in the image under π: $\hat{m} \neq \pi(x)$ where $\hat{m}$ is the actual representation and $\pi(x)$ is the correct one.

Simple error is detectable *within the representational system*. Given access to x or an independent measure of $\pi(x)$, the error can be identified and corrected. The projection architecture itself remains valid; only a particular instance of its application has failed.

Examples include corrupted sensor data, arithmetic mistakes in computation, typographic errors in textual records. These are the failures that debugging, auditing, and quality control address.

Mode II: Projection Mismatch

Projection mismatch occurs when a representational system preserves the wrong invariants: the projection π has been chosen such that $\mathcal{M}$ accurately reflects some structure of $\mathcal{X}$ but not the constraint-relevant structure.

Formally, projection mismatch is a mismatch between the invariants of π and the invariants of : π is accurate but π $6\perp$ (the projection is not aligned with the constraint conditions).

Projection mismatch is detectable *by comparison with external ground truth*. Given access to independent measurements of , the divergence between $\mathcal{M}$-optimization and -resolution can be detected. Goodhart's Law describes exactly this structure: the metric was once a good proxy for the constraint, but optimization pressure decoupled the proxy from the proxied.

Examples include engagement metrics standing in for intellectual value; test scores standing in for learning; GDP standing in for welfare; publication count

standing in for scientific contribution. In each case, the representation is not inaccurate—it faithfully reflects what it measures—but it measures the wrong thing.

The correction is a change of π: a redesign of the measurement system to better align with . This is possible in principle, though institutionally difficult.

Mode III: Ontological Compression Failure

Ontological compression failure is the deepest mode. It occurs when the reduction $\mathcal{R}$ irrecoverably destroys structure in $\mathcal{X}$ that was causally necessary for future constraint resolution, and when π is designed such that $\mathcal{M}$ cannot represent its own incompleteness.

Formally, ontological compression failure is the combination of:

1. Architectural irrecoverability: no admissible decoder exists that can reconstruct the erased fiber structure.
2. Self-opacity: the interface $\mathcal{M}$ has no representational slot for the absence of what was erased. The system cannot ask the question "what did we lose?" because the answer lives entirely in the erased fiber.

Ontological compression failure is *undetectable from within* $\mathcal{M}$. Unlike simple error, there is no comparison that reveals the mistake; unlike projection mismatch, there is no external ground truth accessible to the representational system. The system operates on its compressed representation without any signal that the compression was destructive.

Proposition 8.2 (Undetectability of Mode III Failures)**.** *A system experiencing ontological compression failure cannot detect this condition through any observation or computation within* $\mathcal{M}$*.*

Proof sketch. By definition, the erased structure has no image in $\mathcal{M}$. Any observation the system can make is an observation in $\mathcal{M}$. The erased structure is therefore invisible to any observation the system can make. □

This explains why ontological compression failures are so persistent and so harmful. They cannot be corrected from within the system; they require an external vantage point capable of observing the relationship between $\mathcal{X}$ and $\mathcal{M}$. Systems that have suffered ontological compression failure typically do not know it.

Examples include: the irreversible flattening of individual identity trajectories to engagement profiles; the decontextualization of traditional knowledge systems through institutional classification; the reduction of complex ecological systems to biodiversity counts; the projection of embodied cultural practices into digital archival formats that preserve content but discard embodied enactment; AI hallucination as the confident production of $\mathcal{M}$-plausible outputs in the absence of the $\mathcal{X}$-grounding that would make them true.

The Detectability Hierarchy

The three failure modes form a strict hierarchy by detectability:

Mode	Detectability	Correction
Simple Error	Within $\mathcal{M}$	Recomputation
Projection Mismatch	From external ground truth	Redesign of π
Ontological Compression Failure	Only from outside $\mathcal{M}$	Requires external vantage

The hierarchy implies that as systems become more comprehensively enclosed within their own representational manifolds—as external ground truth becomes less accessible—the dominant failure mode shifts from Mode I toward Mode III. Closed systems optimizing internal metrics are the natural habitat of ontological compression failure.

Toward an Ethics of Representation

Ethics Derived from Stability Conditions

A natural objection to deriving ethics from formal stability conditions is that stability is a descriptive property of systems, not a normative one. Why should we prefer representations that preserve constraint-relevant structure? The answer cannot be purely normative; it must engage with what representational systems are for.

The answer lies in the relationship between representation and continued operation. A representational system exists to support future constraint resolution: to enable the actions, decisions, and interventions that close the conditions $(x_T) = 0$. A representation that irrecoverably destroys structure necessary for

future constraint resolution degrades the capacity of the system it represents to continue resolving constraints.

Exploitative representational systems are therefore not merely morally deficient; they are structurally parasitic. They extract value from the reduction of $\mathcal{X}$ to $\mathcal{M}$ while externalizing the cost of that reduction onto the constraint-resolution capacity of the degraded trajectory space. The exploitation is thermodynamic before it is moral: it is a transfer of constraint load from the extractor to the extracted-from.

Corollary 9.1 (Structural Parasitism)**.** *A representational system that irrecoverably reduces the constraint-resolution capacity of the trajectory space it operates on is structurally parasitic: it optimizes $\mathcal{M}$-valued objectives while degrading the $\mathcal{X}$-valued conditions required for its own continued legitimacy.*

Remark 9.2. Corollary 9.1 is interpretive rather than mathematically derived in the strict sense: it applies the formal structure of irrecoverable reduction to institutional dynamics where the mapping to $(\mathcal{X}, \mathcal{M}, \mathcal{R})$ is established empirically rather than by construction. The claim is formally motivated but requires domain-specific adequacy conditions to be fully warranted. What is formally precise is the underlying structure: a system that optimizes $\pi(x)$ while irrecoverably erasing S_c-relevant fiber content degrades its own future constraint-closure capacity.

This theorem does real work. It explains why extractive platforms face long-run instability: they degrade the user trajectories whose vitality the platform depends on for continued engagement. It explains why identity-flattening institutions lose the capacity to serve the populations they administrate: they have rendered themselves incapable of engaging with the constraint surfaces they are supposed to support. It explains why surveillance capitalism generates systematic misrepresentation: the projection systems optimized for behavioral extraction are misaligned with the constraint resolutions that constitute genuine user value.

Three Derived Ethical Principles

The formal framework generates three ethical principles for representational system design, each derived from stability conditions rather than asserted normatively.

Principle I: Preserve Reconstructability

A representational system is *ethically well-formed with respect to reconstructability* if, for every reduction it performs, either a decoder $\mathcal{D}$ exists such that $\mathcal{D}(\mathcal{R}(x), s) \approx x$, or the erased fiber structure contains no degrees of freedom causally relevant to future constraint closure for the trajectory x.

This principle distinguishes legitimate compression from destructive reduction. Compilation is legitimate when source code is retained. Archiving is legitimate when retrieval conditions are preserved. Demographic aggregation is legitimate when individual records remain accessible for relevant future queries. Identity reduction is illegitimate when the constraint surface of an individual is irrecoverably compressed in ways that foreclose future constraint resolution. Proposition 4.3 makes this precise: a fidelity-insufficient substrate produces a reduced image $\Sigma_{\mathcal{S}}$ from which the original Σ cannot be recovered, and all subsequent institutional interactions will be with $\Sigma_{\mathcal{S}}$, not with the identity it ostensibly represents.

The principle does not prohibit compression; it requires that compression be designed with foresight about what future constraint resolutions will require.

Principle II: Maintain Provenance

A representational system is *ethically well-formed with respect to provenance* if it records the reduction operator $\mathcal{R}$ and the projection π alongside the representation $m = \pi(x)$.

Provenance maintenance is the disclosure of the reduction process. A metric that records not only its value but also the aggregation procedure, the population sampled, and the constraint conditions it was designed to track provides the side information s required for a decoder. An AI system that records the training data distribution, the objective function, and the architectural choices that generated an output provides partial side information for interpreting that output.

Provenance maintenance is the minimal condition for making Mode I errors detectable and Mode II errors correctable. It does not prevent ontological compression failure, but it creates the conditions for external parties to identify it.

Principle III: Acknowledge Irreversibility

A representational system is *ethically well-formed with respect to irreversibility* if it does not represent irrecoverable reductions as reversible, and does not claim the capacity to reconstruct what it has architecturally excluded.

This principle is violated when: platforms claim to represent users while having discarded the constraint surfaces that constitute user identity; AI systems claim to understand human cognition while operating only on linguistic projections of cognitive trajectories; institutions claim to serve individuals while having rendered those individuals legible only through reductive classification systems.

The acknowledgment of irreversibility is not a counsel of despair. It is the condition for honest design: knowing what has been lost, systems can be designed to minimize irreversible loss where future constraint resolution requires the lost degrees of freedom.

Anti-Extractive Infrastructure

The three principles together define the conditions for *anti-extractive infrastructure*: representational systems designed to preserve constraint-resolution capacity rather than optimizing compressed projections.

Anti-extractive infrastructure is characterized by:

- *Fiber transparency*: the reduction operator $\mathcal{R}$ and its fiber-collapsing behavior are disclosed and auditable.
- *Provenance chains*: each representation carries a record of the reductions that produced it.
- *Irreversibility accounting*: irrecoverable reductions are identified as such and their future constraint costs are estimated.
- *Log sovereignty*: the entities whose trajectories are reduced retain access to encodings of their constraint histories not subordinated to the reducing system.
- *Navigational sovereignty*: systems are designed such that the user's self-model is not systematically less accurate than the platform's behavioral model (Proposition 7.3). Where predictive inversion has occurred, users are entitled to access the platform's model of their own trajectory as a con-

dition of continued legitimate operation.

This is not a utopian program. Reduction is unavoidable; all representation involves compression. Anti-extractive infrastructure does not eliminate reduction; it designs reduction systems that are transparent about what they destroy and that minimize architecturally irrecoverable losses.

Conclusion

Civilization increasingly depends on systems that compress reality into manageable interfaces. These compressions are not optional. The trajectory space $\mathcal{X}$ of any sufficiently complex system is too high-dimensional to be acted upon directly. Some reduction operator $\mathcal{R}$ and some projection π are necessary for any cognitive, computational, or institutional operation.

The argument of this essay is that the choice of π—the decision about which invariants to preserve and which to discard—is the constitutive ethical and epistemological act in the design of representational systems. Every projection discards structure. The question is whether what is discarded was needed.

The formal framework developed here offers three contributions.

First, a generalized reduction operator that unifies apparently disparate phenomena—computational normal forms, thermodynamic equilibria, predictive compression, institutional metrics—under a single formal architecture. This unification reveals that these are not analogies but instances, and that their failure modes therefore share formal structure.

Second, a taxonomy of representational failures distinguished by detectability and recoverability. Simple error is detectable within the representational system. Projection mismatch is detectable from external ground truth. Ontological compression failure is detectable only from outside the representational manifold, and its effects are undetectable from within. This hierarchy explains why the deepest failures are also the most persistent: the system cannot represent its own deficit.

Third, an ethics of representation derived from stability conditions rather than asserted normatively. Exploitative representational systems are not merely morally deficient; they are structurally parasitic, degrading the constraint-resolution capacity of the trajectory spaces they reduce. The ethical principles of recon-

structability, provenance, and irreversibility acknowledgment follow from the conditions for sustainable operation, not from external moral supplement.

The implications reach across every domain where compressed representations are treated as authoritative: artificial intelligence systems that confuse linguistic projection with cognitive grounding; economic systems that optimize metric proxies for welfare; social platforms that degrade identity to engagement profile; institutional bureaucracies that reduce persons to cases; software infrastructure that tracks syntax while semantic provenance decays.

In each case, the formal diagnosis is the same. The projection π has been chosen—deliberately or by default—in ways that discard causally relevant fiber structure. The system then operates on the compressed image $m = \pi(x)$ as though it were the trajectory x itself. The gap between $\mathcal{X}$ and $\mathcal{M}$ is invisible from within $\mathcal{M}$. The system cannot correct what it cannot see.

Intelligence, ethics, computation, and institutions must ultimately be understood as processes operating under irreversible constraint within partially observable spaces. The challenge is not to eliminate constraint or achieve full observability— both are impossible in any complex system—but to design the interface between the observable and the unobservable with sufficient foresight that what is discarded is not what will be needed.

The history of civilization is, in part, the history of choosing projections. The projections we choose determine what we can see, what we can act on, and what we will irrecoverably lose.

Mathematical Appendix

Fiber Bundle Formalism

For readers unfamiliar with fiber bundle theory, a brief summary of the relevant formalism is provided here.

A *fiber bundle* (E, B, F, π) consists of: a total space E, a base space B, a typical fiber F, and a projection $\pi : E \to B$ such that every $b \in B$ has a neighborhood U

for which $\pi^{-1}(U) \cong U \times F$. Intuitively, the bundle is locally a product but may be globally twisted.

A *connection* on a bundle is a consistent rule for horizontally lifting paths in the base B to paths in the total space E. Formally, it is a splitting of the tangent bundle TE into horizontal and vertical subspaces: $TE = H \oplus V$ where $V = \ker(d\pi)$.

Parallel transport along a path $\gamma : [0,1] \to B$ is the horizontal lift of γ: the unique path in E that projects to γ and whose tangent vector lies everywhere in H.

Holonomy at $b \in B$ with respect to a connection is the group of transformations of the fiber $\pi^{-1}(b)$ induced by parallel transport around loops based at b. A flat connection (zero curvature) has trivial holonomy: transport around contractible loops returns every fiber element to itself.

In the platform bundle of Section 7, E is the space of content-context pairs, B is the space of platform contexts, $F = \pi^{-1}(b)$ for a fixed b is the content space in context b, and the feed algorithm defines the connection. Context collapse is non-trivial holonomy: content transported across contextual loops does not return to its original semantic position.

Sheaf Cohomology and Obstruction Theory

A *presheaf* $\mathcal{F}$ on a topological space U assigns to each open set $V \subseteq U$ an abelian group $\mathcal{F}(V)$ (or set, ring, module, etc.) with restriction homomorphisms $\rho_{VW} : \mathcal{F}(V) \to \mathcal{F}(W)$ for $W \subseteq V$.

A presheaf is a *sheaf* if it satisfies the gluing axiom: for any cover $\{V_i\}$ of V and compatible sections $s_i \in \mathcal{F}(V_i)$ (meaning $\rho_{V_i, V_i \cap V_j}(s_i) = \rho_{V_j, V_i \cap V_j}(s_j)$ for all i, j), there exists a unique $s \in \mathcal{F}(V)$ with $\rho_{V, V_i}(s) = s_i$ for all i.

The *Čech cohomology* $\check{H}^n(U, \mathcal{F})$ of a sheaf measures the obstruction to the gluing condition: $\check{H}^0(U, \mathcal{F})$ is the group of global sections; $\check{H}^1(U, \mathcal{F})$ classifies the ways gluing can fail (i.e., the obstruction to finding a global section given compatible local sections).

In the semantic infrastructure of Section 6, U is the space of semantic contexts, $\mathcal{F}$ assigns to each context the set of semantically consistent code states, and a merge conflict is a non-trivial class in $\check{H}^1(U, \mathcal{F})$: the local sections (branch

states) are individually consistent but fail to glue globally.

The Reduction Operator and Fiber Contraction

Let $\mathcal{X}$ be a measurable space with a σ-finite measure μ and $\pi : \mathcal{X} \to \mathcal{M}$ a measurable surjection. The fiber over $m \in \mathcal{M}$ is the pre-image $\pi^{-1}(m) = \pi^{-1}(m)$.

A reduction operator $\mathcal{R} : \mathcal{X} \to \mathcal{X}'$ contracts fibers in the sense that:

$$\mu(\pi^{-1}(m) \cap \mathcal{X}') \leq \mu(\pi^{-1}(m)) \tag{11}$$

for all $m \in \mathcal{M}$, with strict inequality for at least one m.

The *recoverable information* in a reduction is:

$$I_{\text{rec}}(\mathcal{R}) = I(\mathcal{X}; \mathcal{X}' \times S) \tag{12}$$

where S is the space of admissible side information and $I(\cdot;\cdot)$ denotes mutual information. A reduction is recoverable if $I_{\text{rec}}(\mathcal{R}) = H(\mathcal{X})$ (the full entropy of the original space is recoverable); irrecoverable if $I_{\text{rec}}(\mathcal{R}) < H(\mathcal{X})$ regardless of the choice of S within admissible bounds.

The degree of irrecoverability:

$$\Delta I(\mathcal{R}) = H(\mathcal{X}) - I_{\text{rec}}(\mathcal{R}) \tag{13}$$

quantifies the structural information permanently lost by the reduction. Ontological compression failure corresponds to $\Delta I(\mathcal{R}) > 0$ for degrees of freedom in $\mathcal{X}$ that are causally relevant to future constraint closure $(x_T) = 0$.

References

[1] Claude E. Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, 1948.

[2] Alonzo Church, *An Unsolvable Problem of Elementary Number Theory*, American Journal of Mathematics, vol. 58, no. 2, pp. 345–363, 1936.

[3] Alan M. Turing, *On Computable Numbers, with an Application to the Entscheidungsproblem*, Proceedings of the London Mathematical Society, vol. 42, no. 1, pp. 230–265, 1936.

[4] Alonzo Church and J. Barkley Rosser, *Some Properties of Conversion*, Transactions of the American Mathematical Society, vol. 39, no. 3, pp. 472–482, 1936.

[5] Karl Friston, *The Free-Energy Principle: A Unified Brain Theory?*, Nature Reviews Neuroscience, vol. 11, no. 2, pp. 127–138, 2010.

[6] Karl Friston, *A Free Energy Principle for Biological Systems*, Entropy, vol. 11, no. 3, pp. 293–301, 2009.

[7] Charles Goodhart, *Problems of Monetary Management: The U.K. Experience*, in *Papers in Monetary Economics*, Reserve Bank of Australia, 1975.

[8] Donald T. Campbell, *Assessing the Impact of Planned Social Change*, Dartmouth College Public Affairs Center, 1976.

[9] David Graeber, *Bullshit Jobs: A Theory*, Simon & Schuster, 2018.

[10] Donald MacKenzie, *An Engine, Not a Camera: How Financial Models Shape Markets*, MIT Press, 2006.

[11] N. Katherine Hayles, *How We Became Posthuman*, University of Chicago Press, 1999.

[12] Martin Heidegger, *The Question Concerning Technology*, in *The Question Concerning Technology and Other Essays*, Harper & Row, 1977.

[13] Maurice Merleau-Ponty, *Phenomenology of Perception*, Routledge, 1962.

[14] Andy Clark, *Being There: Putting Brain, Body, and World Together Again*, MIT Press, 1997.

[15] Francisco J. Varela, Evan Thompson, and Eleanor Rosch, *The Embodied Mind*, MIT Press, 1991.

[16] Evan Thompson, *Mind in Life*, Harvard University Press, 2007.

[17] Robert Rosen, *Anticipatory Systems*, Pergamon Press, 1985.

[18] Stuart A. Kauffman, *The Origins of Order*, Oxford University Press, 1993.

[19] W. Ross Ashby, *An Introduction to Cybernetics*, Chapman & Hall, 1956.

[20] Norbert Wiener, *Cybernetics: Or Control and Communication in the Animal and the Machine*, MIT Press, 1948.

[21] Gregory Bateson, *Steps to an Ecology of Mind*, University of Chicago Press, 1972.

[22] Gilles Deleuze, *Difference and Repetition*, Columbia University Press, 1994.

[23] Gilbert Simondon, *L'individuation psychique et collective*, Aubier, 1989.

[24] Bruno Latour, *Reassembling the Social*, Oxford University Press, 2005.

[25] Karen Barad, *Meeting the Universe Halfway*, Duke University Press, 2007.

[26] James Gleick, *Chaos: Making a New Science*, Viking, 1987.

[27] Ilya Prigogine and Isabelle Stengers, *Order Out of Chaos*, Bantam Books, 1984.

[28] Rolf Landauer, *Irreversibility and Heat Generation in the Computing Process*, IBM Journal of Research and Development, vol. 5, no. 3, pp. 183–191, 1961.

[29] Andrey N. Kolmogorov, *Three Approaches to the Quantitative Definition of Information*, Problems of Information Transmission, vol. 1, no. 1, pp. 1–7, 1965.

[30] Ray Solomonoff, *A Formal Theory of Inductive Inference*, Information and Control, vol. 7, no. 1, pp. 1–22, 1964.

[31] Judea Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2009.

[32] Michael Spivak, *A Comprehensive Introduction to Differential Geometry*, Publish or Perish, 1999.

[33] Allen Hatcher, *Algebraic Topology*, Cambridge University Press, 2002.

[34] Glen E. Bredon, *Sheaf Theory*, Springer, 1997.

[35] Saunders Mac Lane, *Categories for the Working Mathematician*, Springer, 1998.

[36] Steve Awodey, *Category Theory*, Oxford University Press, 2010.

[37] Robert Ghrist, *Elementary Applied Topology*, Createspace, 2014.

[38] Russell T. Hurlburt and Eric Schwitzgebel, *Describing Inner Experience?*, MIT Press, 2007.

[39] Adam Zeman et al., *Lives Without Imagery: Congenital Aphantasia*, Cortex, vol. 73, pp. 378–380, 2015.

[40] Ned Block, *On a Confusion About a Function of Consciousness*, Behavioral and Brain Sciences, vol. 18, no. 2, pp. 227–247, 1995.

[41] Randall D. Beer, *Dynamical Approaches to Cognitive Science*, Trends in Cognitive Sciences, vol. 4, no. 3, pp. 91–99, 2000.

[42] Adam Frank, Marcelo Gleiser, and Evan Thompson, *The Blind Spot: Why Science Cannot Ignore Human Experience*, MIT Press, 2025.