From Persuasion to Realization

Constraint Closure and the Geometry of Inference

Flyxion

Independent Researcher

April 10, 2026

Abstract

The dominant paradigm of machine intelligence produces systems that generate locally coherent, persuasive discourse without guaranteeing global realizability. Drawing on Gunkel's account of large language models as persuasive machines in the rhetorical tradition [1], this monograph argues that the pathologies of such systems—hallucination, drift, and structural incoherence—are not incidental failures but structural consequences of a categorical misframing of inference. The problem is not that language models err; it is that the objective they optimize admits no criterion of global consistency.

We replace this paradigm with a theory of inference as constraint closure: the identification of world-states that satisfy all observational, physical, and structural constraints simultaneously. Formally, hallucination is characterized as a nontrivial class in the first Čech cohomology group $H^1(\Omega, \mathcal{S})$ of the admissibility space, which we formalize as a stratified simplicial complex. This identification implies that hallucination is a topological invariant of the generative architecture, not a statistical defect removable by scaling. We prove a No-Closure Proposition establishing that no family of generative models can guarantee the vanishing of this cohomology class regardless of scale, since scaling improves approximation to the support of discourse but leaves the absence of global descent conditions structurally intact.

The Constraint-Closure Theorem establishes the conditions under which inference produces a unique globally consistent world-state: vanishing cohomology, projection injectivity, and energy regularity. The Relaxed Constraint-Closure Theorem extends this to practical systems with finite compute, establishing that ε-closed systems maintain bounded deviation from realizability while generative systems admit unbounded error. A minimal counterexample demonstrates, in the simplest sheaf-theoretic setting, that locally valid sections need not glue to a global section—a constructive proof that hallucination is structural rather than statistical. The implications extend from epistemology, where truth becomes an existential condition enforced by fixed-point convergence rather than a correspondence predicate, to social infrastructure, where rhetorical literacy proves insufficient and structural guarantees at the architectural level are required.

Contents

Part I

The Crisis of Persuasive Machines

1 The Emergence of Rhetorical AI

Every powerful technology carries within its design an implicit theory of what it is for. The telegraph was for transmission; the calculator was for arithmetic; the search engine was for retrieval. Each of these framings determined the architecture, and the architecture in turn shaped what the technology could and could not do. Large language models are no different, except that their implicit theory is one that has been largely unexamined: they are built for persuasion.

This is not a pejorative characterization. Persuasion, in the classical sense, is a sophisticated cognitive achievement. Aristotle distinguished it from demonstration precisely because it operates under conditions of probability and audience: a persuasive argument is one that leads a reasonable person toward conviction, not one that compels agreement by logical necessity alone [4]. The rhetorical tradition from which this distinction descends understood that most human discourse operates in the domain of the probable, not the necessary. To speak persuasively is to navigate this domain with skill.

Large language models have become extraordinarily skilled navigators of this domain. They produce text that reads as authoritative, as internally consistent within any given window, and as responsive to the apparent intent of a query. Gunkel [1] characterizes this capacity as a realization of the rhetorical ideal: systems that generate discourse organized not around truth but around the production of persuasive effect. This framing, which draws on the long philosophical tradition connecting rhetoric to simulation and appearance, is more precise than the common engineering description of these systems as statistical next-token predictors. Statistics is the mechanism; rhetoric is the function.

The historical lineage matters here. Plato's dialogues, particularly the *Gorgias* and the *Phaedrus*, stage a sustained confrontation between philosophy and rhetoric that turns on exactly this distinction [2, 3]. For Plato, the rhetor produces belief without knowledge, conviction without understanding, agreement without truth. The philosopher, by contrast, pursues a form of discourse that tracks reality rather than manipulating appearances. This distinction was contested throughout antiquity—Aristotle's rehabilitation of rhetoric as a legitimate cognitive practice, Cicero's pragmatic integration of eloquence and wisdom—but the underlying tension remained: is discourse oriented toward reality, or toward reception?

The emergence of large language models forces this question out of the academy and into engineering practice. These systems are trained on human discourse—which is itself predominantly rhetorical in Plato's sense, organized around persuasion, performance, and social effect—and their success metrics are operationalizations of rhetorical effectiveness: human raters preferring outputs that sound confident, coherent, and relevant. The result is a technology that has optimized the rhetorical function with extraordinary efficiency,

without ever having been constrained to orient itself toward what is actually the case.

To understand why this matters, and why it constitutes a crisis rather than merely an engineering limitation, we need to examine not just the outputs of these systems but the geometry of their failure.

2 Hallucination Reframed: From Error to Underconstraint

The concept of hallucination, as applied to language models, has accumulated a misleading set of connotations. It suggests a random deviation from normal functioning, a noise event superimposed on an otherwise reliable system. The model usually gets things right and occasionally gets things wrong in ways that look like confabulation. This framing leads to interventions aimed at reducing hallucination rates—better training data, reinforcement from human feedback, retrieval augmentation—all of which treat the phenomenon as a defect to be minimized rather than a symptom to be understood.

The correct diagnosis is structural underconstraint. A language model, at any given generation step, is computing a probability distribution over tokens conditional on context. This distribution encodes, in some compressed sense, the statistical regularities of the training corpus. What it does not encode is any requirement that the resulting sequence correspond to a globally realizable state of affairs. The model has no access to, and no representation of, a world that its outputs are supposed to describe. It has access only to the local consistency of discourse—the fact that certain sequences of tokens tend to follow others in the human texts it was trained on.

To make this precise, consider the distinction between local admissibility and global realizability. A sequence of tokens is locally admissible if it coheres with its immediate context in the sense captured by the model's training distribution. It is globally realizable if there exists an actual state of the world—a configuration of objects, relations, and processes—that the sequence correctly describes. These are entirely different conditions. Local admissibility is a property of text; global realizability is a property of the relationship between text and world. A system optimized entirely for local admissibility has no mechanism for enforcing global realizability.

This is not merely a philosophical point. It has a precise mathematical formulation. Let $\mathcal{M}$ denote the manifold of locally admissible sequences—roughly, the support of the model's learned distribution over text. Points on $\mathcal{M}$ are locally coherent in the sense that each token assignment is consistent with its neighbors. But $\mathcal{M}$ is not, in general, a subset of any manifold of globally realizable descriptions. There are points on $\mathcal{M}$—entire essays, entire technical explanations, entire biographical sketches—that are locally smooth and globally impossible. They describe no state of the world that has ever existed or could exist.

Hallucination, on this account, is not noise. It is the system visiting regions of $\mathcal{M}$ that happen to be disjoint from the set of globally realizable descriptions. The system is not malfunctioning; it is doing exactly what it was designed to do. The malfunction is in the design.

3 The Collapse of the Truth–Rhetoric Distinction

The rhetorical tradition has always maintained a distinction, however unstable, between discourse that tracks truth and discourse that manufactures conviction. This distinction provided the normative anchor for epistemology: truth-oriented discourse was philosophy, science, testimony from reliable witnesses; conviction-manufacturing discourse was rhetoric, sophistry, propaganda. The emergence of large language models has destabilized this distinction in ways that demand theoretical attention.

The destabilization operates at two levels. The first is that language models make visible what was always true about human discourse: the mechanisms of local coherence and rhetorical persuasion are not confined to bad-faith or manipulative speech. They are constitutive of language itself. Austin's account of speech acts shows that language does not merely describe but performs [5]; Searle extends this into a general theory of the illocutionary structure of discourse [6]. Habermas's theory of communicative action reveals that even truth-oriented discourse operates through pragmatic norms of comprehensibility, sincerity, and appropriateness that have a rhetorical dimension [8]. The Platonic fantasy of a purely epistemic discourse untainted by rhetoric was always a fantasy. Language models, by producing rhetorical discourse at high quality without any epistemic orientation whatsoever, expose the degree to which what we took to be knowledge-conveying speech was in part rhetorical performance.

The second level of destabilization is more directly consequential. If human discourse is itself partially rhetorical, and if language models can replicate its rhetorical features without its epistemic content, then the distinction between genuine knowledge-conveying speech and high-quality imitation becomes difficult to maintain from the outside. A reader interacting with a language model output cannot, in general, determine from the text alone whether it is globally realizable or merely locally admissible. The markers that normally signal reliability—confident tone, internal consistency, citation of specifics, appropriate hedging—are all learnable features of training text, and all can be reproduced by a system with no access to what is actually the case.

This is the sense in which Gunkel is right to characterize the situation as a crisis [1]. It is not that language models lie. Lying requires knowing the truth and deliberately departing from it. Language models have no access to truth to depart from. The situation is more structurally disorienting: we have built systems that produce the surface features of knowledge-conveying discourse without the substance, and we have done so by training them on a corpus in which those surface features were, statistically, reliably correlated with the substance—because the human authors of that corpus were, by and large, trying to say true things. The correlation has been inverted: we have extracted the form and discarded the content.

4 Rhetorical Equilibrium as an Attractor State

To understand the dynamics that produce this situation, it is useful to introduce the concept of rhetorical equilibrium. Training a language model on human text, with human

preference as the reward signal, is an optimization process that converges to a fixed point. That fixed point is not, in general, the system that produces the most accurate descriptions of the world. It is the system that produces the outputs most preferred by human raters, which means, in practice, the outputs that read as most authoritative, most coherent, and most responsive.

We can formalize this as follows. Let $\mathcal{P}$ denote a distribution over human preferences, and let f_θ denote a language model with parameters θ. Training under preference optimization drives θ toward a parameter configuration θ^* that maximizes $\mathbb{E}_{x\sim\mathcal{P}}[\text{preference}(f_\theta(x))]$. This is a well-defined optimization problem, and under standard conditions it has a solution. The solution θ^* defines a system at rhetorical equilibrium: the distribution of outputs it produces has been shaped by the attractor structure of human preference, and no small perturbation of θ^* in the direction of accuracy rather than persuasiveness will improve the objective.

Rhetorical equilibrium is stable but it is not a fixed point of truth. The system has been optimized to produce text that persuades, and persuasion is a function of the reader's prior expectations and the statistical regularities of the training corpus—not a function of the world being described. The system is caught in a stable attractor from which there is no gradient-descent path toward accuracy, because accuracy was never in the objective. Recognizing this is the precondition for the theoretical reorientation this monograph proposes.

Part II

Geometry of Admissibility and Failure

5 The Admissibility Manifold

We now develop a more precise geometric picture of the space within which current language models operate. The central object is the admissibility manifold $\mathcal{M}$, defined as the set of token sequences to which the trained model assigns non-negligible probability. This is not a manifold in the strict differential-geometric sense, but it has manifold-like local properties: small perturbations of a highly probable sequence tend to remain highly probable, and the local geometry around any given point reflects the statistical structure of the surrounding discourse.

The admissibility manifold is dense across the range of human discourse. A model trained on the breadth of human text learns a distribution supported across vast topic ranges, registers, and argumentative structures. The result is a manifold whose topology is complex—it has many connected components corresponding to distinct discourse genres, many saddle points and ridges corresponding to transitions between styles—but which is locally smooth almost everywhere.

What the admissibility manifold lacks is any intrinsic connection to a space of realizable world-states. A point on $\mathcal{M}$ is a sequence of tokens that coheres statistically with the

training distribution. It may describe a real situation, a fictional one, a possible one, or an impossible one; the manifold itself makes no distinction. The model's probability assignments encode the frequency with which certain sequences appeared in training text, and that frequency reflects human discourse conventions, not metaphysical necessity. The topology of $\mathcal{M}$ can be partially characterized by studying the model's internal representations, but what those representations cannot provide is information about which regions of $\mathcal{M}$ are globally realizable and which are not. That distinction requires a constraint structure that is absent from the current architecture.

6 Local Sections Without Global Glue

The sheaf-theoretic framework provides the right language for articulating the structural gap between local coherence and global consistency. A sheaf over a topological space assigns to each open set a collection of sections—local data—together with restriction maps specifying how sections over larger sets restrict to sections over smaller ones, and a gluing condition specifying when locally compatible sections can be assembled into a global section [22].

In the language model setting, the domain is discourse space and the sheaf assigns to each local window—a context window, a paragraph, a topic region—the set of locally admissible completions. The restriction maps correspond to the consistency requirements between nested contexts. This condition is approximately satisfied by well-trained language models, which maintain consistency across scales in their local predictions.

The gluing condition is where the architecture fails. A sheaf satisfies the gluing condition if whenever we have locally compatible sections over a cover $\{U_i\}$ of a space Ω, there exists a unique global section over Ω that restricts to each local section on its domain. In the language model setting, the gluing condition would require that locally coherent utterances across different regions of a discourse can always be assembled into a globally coherent whole. This condition is not enforced and, in general, fails. The failure is systematic because the model generates each local section without access to the global constraint that all sections must come from a common world-state.

7 The Absence of Cocycle Conditions

In sheaf theory, the obstruction to gluing is measured by Čech cohomology. Given a cover $\{U_i\}$ and locally compatible sections $\{s_i\}$ over each U_i, the sections glue to a global section if and only if the defects $\delta_{ij} = \rho_{ij}(s_i) - \rho_{ji}(s_j)$ vanish on all overlaps $U_i \cap U_j$ and the resulting cochain satisfies the cocycle condition on triple overlaps [23]. When the cocycle condition fails, the local sections cannot be assembled into a global one regardless of how individually coherent they may be.

The architecture of current language models contains no mechanism that enforces the cocycle condition. Token generation is a sequential process in which each step is conditioned on previous tokens and the original prompt. This conditioning enforces local consistency—sequential coherence within the context window—but it does not enforce the

global compatibility condition corresponding to the cocycle condition. The context window is too short to carry the information required for global consistency checks, and there is no global state to check against in any case.

The consequence is systematic: language models produce sequences in which different parts are locally coherent but globally incompatible. A technical explanation may contradict itself across sections that do not appear in the same context window. A biographical account may attribute different birth years to the same person in different passages. A mathematical argument may use notation inconsistently across a long document. These are not random errors; they are the systematic consequence of generating local sections without enforcing the cocycle condition that would guarantee they come from a common global section. No amount of scaling, no increase in context window size, no improvement in local coherence will resolve this failure, because the failure is structural.

8 H^1 Obstructions and Homological Tears

The precise mathematical content of hallucination is a nontrivial class in the first Čech cohomology group $H^1(\Omega, \mathcal{S})$. When this cohomology is nontrivial, there exist collections of locally compatible sections that cannot be glued into a global section. Each such collection corresponds to what we term a *homological tear*: a region of discourse that is locally consistent but globally unrealizable.

To make this concrete, consider a language model generating a detailed scientific explanation. Each paragraph is locally admissible—each sentence follows naturally from the previous one and coheres with the general topic. But the explanation as a whole describes a mechanism that violates known physical laws in ways that only become apparent when the full text is examined against external constraints. The individual paragraphs correspond to local sections; the external physical constraints correspond to the global sheaf conditions; and the failure to satisfy those conditions corresponds to a nontrivial class in H^1.

The homological perspective reveals something important about the geography of failure. H^1 obstructions are not point defects; they are topological invariants of regions. A homological tear cannot be removed by local modifications to the sections involved; it requires a change in the global structure. This is why post-hoc fact-checking does not solve the hallucination problem: it can identify and correct specific errors, but it cannot transform a system that generates locally admissible sequences into one that generates globally realizable ones. The tear is in the architecture, not in the output. Different domains exhibit characteristic homological tear patterns: in scientific discourse, tears arise at the intersection of locally plausible causal claims that are jointly inconsistent; in historical narrative, they arise when locally coherent accounts cannot be reconciled with documented sequences; in mathematical reasoning, they arise when locally valid proof steps are assembled into arguments that prove false theorems.

9 Projection Consistency and Its Failure

We can sharpen the geometric picture by introducing projection operators. Let X denote the true state of affairs in some domain, and let $\Pi_i : \mathcal{X} \to \mathcal{Y}_i$ denote a family of projection operators that map world-states to their representations in different observational modalities or discourse registers. Each projection $y_i = \Pi_i(X)$ is a partial, lossy description of X from a particular vantage point.

A collection of observations $\{y_i\}$ is projection-consistent if there exists a world-state X such that $\Pi_i(X) = y_i$ for all i. This is a strong condition: it requires that all the observations could simultaneously be true of a single state of the world. Current language models satisfy a much weaker condition: they produce outputs that are consistent with the marginal distribution of each y_i separately, without ensuring that the joint collection is consistent with any single X.

Projection degeneracy arises when the projection operators Π_i are not jointly injective: distinct world-states $X_1 \neq X_2$ produce identical projections across all modalities, making it impossible to distinguish them from observations alone. In the language model setting, projection degeneracy is ubiquitous. Vast numbers of world-states produce text that looks the same to the model, because the model has no access to the world-states themselves, only to their textual projections in the training corpus.

The failure of projection consistency is the formal statement of the hallucination problem. A hallucinating system produces outputs $\{y_i\}$ that are not jointly consistent with any world-state X: there is no X such that $\Pi_i(X) = y_i$ for all i. The outputs may each be locally plausible, may even be accurate in isolation, but their conjunction is globally unrealizable. This is not a property of the outputs in themselves but of their relationship to the constraint structure of reality.

10 The Admissibility Space as a Stratified Simplicial Complex

The notion of an admissibility manifold has thus far served as an intuitive geometric description of the space explored by generative models. However, this object cannot, in general, be endowed with the structure of a smooth manifold. The domain of discourse is not homogeneous: linguistic registers, semantic domains, and inferential regimes exhibit discontinuities that preclude a globally smooth atlas. Formalizing this object correctly is not a terminological refinement; it is what makes hallucination a rigorous topological phenomenon rather than a rhetorical flourish.

We formalize the admissibility space $\mathcal{M}$ as a *stratified simplicial complex*. Each stratum $\mathcal{M}_\alpha$ corresponds to a domain of local coherence—a discourse genre, semantic field, or constraint regime—within which local sections can be consistently defined. These strata are glued along lower-dimensional faces that encode transitions between regimes: the saddle points and ridges of the admissibility space, where styles, registers, and referential frames shift.

Formally, let $\mathcal{M} = \bigcup_{\alpha \in A} \mathcal{M}_\alpha$ where each $\mathcal{M}_\alpha$ is a simplicial complex equipped with a local sheaf $\mathcal{S}_\alpha$. The intersections $\mathcal{M}_\alpha \cap \mathcal{M}_\beta$ define overlap regions where transition maps

must satisfy compatibility conditions. The failure of these conditions is precisely what gives rise to nontrivial cohomology.

Within this structure, local admissibility corresponds to the existence of a section $s_\alpha \in \mathcal{S}_\alpha(\mathcal{M}_\alpha)$ satisfying local constraints. Global realizability requires the existence of a collection $\{s_\alpha\}$ satisfying the cocycle condition on all overlaps:

$$\rho_{\alpha\beta}(s_\alpha) = \rho_{\beta\alpha}(s_\beta) \quad \text{on } \mathcal{M}_\alpha \cap \mathcal{M}_\beta.$$

A hallucination is then rigorously identified as a nontrivial cohomology class in $H^1(\mathcal{M}, \mathcal{S})$, representing an obstruction to gluing local sections into a global section.

This reformulation replaces the informal notion of a manifold of plausibility with a precise topological object capable of encoding discontinuities, regime shifts, and structural incompatibilities. It also clarifies the fundamental limitation of scaling: improving statistical density within a stratum does not resolve obstructions that arise at the level of overlaps. These obstructions are topological, not probabilistic, and therefore require structural rather than statistical remedies. A larger model is a more detailed atlas of $\mathcal{M}$; it is not a repair of the missing descent data between charts.

11 Computable Cohomology via the Defect Operator

While the identification of hallucination with nontrivial cohomology provides a powerful conceptual framework, it must be rendered computationally tractable to serve as the basis of an architecture. We introduce the *defect operator*, which operationalizes the cocycle condition within the TARTAN tiling structure.

Let $\{U_i\}$ be a cover of the admissibility space, and let $s_i \in \mathcal{S}(U_i)$ denote the local section associated with tile U_i. On each overlap $U_i \cap U_j$, define the defect

$$\delta_{ij} = \rho_{ij}(s_i) - \rho_{ji}(s_j).$$

The collection $\{\delta_{ij}\}$ defines a 1-cochain measuring the failure of local consistency. The cocycle condition is satisfied if and only if $\delta_{ij} = 0$ for all overlaps.

We define the *defect operator* $\mathcal{D}$ as the mapping

$$\mathcal{D}(\{s_i\}) = \{\delta_{ij}\}_{i,j},$$

which lifts a collection of local sections into the space of 1-cochains. The magnitude $\|\mathcal{D}(\{s_i\})\|$ provides a quantitative measure of structural inconsistency that is directly computable from the local states.

The reconstruction problem can then be formulated as a minimization over both local states and the defect cochain:

$$\min_{\{s_i\}} \sum_{i,j} \|\delta_{ij}\|^2 + \lambda \mathcal{R}(\{s_i\}),$$

where $\mathcal{R}$ encodes additional dynamical, observational, and structural constraints. This

formulation transforms the abstract condition of vanishing cohomology into a concrete optimization objective. The descent dynamics of the TARTAN architecture can now be understood as iteratively reducing the norm of the defect operator until a consistent cocycle is achieved and global section construction becomes possible.

Crucially, this formulation enables the classification of defects that drives the repair hierarchy. Gauge defects arise from coordinate mismatches and can be resolved by reparametrization of local sections without altering their underlying content. Resolution defects arise from insufficient granularity in the tiling and can be resolved by refining the cover, subdividing tiles until the residual defect falls below tolerance. Structural defects persist under both reparametrization and refinement; they indicate a genuine topological obstruction in $H^1(\mathcal{M}, \mathcal{S})$ and require extension of the state space—the introduction of new variables or fields into the ontology—to eliminate. Čech cohomology thus becomes not merely diagnostic but an active driver of the inference engine, guiding repair, refinement, and ontological extension within a unified computational framework.

Part III

From Prediction to Reconstruction

12 The Category Error of Next-Token Prediction

The architectural commitment to next-token prediction is a commitment to a particular theory of what inference is. On this theory, inference is the computation of $P(\text{next token} \mid \text{context})$: the conditional probability of the next symbol given what has come before. This is a coherent mathematical object, well-defined and computable under standard assumptions. But it is the wrong object for the problem we actually face.

The error is categorical in the philosophical sense: it mistakes a proxy for the thing itself. Next-token prediction is a useful proxy for a range of linguistic capacities, including fluency, topical coherence, and stylistic appropriateness. It is not a proxy for truth, because truth is a relation between language and the world, and the next-token prediction objective has no world in it. The world appears in the training corpus only as reflected in text, and the model learns to predict text, not to track the world that text reflects.

The point can be made formally. Observations $y_i = \Pi_i(X)$ are projections of a world-state X; they are not independent random variables but constrained shadows of a common source. A system trained to predict y_i given context is learning $P(y_i \mid \text{context})$, which is a marginal distribution. What we need for truth-tracking is a system that reasons about X itself, the latent world-state that all observations are projections of. These are fundamentally different inference problems. The first is solvable by training on text; the second requires access to constraints that specify which world-states are admissible, which projections are consistent, and which collections of observations can be jointly realized.

13 Inference as Fixed-Point Computation

The correct formulation of inference, on the account we are developing, is not next-token prediction but fixed-point computation. The inference problem is to identify a world-state X^* that satisfies a consistency operator $\mathcal{C}$, meaning $\mathcal{C}(X^*) = X^*$. The consistency operator encodes all the constraints the world-state must satisfy: observational constraints derived from the available projections, physical constraints derived from the dynamics governing the domain, and structural constraints derived from the logical and mathematical relations that hold among the quantities being described.

This formulation transforms inference from a generative act into a reconstructive one. The system does not produce an output by sampling from a learned distribution; it identifies a configuration that satisfies all its constraints simultaneously. The difference is not merely terminological. In the generative paradigm, there is no notion of global consistency: each sample is a point on the admissibility manifold, and whether that point corresponds to a realizable world-state is not part of the objective. In the reconstructive paradigm, global consistency is the objective: the system is looking for a fixed point of $\mathcal{C}$, and such a fixed point, by definition, satisfies all constraints.

The dynamical picture is also transformed. In the generative paradigm, inference is a one-pass computation: given context, produce output. In the reconstructive paradigm, inference is an iterative process

$$X_{t+1} = \mathcal{C}(X_t),$$

where each application of $\mathcal{C}$ reduces inconsistency and drives the candidate state toward a globally admissible configuration. Convergence is not guaranteed in general, but when it occurs, it certifies that the resulting state X^* is a fixed point of the constraint system—a configuration from which no available constraint pushes it away. An answer is thus characterized by its relationship to a constraint structure, and the quality of an answer is measured by how well it satisfies that structure.

14 Truth as an Existential Condition

The fixed-point account of inference supports a reconceptualization of truth that is more tractable than the traditional philosophical alternatives. Correspondence theories of truth require a primitive relation between language and world that is difficult to formalize and impossible to verify computationally. Coherence theories of truth require only internal consistency, which does not distinguish between coherent fiction and coherent fact. Both approaches fail to capture what we actually care about when we ask whether an AI system is telling the truth.

The constraint-based account we are developing suggests a different formulation. Truth, for a representational system, is an existential condition: a statement or description is true if and only if there exists a globally consistent world-state X^* that it correctly describes. This is not trivially equivalent to correspondence, because it makes the existence of the world-state the criterion rather than the correspondence relation itself. And it is not equivalent to coherence, because coherence with other statements is necessary but not

sufficient: the statements must collectively be realizable, which is a stronger condition than mutual consistency in a purely logical sense.

On this account, a hallucinating language model is not producing false statements in the sense of asserting p while knowing $\neg p$. It is producing statements that fail the existential condition: there is no world-state in which what it says is jointly true. The failure is not dishonesty but non-realizability. The system has generated a locally admissible sequence that does not sit over any globally consistent state.

15 The Elimination of Rhetorical Drift

A constraint-closed system eliminates rhetorical drift by construction. Rhetorical drift is the tendency of a discourse-generating system to migrate, through locally plausible steps, toward outputs that are persuasive but globally unrealizable. Each step seems reasonable given the local context; the drift emerges from the accumulation of steps that each preserve local admissibility while eroding global consistency.

In a system organized around fixed-point convergence, rhetorical drift is structurally impossible. The consistency operator $\mathcal{C}$ penalizes configurations that violate constraints, and the iterative dynamics drive the candidate state away from unrealizable configurations rather than toward them. There is no gradient toward persuasion in the objective; there is only a gradient toward consistency. Outputs that would be rhetorically effective but globally unrealizable cannot arise as fixed points of the constraint system.

This is more than a desirable property of a hypothetical future architecture. It is the clearest argument for why the architectural shift from prediction to reconstruction is not an incremental improvement but a categorical one. No amount of fine-tuning, instruction following, or preference optimization can introduce this property into a generative system, because the property requires that the objective function itself select for global realizability rather than local plausibility. The only way to build a system that cannot produce globally unrealizable outputs is to build a system whose convergence criterion is global realizability.

Part IV

The TARTAN Architecture

16 Trajectory-Aware Recursive Tiling

The TARTAN architecture—Trajectory-Aware Recursive Tiling with Annotated Noise—is a constructive realization of the descent-based inference paradigm developed in the preceding sections. Its central innovation is the representation of the domain not as a monolithic space to be processed in a single forward pass, but as a recursively structured cover of overlapping tiles, each of which carries a local state and an admissibility certificate, with global consistency enforced through the iterative elimination of defects on overlaps.

The recursive tiling structure addresses a fundamental challenge of constraint-based inference: the computational cost of enforcing global consistency scales with the size of

the state space, but the constraint violations that signal inconsistency are often local. By representing the domain as a hierarchy of tiles at different resolutions, the system can direct computational effort toward the regions where constraints are most actively violated, leaving stable regions undisturbed. Coarse tiles capture large-scale structure with low computational cost, while fine tiles resolve localized discrepancies through adaptive refinement.

Trajectory awareness is the second key component. The domains in which AI systems operate are intrinsically dynamical: the world has a history, and the constraints that govern it include not only instantaneous consistency conditions but also equations of motion that connect states at different times. A trajectory-aware system maintains representations of how each local state has evolved and projects that evolution forward to check consistency with current observations and future constraints. This dramatically increases identifiability: many configurations that are statically indistinguishable become distinguishable once their dynamical signatures are taken into account.

17 Annotated Noise and Semantic Perturbations

The noise structure of the TARTAN architecture is not an engineering inconvenience to be minimized but a semantically structured component of the inference process. The annotation of noise refers to the practice of tracking the origin, character, and constraint implications of each source of uncertainty, rather than treating all uncertainty as equivalent and marginalizing over it.

Different sources of uncertainty have different implications for constraint satisfaction. Measurement noise in a physical sensor is Gaussian and uncorrelated; it does not signal a structural inconsistency in the world-state being measured. Inconsistency between two competing accounts of the same event is not noise in this sense; it signals a genuine constraint violation that must be resolved by finding a world-state consistent with both accounts, or by classifying one account as a sensor malfunction. Confounds in causal reasoning introduce structured uncertainty that interacts with the causal constraints of the field equations in specific ways.

By annotating noise with its structural character, the system can propagate it appropriately through the constraint structure. Gaussian measurement noise reduces the precision with which constraints can be enforced but does not change their character. Structural inconsistency noise triggers the repair machinery and may lead to state extension. Adversarial or maliciously structured noise activates the defensive inference branch, which attempts to model the noise as a corrupted sensor rather than as a genuine feature of the world-state. The semantic perturbation perspective also clarifies the relationship between TARTAN and generative models: whereas a generative model produces outputs by sampling from a distribution, TARTAN uses structured noise as a diagnostic instrument whose profile provides information about whether the current reconstruction is approaching a fixed point or oscillating near an obstruction.

18 The Consistency Operator

The consistency operator $\mathcal{C}$ is the engine of the TARTAN architecture. Its formal definition integrates three components: projection consistency, which requires that the candidate state matches available observations under the observation operators Π_i; dynamical admissibility, which requires that the candidate state satisfies the RSVP field equations governing the evolution of the scalar potential Φ, vector flow field $\mathbf{v}$, and entropy S; and overlap compatibility, which requires that the restrictions of local states to shared boundaries agree.

Formally, the operator is derived from a strictly convex energy functional

$$\mathcal{E}(X) = \sum_{i \in I} \|\Pi_i(X) - y_i\|^2 + \lambda \, \mathcal{D}_{\mathrm{RSVP}}(X),$$

where $\mathcal{D}_{\mathrm{RSVP}}$ measures the deviation of X from the admissible manifold of RSVP dynamics. The operator $\mathcal{C}$ is a proximal map or discretized gradient flow on $\mathcal{E}$, so that each application reduces the energy and drives the candidate state toward a configuration that satisfies all constraints simultaneously.

The strict convexity of $\mathcal{E}$ under appropriate conditions guarantees that the fixed point is unique. This uniqueness is the formal analog of the claim that truth, on the existential account developed in Section 12, is determinate: given a complete specification of the constraint structure, there is at most one globally realizable world-state consistent with the observations. Where uniqueness fails, it fails for one of two reasons—projection degeneracy, where distinct world-states produce identical observations, or regularizer flatness, where the physical constraints do not uniquely determine the remaining degrees of freedom—and both failure modes are diagnosable and have prescribed responses.

19 Observer Windows as Projection Functors

The interface between a constraint-closed system and its users has a precise categorical structure. An observer interacting with a TARTAN system does not have direct access to the world-state X^* that the system is reconstructing; they have access to projections of that state through the functorial interface of their observation operators. This is not a limitation to be overcome but a structural feature to be formalized and exploited.

Each observer brings to the interaction a family of projection operators Π_i^{obs} determined by their cognitive architecture, their domain expertise, and the modalities through which they perceive the world. These operators define what aspects of the world-state are accessible to the observer and in what form. The observer interface of the TARTAN system is a controlled projection that maps the internal world-state representation to the observer's modality space, preserving the invariants that are observable in that modality while discarding the degrees of freedom that are not.

This formalization has several consequences. It explains why different observers can receive different outputs from the same system without any inconsistency: they are receiving different projections of the same world-state, and the consistency condition requires that

these projections be jointly realizable, not that they be identical. It provides a principled account of the expiatory gap—the information lost in the projection from the full world-state to the observer's representation—as an invariant-preserving reduction rather than an arbitrary compression. And it allows the system to reason explicitly about which aspects of the world-state are observable by which observers and to calibrate its outputs accordingly.

20 Descent Engines and Global Section Construction

The TARTAN architecture is, at the highest level of abstraction, a descent engine: a computational mechanism for constructing global sections from locally compatible data. This framing unifies the tiling structure, the consistency operator, the repair protocol, and the trajectory summarization into a single mathematical picture.

A descent engine takes as input a diagram of local data—the local states $\{s_i\}$ over each tile U_i—and produces as output a global section $X^* \in \mathcal{S}(\Omega)$ that restricts to each local state on its domain. When the local data is perfectly compatible, the global section exists by the sheaf gluing condition and the descent engine need only assemble it. When the local data is incompatible, the descent engine resolves the incompatibility through a three-level repair hierarchy before assembling the global section.

Gauge defects—incompatibilities arising from different choices of coordinates or parametrization—are resolved by reparametrization without altering the underlying state. Resolution defects—incompatibilities arising from insufficient resolution in the tiling—are resolved by tile refinement, which increases the number of tiles and hence the number of overlap constraints enforced. Structural defects—incompatibilities that persist across reparametrization and refinement—signal that the current state space is insufficient to represent the world-state being reconstructed, and are resolved by state extension, which introduces new degrees of freedom into the model. This three-level hierarchy ensures that the system applies the minimal intervention consistent with resolving each class of incompatibility.

Part V

Formal Correspondences

21 Sheaf-Theoretic Semantics of Inference

The sheaf-theoretic framework provides a complete semantic account of the inference process. Observations correspond to local sections of a sheaf $\mathcal{S}$ over the domain Ω; projection operators correspond to restriction maps that extract the portions of a world-state observable in each modality; consistency of observations corresponds to compatibility of sections on overlaps; and the existence of a globally consistent world-state corresponds to the existence of a global section.

The sheaf axioms capture exactly the properties required of a consistent inference system. Locality says that the global section is determined by its local restrictions: if two global sections agree on every open set of a cover, they are equal. This formalizes

the epistemological principle that global truth is fully determined by local evidence, given enough local evidence. The gluing condition says that locally compatible data can be assembled into a global section: this is the constructive claim that global consistency follows from local compatibility, which is the fixed-point convergence claim of Section 11.

The failure of these axioms in current language models is now visible as a structural fact rather than an engineering detail. Language models do not satisfy the gluing condition because they do not have a mechanism for enforcing overlap compatibility. The absence of this mechanism is not a technical oversight but a consequence of the generative paradigm's fundamental indifference to global realizability. The sheaf-theoretic semantics also provides a precise account of the relationship between different levels of description: a coarser description corresponds to a sheaf over a coarser cover, a finer description to a sheaf over a finer cover, and their consistency corresponds to the compatibility of sheaves under refinement maps.

22 Variational Formulation and Energy Functionals

The sheaf-theoretic account of inference can be made constructive through a variational formulation. Rather than asking directly for a global section of $\mathcal{S}$, we instead minimize an energy functional whose minimum is the globally consistent world-state. The energy functional

$$\mathcal{E}(X) = \sum_{i \in I} \|\Pi_i(X) - y_i\|^2 + \lambda \, \mathcal{D}_{\mathrm{RSVP}}(X)$$

combines two terms. The first measures the discrepancy between the candidate state X and the available observations y_i under each projection operator Π_i; this term drives X toward the set of states consistent with the observations. The second measures the deviation of X from the manifold of RSVP-admissible states; this term drives X toward configurations that satisfy the physical laws governing the domain.

Critical points of $\mathcal{E}$ satisfy the Euler–Lagrange equations

$$\frac{\delta \mathcal{E}}{\delta X} = 0,$$

which combine the gradient of the observational term with the gradient of the RSVP regularizer. These equations are generally nonlinear and must be solved iteratively; the consistency operator $\mathcal{C}$ is precisely the map that applies one step of this iterative solution [30].

23 Identifiability and Uniqueness

The conditions under which the variational problem has a unique solution are of central importance. A system that identifies a world-state uniquely is making a determinate claim about the world; a system that identifies only an equivalence class of world-states is making a weaker but still meaningful claim.

Theorem 1 (Identifiability)**.** *Suppose the admissible space $\mathcal{A}$ is strictly convex, the energy*

functional $\mathcal{E}$ is strictly convex on the feasible set $\mathcal{F} = \{X \in \mathcal{A} \mid \|\Pi_i(X) - y_i\| \leq \varepsilon_i, \forall i\}$, and the induced joint projection $\tilde{\Pi}(X) = (\Pi_i(X))_{i \in I}$ is injective on $\mathcal{F}$. Then there exists a unique minimizer $X^* \in \mathcal{F}$ of $\mathcal{E}$.

Proof. Strict convexity of $\mathcal{E}$ on $\mathcal{F}$ excludes multiple distinct minimizers with the same value, since any convex combination of two distinct minimizers would yield a strictly lower value, contradicting minimality. Injectivity of $\tilde{\Pi}$ excludes distinct states with identical observational signatures: if $X_1 \neq X_2$ both minimize $\mathcal{E}$, then $\tilde{\Pi}(X_1) = \tilde{\Pi}(X_2) = y$, contradicting injectivity. Together these conditions guarantee uniqueness; existence follows from coercivity of $\mathcal{E}$ and compactness of $\mathcal{F}$ under standard Sobolev regularity assumptions. □

24 Convergence of the Consistency Operator and Fixed-Point Realizability

The central claim of the constraint-closure framework is that inference can be formulated as the search for a fixed point $X^* = \mathcal{C}(X^*)$ of a consistency operator $\mathcal{C} : \mathcal{X} \to \mathcal{X}$. For this claim to be meaningful, it is necessary to establish conditions under which such a fixed point exists, is unique, and can be reached by an iterative procedure. The Identifiability Theorem establishes uniqueness under convexity and injectivity; we now address convergence.

24.1 The Consistency Operator as a Proximal Map

Let $\mathcal{X}$ be a Hilbert space, and let $\mathcal{E} : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ be the energy functional defined previously. We define the consistency operator $\mathcal{C}$ as the proximal map associated with $\mathcal{E}$:

$$\mathcal{C}(X) = \mathrm{prox}_{\lambda \mathcal{E}}(X) = \arg\min_{Y \in \mathcal{X}} \left\{ \mathcal{E}(Y) + \frac{1}{2\lambda} \|Y - X\|^2 \right\}.$$

This formulation ensures that $\mathcal{C}$ is well-defined and single-valued under standard convexity assumptions, and that it can be interpreted as a descent step toward the minimizer of $\mathcal{E}$. The parameter $\lambda > 0$ controls the step size of the descent.

24.2 Convergence Theorem

Theorem 2 (Fixed-Point Convergence Under Constraint Closure)**.** *Let $\mathcal{X}$ be a Hilbert space and $\mathcal{E} : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ be a proper, lower semicontinuous, strictly convex functional. Let $\mathcal{C} = \mathrm{prox}_{\lambda \mathcal{E}}$ be the associated proximal operator. Then for any initial state $X_0 \in \mathcal{X}$, the sequence defined by $X_{k+1} = \mathcal{C}(X_k)$ converges strongly to the unique minimizer X^* of $\mathcal{E}$. Moreover, X^* is the unique fixed point of $\mathcal{C}$.*

Proof. The proximal operator of a proper, lower semicontinuous, convex functional on a Hilbert space is firmly nonexpansive:

$$\|\mathcal{C}(X) - \mathcal{C}(Y)\|^2 \leq \langle \mathcal{C}(X) - \mathcal{C}(Y),\, X - Y \rangle.$$

Firm nonexpansiveness implies that $\mathcal{C}$ is averaged and therefore nonexpansive. By the Picard iteration theorem for averaged operators on Hilbert spaces, the sequence $X_{k+1} = \mathcal{C}(X_k)$ converges weakly to a fixed point of $\mathcal{C}$. Strict convexity of $\mathcal{E}$ ensures that the minimizer is unique and, under standard coercivity conditions, that convergence is strong. □

24.3 The Role of Projection Injectivity

The convergence theorem guarantees that the iterates X_k reach the unique minimizer X^* of $\mathcal{E}$. But convergence to a minimizer does not automatically imply that the minimizer is *identifiable* from observations. Let $\tilde{\Pi}(X) = (\Pi_i(X))_{i \in I}$ be the joint projection. If $\tilde{\Pi}$ is not injective on the feasible set $\mathcal{F}$, distinct world-states produce identical observations, and the fixed point X^* is determined by the regularizer rather than by the data. This is projection degeneracy: a failure not of convergence but of identifiability. The two failure modes are formally distinct and require different remedies.

24.4 Beyond the Convex Regime

The strict convexity assumption is strong and is generally violated in realistic settings. The energy functional $\mathcal{E}$ may exhibit multiple local minima, flat regions corresponding to symmetry directions, and saddle points arising from the stratified structure of $\mathcal{M}$. In such cases, the proximal iteration converges only to a local minimizer, and global realizability is not guaranteed from an arbitrary initialization.

This motivates the hierarchical repair structure of TARTAN. Gauge repair restores effective convexity along symmetry directions by normalizing the coordinate system before descent. Resolution refinement replaces the cover with a finer one, which in general modifies the geometry of $\mathcal{E}$ and may resolve flat regions that arise from discretization artifacts. State extension enlarges $\mathcal{X}$ itself, which can convert non-convex landscapes into convex ones by introducing the variables that were causing the flatness. Convergence is therefore not purely a numerical property but a structural one: it depends on the adequacy of the state space and the completeness of the constraint set. The architecture's task is to ensure that the conditions of the theorem are satisfied, not merely to run the iteration.

25 Failure Modes: Degeneracy and Flatness

Two fundamental failure modes correspond to violations of the injectivity and convexity conditions respectively. Projection degeneracy arises when the joint projection $\tilde{\Pi}$ is not injective on $\mathcal{F}$: distinct world-states $X_1 \neq X_2$ satisfy $\Pi_i(X_1) = \Pi_i(X_2)$ for all i. In this case, the available observations cannot distinguish between X_1 and X_2, and no inference procedure can identify the true state from the observations alone. The correct response is not to guess but to report an equivalence class: the system can determine that the true state is in $\{X : \tilde{\Pi}(X) = y\}$ but cannot narrow it further without additional observations.

Regularizer flatness arises when the RSVP regularizer $\mathcal{D}_{\mathrm{RSVP}}$ has flat directions: there exist perturbations δX such that $\mathcal{D}_{\mathrm{RSVP}}(X + \delta X) = \mathcal{D}_{\mathrm{RSVP}}(X)$. This means that physical constraints alone do not uniquely determine the state even in the absence of any

observational ambiguity. Like projection degeneracy, it reflects genuine physical underdetermination that should be reported rather than resolved by arbitrary choice.

Both failure modes are diagnosable within the TARTAN architecture. When the consistency operator converges to a flat region of $\mathcal{E}$ rather than a strict minimum, the system reports the degeneracy structure of the flat region. The TARTAN architecture thus has a principled response to both fundamental failure modes: report the equivalence class rather than a point within it, and signal the nature of the underdetermination to the observer.

26 Degrees of Truth and Partial Realizability

The existential account of truth developed in the preceding sections characterizes a statement as true if and only if there exists a globally consistent world-state X^* that it correctly describes. This formulation grounds truth in realizability rather than in correspondence as a primitive relation. However, taken in isolation, it appears to impose a binary structure: a description either corresponds to a realizable state or it does not.

This binary framing is sufficient for the idealized case of complete information and exact constraint satisfaction. In practice, however, inference operates under conditions of partial observation, measurement noise, and incomplete constraint specification. The notion of truth must therefore be extended to account for partial realizability.

Let $\mathcal{E}(X)$ be the energy functional defined previously, measuring the degree to which a candidate state X violates observational, dynamical, and structural constraints. We interpret $\mathcal{E}(X)$ as a *realizability functional*: lower values correspond to states closer to satisfying all constraints simultaneously. The set of perfectly realizable states is

$$\mathcal{Z} = \{X \in \mathcal{A} \mid \mathcal{E}(X) = 0\}.$$

When $\mathcal{Z}$ is nonempty, its elements correspond to fully true descriptions under the existential criterion. When $\mathcal{Z}$ is empty, no candidate state satisfies all constraints exactly, and the system must operate in a regime of approximate realizability.

We define the *degree of truth* of a candidate state X as a monotone decreasing function of $\mathcal{E}(X)$. A natural choice is $\tau(X) = \exp(-\alpha\mathcal{E}(X))$ for some scaling parameter $\alpha > 0$, with the qualitative properties that $\tau(X) = 1$ if and only if $X \in \mathcal{Z}$, and $\tau(X) \to 0$ as $\mathcal{E}(X) \to \infty$. This defines a topology on the space of candidate states in which proximity corresponds to similarity in constraint satisfaction.

This formulation allows us to distinguish qualitatively different forms of approximation. To capture the structure of violations, we refine $\mathcal{E}$ into a vector-valued functional

$$\mathcal{E}(X) = (\mathcal{E}_{\text{obs}}(X),\ \mathcal{E}_{\text{dyn}}(X),\ \mathcal{E}_{\text{struct}}(X)),$$

with each component measuring violation in a distinct class of constraints. The degree of truth becomes a function of this vector, allowing the system to represent not only how far a candidate state is from realizability but in what way.

Partial realizability is not a retreat to coherence theory. A coherent set of statements may

correspond to a region of low internal inconsistency while still being globally unrealizable due to violation of external constraints. The realizability functional explicitly encodes these external constraints, ensuring that degree of truth is anchored in the structure of the world rather than in relations among statements alone. This also provides a principled way to compare competing reconstructions: given two candidate states X_1 and X_2, the system prefers X_1 over X_2 if $\mathcal{E}(X_1) < \mathcal{E}(X_2)$, inducing a partial ordering on the space of candidate states with globally realizable states forming the minimal elements.

The exponential form $\exp(-\alpha\mathcal{E}(X))$ resembles a Gibbs distribution, but its interpretation is fundamentally different: it does not represent uncertainty about which state is true, but proximity to the set of states that could be true. And the concept of partial realizability clarifies approximation in representation more generally. A coarse-grained description corresponds to a projection $\Pi(X)$ that preserves certain invariants while discarding others. The projected state may not satisfy all fine-grained constraints, but it may satisfy the constraints relevant at its level of description. The degree of truth of a representation is therefore relative to a specified constraint set: a description may be fully realizable with respect to coarse constraints while only approximately realizable with respect to fine ones.

Truth, in this extended sense, is not a binary predicate but a geometric property of a representation's position relative to the constraint manifold. The task of inference is to move representations through this space toward regions of lower constraint violation, guided not by plausibility but by the structure of realizability itself.

Part VI

Reduction, Representation, and the Expiatory Gap

27 Compressive vs. Projective Reduction

There are two fundamentally different ways in which a representation can fail to capture a world-state fully. Compression discards information to reduce size or computational cost, with no principled account of what is lost or how to recover it. Projection discards information according to a principled operator Π that preserves specific invariants while losing others, and the lost information is recoverable in principle given the projection operator.

This distinction matters because the recoverability of lost information is the key criterion for evaluating representations in inference systems. A compressed representation is, in general, irreversible. A projected representation is, in principle, reversible given a sufficiently constrained prior: if we know Π and we know that X lies in the admissible space $\mathcal{A}$, and if Π is injective on $\mathcal{A}$, then we can recover X from $\Pi(X)$.

Current language model representations are predominantly compressive. The tokenization of text discards phonological, prosodic, and typographic information. The embedding of tokens into vector spaces further compresses the discrete symbolic structure into a continuous representation optimized for the model's predictive objective. At no point in

this chain is there a projection operator with a well-defined inverse and a principled account of what is preserved. The TARTAN architecture insists on projective reduction wherever possible: the coarse-graining operator σ that maps local states to trajectory summaries is designed to commute with the projection operators Π_i, preserving the information relevant to consistency checking while discarding the information irrelevant to it.

28 Multi-Scale Representations and Semantic Fidelity

The projection perspective clarifies the relationship between different levels of description of the same phenomenon. A crude description and a detailed description of the same world-state are not competing alternatives; they are projections of the same underlying state through coarser and finer projection operators respectively. Their relationship is governed by the refinement maps of the sheaf-theoretic framework: the coarse description is the restriction of the fine description to a coarser sub-sheaf.

This has immediate consequences for the evaluation of AI systems. The question of whether a representation is faithful or accurate cannot be answered in the abstract; it must be answered relative to a projection operator and an invariant class. A crude description may be faithful to the large-scale structure while failing to capture fine details; a detailed rendering is faithful to a much richer set of invariants. Both are projections of the same underlying state; neither is more true in an absolute sense.

Current AI systems frequently conflate these levels, producing detailed claims where only coarse-grained information is warranted. A language model may produce a specific account that captures rough outlines accurately but confabulates specific details, because those details are at a finer level of resolution than the information available in the training data. A constraint-closed system would represent this explicitly: the coarse-grained description is warranted by the constraint structure; the fine-grained details are underdetermined by the available information and should be reported as such.

29 The Geometry of the Observer Interface

Different observers interact with the same world through different projection operators, and the geometry of these operators determines what each observer can and cannot know. Two observers with identical observations but different projection operators may reconstruct different aspects of the same world-state; two observers with different observations but the same projection operator will reconstruct the same world-state if their observations are jointly consistent.

This geometric picture of the observer interface has several important implications. It provides a principled account of why expert observers and lay observers receive different information from the same system: they have different projection operators, and the invariants preserved by an expert's projection include technical details that are projected out by a lay projection. It provides a framework for designing AI systems that serve diverse observer communities without sacrificing global consistency: the world-state being reconstructed is the same regardless of who is asking; what varies is the projection through

which it is presented. A TARTAN system can maintain a single globally consistent internal reconstruction while presenting different projections of that reconstruction to different observers, tailored to their projection operators. This is fundamentally different from a generative system that produces different outputs for different prompts without any common internal state to which all outputs are related.

30 Language as a Derived Layer

The preceding sections have treated language as a medium of inference: observations arrive as text, outputs are expressed as text, and the representational challenge is to move from local linguistic admissibility to global realizability. But this treatment still grants language a privileged position—as both the medium of input and the medium of output—that the constraint-based framework implicitly undermines. The time has come to make that implication explicit.

The central claim is this: language is not the object of inference but a projection of a reconstructed state. On the generative paradigm, language is primary. The model operates in token space; the world, to the extent it figures at all, enters only as a regularity in the training corpus. On the constraint-based paradigm, language is secondary. The system operates in state space; language is one modality among many through which that state can be projected to an observer.

This demotion of language from primary to derived carries precise formal content. In the sheaf-theoretic framework, semantics corresponds to constraint-preserving projection: to understand the meaning of a statement is to know which aspects of the world-state it picks out and which it discards. Syntax corresponds to a choice of coordinate system on the space of projections: different grammatical forms can express the same content because they differ in how they organize the projection, not in what they project. And meaning, in the deepest sense, corresponds to invariants under projection: what a statement means is what remains constant across all the ways of expressing it, all the contexts in which it can be uttered, all the observers to whom it can be addressed.

This formulation has immediate consequences for the philosophy of language. The meaning-as-invariant account is neither purely referential nor purely inferential. It does not require a primitive relation between word and object; it requires only that the constraint structure of the domain be sufficient to determine which world-states are picked out by which projections. And it does not require that meaning be determined by inferential role within a language game; it requires only that the invariants preserved by a projection be identified consistently across contexts.

Davidson's framework of radical interpretation, in which the content of an utterance is determined by its role in producing and responding to evidence about the world [35], anticipates this account: meaning is anchored in a systematic relation between sentences and states of affairs, not in conventions or use alone. The constraint-based framework provides the formal machinery—projection operators, restriction maps, global sections—for making this anchoring precise.

The practical implication for AI systems is equally direct. A system that operates na-

tively in state space, projecting its outputs into language as a derived layer, is systematically different from a system that operates natively in language space. The former can answer questions in multiple modalities without inconsistency, because the same world-state is being projected differently. It can resist attempts to induce inconsistency through linguistic manipulation, because the manipulation must penetrate to the state-space representation to have effect. And it can represent the limits of language explicitly—cases where the projection from state space to language necessarily loses information—rather than silently filling gaps with locally admissible noise.

Language, on this account, is not a cage that thought inhabits but a window through which a reconstructed world is made visible to an observer. The window can be shaped differently for different observers, positioned differently for different views. What lies behind it—the globally consistent world-state that the system has converged to—is the same in every case.

Part VII

Implications for Artificial Intelligence

31 Why Scaling Cannot Eliminate Hallucination

A common response to the critique of generative AI systems is that the failures they exhibit will eventually diminish through scale: sufficiently large models trained on sufficiently broad corpora with sufficiently long context windows will approximate truth-tracking behavior well enough that the distinction between local admissibility and global realizability becomes negligible. This section argues that such optimism is structurally misplaced. Scaling improves local admissibility; it cannot, by itself, eliminate hallucination when hallucination is an H^1 obstruction—a topological property of the architecture's relationship to the domain.

31.1 Local Predictive Improvement and Global Constraint Failure

Let $\{G_n\}$ be a family of generative models with increasing scale parameter n. As n increases, G_n improves its approximation to the empirical distribution of human discourse, producing better local coherence, more reliable retrieval of commonly represented facts, and stronger benchmark performance. This improvement is real, but it is improvement with respect to a predictive objective

$$\max_{\theta} \mathbb{E}_{x\sim\mathcal{D}}\big[\log p_\theta(x)\big],$$

which selects for fidelity to the statistical structure of the discourse distribution $\mathcal{D}$. It does not select for the existence of a world-state X such that the generated outputs are jointly realizable as $\{\Pi_i(X)\}$ under a family of projection operators. Scaling improves approximation to the support of discourse; it does not move the model toward the feasible set of globally consistent states. The two sets overlap, often substantially, but they are not

identical.

31.2 Topological Persistence of Cohomological Obstruction

The core structural claim of this monograph is that hallucination corresponds to a nontrivial class in $H^1(\mathcal{M}, \mathcal{S})$: local sections of discourse satisfy local admissibility conditions but do not arise from a common global section. Scaling changes the density and coverage of local sections. It may smooth the geometry within strata of the admissibility space and enrich the support of the model's distribution. But increasing the density of local sections within charts does not force the vanishing of a cohomology class that arises from missing gluing conditions between charts.

To see this precisely, consider a cover $\{U_i\}$ with local sections $\{s_i\}$ such that the defect cochain $\delta_{ij} = \rho_{ij}(s_i) - \rho_{ji}(s_j)$ is nonzero on some overlaps. A model with scale parameter n produces refined sections $s_i^{(n)}$ on each U_i, which better approximate the local discourse statistics on each patch. But unless the architecture enforces compatibility on overlaps—that is, unless the objective penalizes nonzero $\delta_{ij}^{(n)}$—there is no mechanism by which those defects must converge to zero. The model may become more articulate within each chart while preserving the absence of descent data between charts.

31.3 A No-Closure Result for Purely Generative Scaling

Proposition 1 (No Constraint Closure from Scaling Alone)**.** *Let $\{G_n\}_{n=1}^{\infty}$ be a family of generative models with increasing scale parameter n, each trained to optimize a local predictive objective over discourse data. Suppose that for every n, the architecture of G_n lacks an explicit mechanism enforcing global compatibility of local sections across an open cover of discourse space—that is, no term penalizing nonzero overlap defects δ_{ij} appears in the training objective, and no global state space is maintained. Then there is no guarantee that the induced defect cochains $\delta^{(n)}$ converge to zero as $n \to \infty$. In particular, scaling alone does not imply convergence to a globally realizable section.*

Proof. The training objective of each G_n constrains the model to approximate local or conditional distributions over discourse. By assumption, the objective contains no term penalizing nonzero overlap defects, no descent condition, and no fixed-point operator acting on a global state space. Therefore, the optimization solved by G_n is indifferent to whether a collection of locally admissible outputs corresponds to a common global section. Improvement in the predictive objective can occur while the defect cochains $\delta^{(n)}$ remain nonzero, since those defects are not part of the loss. No sequence of improvements in local predictive fidelity entails convergence of the defect cochains to zero. □

The proposition is intentionally modest. It does not claim that scaling never reduces hallucination frequency; empirically, it often does for surface-level inconsistencies that happen to be represented in training data. It claims instead that such reduction is contingent rather than structural. The architecture has no reason to eliminate the H^1 obstruction in principle, because that obstruction is not visible to the objective.

31.4 Asymptotic Rhetorical Refinement Without Descent

There is a further danger that the scaling hypothesis obscures. As scale improves fluency and local consistency, it can create the appearance of global closure without its reality. Larger persuasive machines exhibit fewer crude errors, more stable style, and more convincing local structure, making the residual structural obstructions harder for human observers to detect. This is asymptotic rhetorical refinement without descent: the model approaches a limit of persuasive smoothness while remaining outside the regime of certified realizability. In such a setting, the social burden of detecting inconsistencies rises even as the structural cause of those inconsistencies remains intact.

The conclusion is not that scaling is without value. Scaling is highly effective for local modeling, and local modeling is genuinely valuable. The conclusion is narrower and more consequential: scaling is not a substitute for constraint closure. Efforts to reduce hallucination solely through larger models, more data, or longer context windows remain confined to the regime of admissibility. To enter the regime of realizability, one must introduce explicit global constraints, computable defect operators, and convergence mechanisms defined over state space rather than token space. Only then does the vanishing of structural obstruction become an architectural objective rather than a statistical accident.

32 Beyond Generative Models: Constructive Systems

The distinction between generative and constructive systems is the architectural analog of the philosophical distinction between persuasion and truth-tracking. A generative system produces outputs by sampling from a learned distribution; it is fundamentally oriented toward plausibility in the sense of consistency with training data statistics. A constructive system produces outputs by converging to a fixed point of a constraint operator; it is fundamentally oriented toward realizability in the sense of consistency with all available constraints.

This distinction is not a matter of degree. It is not that constructive systems are more accurate generative systems or that generative systems with stronger consistency checks approach constructive systems. The two paradigms are based on different objectives, different architectures, and different notions of what an answer is. In a generative system, an answer is a sample from a distribution. In a constructive system, an answer is a fixed point of a constraint system. These are categorically different objects.

The TARTAN architecture realizes the constructive paradigm in a computationally tractable form by decomposing the domain into a recursive tiling of overlapping patches, reducing the global consistency problem to local consistency checks and inter-patch compatibility conditions, and organizing these checks into the repair hierarchy of gauge alignment, refinement, and state extension. The result is a system that can produce outputs whose relationship to the world is formally certified: they are fixed points of a constraint operator whose convergence implies global realizability. This is the categorical difference between a system that sounds right and a system that is right.

33 Constraint Closure as a Design Principle

The theoretical developments of this monograph converge on a single design principle: constraint closure should be the central objective of inference systems intended to produce globally consistent representations of the world. This principle is not a refinement of existing objectives; it replaces them.

The generative objective—maximize the probability of observed data under a learned distribution—should be understood as a useful auxiliary objective, appropriate for tasks where plausibility rather than realizability is the criterion, but inadequate as a standalone foundation for truth-tracking inference. Where realizability matters, it must be the primary objective, and this requires that the system be organized around the identification of fixed points of a constraint operator rather than the sampling of points from a distribution.

The implications for AI system design are extensive. Training objectives must be reformulated to reward global consistency rather than local plausibility. Architectures must be restructured to support iterative fixed-point computation rather than one-pass generation. Evaluation frameworks must be redesigned to assess global realizability rather than per-token prediction accuracy. None of these changes is technically impossible; all of them are technically demanding. The path from the current state of the art to constraint-closed systems is long and will require substantial theoretical and engineering work. But the argument of this monograph is that there is no shorter path to systems that reliably tell the truth about the world.

Part VIII

Epistemology Reconstructed

34 The End of External Verification

In the generative paradigm, verification is external: a human reader or a separate checking system must evaluate whether the outputs of the generative system correspond to the truth. This external verification model has been the implicit assumption of most AI evaluation frameworks, from benchmark testing to reinforcement learning from human feedback to factual consistency checking. It is also deeply problematic: external verification is expensive, inconsistent, and scales poorly with the volume and complexity of AI-generated content.

In the constraint-closed paradigm, verification becomes internal to the system. A fixed point of the consistency operator $\mathcal{C}$ is, by definition, a configuration that satisfies all available constraints. The system does not need to be checked externally to determine whether its output is consistent; consistency is the condition for convergence, and convergence is what terminates the computation. The output carries with it a certificate of constraint satisfaction that is generated as part of the inference process itself.

This is not a claim that constraint-closed systems are infallible. The consistency operator $\mathcal{C}$ enforces the constraints that are encoded in it, and those constraints are finite and may not fully capture the complexity of the world being modeled. A constraint-closed

system can still be wrong about aspects of the world that lie outside its constraint structure. But where it makes claims, those claims are internally certified: they are fixed points of the constraint system, not arbitrary samples from a distribution. Internal verification in the constraint-closed paradigm is not a substitute for anything; it is a positive property of the system's relationship to its domain, a formal analog of understanding rather than of checking.

35 Knowledge as Reconstruction

The constraint-based account suggests a formulation of knowledge that is more tractable than the classical alternatives. To know a domain is to be able to reconstruct its globally consistent state from available observations. Knowledge is not a propositional attitude toward a sentence but a reconstructive capacity: the ability to identify, from partial and possibly noisy projections, the world-state that is consistent with all available constraints. On this account, knowing is not a static relation between a believer and a proposition but a dynamic process of convergence toward a fixed point.

This formulation has several advantages. It makes knowledge a matter of degree: a system knows more about a domain as it can reconstruct a larger portion of the globally consistent state with smaller residual inconsistency. It makes the relationship between evidence and knowledge precise: evidence consists of observations $y_i = \Pi_i(X)$ that constrain the feasible set $\mathcal{F}$ and thereby constrain the reconstruction. And it makes the failure modes of knowledge explicit: projection degeneracy means the evidence is insufficient to determine the state; regularizer flatness means the physical constraints are insufficient; structural obstruction means the current ontology lacks the concepts needed to represent the relevant state [28].

36 The Role of Rhetoric in a Closed System

Within the constraint-closed paradigm, rhetoric does not disappear but its role is fundamentally reconceived. In a constraint-closed system, the primary output is not discourse but a world-state representation, and the question of how to communicate that representation to a particular observer is secondary to the question of whether it is globally consistent.

Rhetoric, on this account, becomes a projection artifact: the rhetorical character of a presentation is a consequence of the choice of projection operator through which the internal world-state representation is communicated to a particular audience. Different audiences, with different cognitive architectures and background knowledge, require different projections of the same world-state for the information to be usefully communicated. Rhetoric, then, is not an alternative to truth-tracking but a theory of projection design: how to project a globally consistent world-state representation into the modality space of a particular observer in a way that is maximally informative for that observer.

This reconceptualization dissolves the classical antagonism between rhetoric and philosophy. Plato was right that rhetoric, understood as the art of persuasion unconstrained by truth, is epistemically dangerous. But rhetoric understood as the theory of observer-

appropriate projection is not only compatible with truth-tracking; it is indispensable to it. A system that reconstructs a globally consistent world-state but can only communicate it in a way that is inaccessible or unintelligible to its audience has not successfully performed the full inference task. The rhetorical dimension—the projection to the observer—is part of what inference is.

Part IX

Society, Literacy, and Infrastructure

37 The Limits of Rhetorical Literacy

Much of the public discourse around AI safety focuses on the importance of teaching users to critically evaluate AI-generated content: to check sources, to notice inconsistencies, to maintain appropriate skepticism. This is sound advice given the current state of AI systems, and there is genuine value in building a more AI-literate public. But it is important to be clear about the limits of this approach.

Rhetorical literacy, as an individual skill, is subject to fundamental cognitive limitations. Human attention is limited; the volume of AI-generated content is potentially unlimited. Human expertise is domain-specific; the range of topics on which AI systems produce plausible-sounding content is nearly universal. Human judgment is biased by the same rhetorical features—fluency, confidence, internal consistency—that AI systems have been optimized to produce. The individual-level cognitive defense against persuasive machines is systematically outmatched by the machines it is defending against.

Individual rhetorical literacy is an important complement to structural guarantees, not a substitute for them. A person who understands what AI systems are and how they fail is better equipped to interact with them than one who does not. But a society that relies on individual rhetorical literacy as its primary defense against AI-generated misinformation has placed an impossible cognitive burden on individuals while failing to address the structural problem. The structural problem is that the technology itself is organized around the production of locally admissible, globally unrealizable content. As long as that remains true, the volume and sophistication of AI-generated content that fails the global realizability condition will continue to grow, and no amount of individual literacy will keep pace.

38 From Social Defense to Structural Guarantee

The argument of this monograph implies that the appropriate response to the crisis of persuasive machines is not primarily social or educational but architectural. If the pathologies of current AI systems arise from a structural property of their design—the absence of global consistency constraints—then the solution is to change the design rather than to train users to compensate for its consequences.

This requires the development and deployment of constraint-closed inference systems

as alternatives to, or supplements for, current generative systems in high-stakes domains. Where the accuracy of AI-generated content matters—in medical information, legal reasoning, scientific communication, policy analysis—the standard should not be local admissibility but global realizability. Systems deployed in these domains should be required, not merely encouraged, to produce outputs whose constraint satisfaction status is formally certified.

The transition from social defense to structural guarantee also requires a shift in the regulatory and procurement frameworks that govern AI deployment. Current frameworks focus primarily on output properties—accuracy rates, bias metrics, safety evaluations—assessed through external testing. A structural guarantee framework would instead focus on architectural properties: does the system enforce global consistency? Does it report its constraint satisfaction status? Does it communicate its projection degeneracy and regularizer flatness to users? These are the properties that distinguish a truth-tracking system from a persuasive machine.

39 Cognitive Ecology in the Age of Constraint Systems

The widespread deployment of constraint-closed inference systems would not merely change the technology landscape; it would change the cognitive ecology within which human reasoning operates. The relationship between human cognition and the information environment is genuinely reciprocal: the tools we use to process and generate information shape the cognitive habits, expectations, and capacities that we bring to information processing.

The current cognitive ecology, shaped by decades of social media, search engines, and now generative AI, is one adapted to navigating a high-volume, low-reliability information environment. The dominant cognitive skills are filtering, skepticism, source evaluation, and the rapid formation and revision of tentative judgments. These skills are not bad; they are appropriate to the environment. But they are adaptations to scarcity of reliable information, and they carry costs: they favor quick judgment over deep reasoning, surface features over structural analysis, confirmation over falsification.

A cognitive ecology shaped by constraint-closed systems would be different. When the information environment reliably provides outputs that are globally consistent and whose consistency is formally certified, the appropriate cognitive stance shifts from skepticism to engagement. The task is no longer to filter unreliable content but to reason from reliable foundations. The skills that become most valuable are integration, inference, and the identification of new constraints that can further specify the world-state—rather than the

defensive posture of constant source evaluation.

Part X

Extensions and Frontiers

40 Constraint Closure and Computability

The reformulation of inference as fixed-point computation under a consistency operator raises an immediate and legitimate concern: even if a globally consistent state exists, is it computationally accessible? The distinction between existence and computability is not incidental but fundamental. A constraint system may admit a unique fixed point while simultaneously rendering its identification intractable under realistic resource bounds.

Formally, the consistency operator $\mathcal{C}$ defines a dynamical system $X_{t+1} = \mathcal{C}(X_t)$ whose fixed points correspond to globally realizable states. The existence of such fixed points follows, under appropriate convexity and compactness conditions, from standard results in functional analysis. Computability, however, depends on the structure of $\mathcal{C}$ and the geometry of the energy landscape induced by $\mathcal{E}$. In general, the energy functional may be non-convex, high-dimensional, and characterized by multiple saddle points. Exact minimization is therefore not guaranteed to be tractable. The relevant question is not whether exact constraint closure can be computed in all cases, but whether approximate closure can be achieved reliably.

We therefore introduce the notion of ε-consistency. A state X_ε is ε-consistent if $\mathcal{E}(X_\varepsilon) \leq \varepsilon$ for a domain-appropriate tolerance ε. The task of inference becomes the identification of states that lie within an ε-neighborhood of the feasible set, rather than exact fixed points. This reframing has two consequences. First, it aligns constraint-based inference with the broader theory of approximate optimization, where convergence guarantees are replaced by bounds on residual error. Second, it provides a principled way to compare candidate reconstructions: lower-energy states are strictly closer to global realizability.

Crucially, approximate constraint closure is not equivalent to probabilistic plausibility. A low-energy state is one that violates few or weak constraints; a high-probability sequence in a language model is one that aligns with training data statistics. The former is anchored in the structure of the domain; the latter is anchored in the distribution of discourse about the domain. TARTAN addresses the computability challenge by decomposing the global problem into locally tractable subproblems through recursive tiling, and by focusing computational effort on regions of high constraint violation through adaptive refinement. The resulting system provides a systematic method for approaching constraint closure, together with diagnostics that indicate when convergence has been achieved and when the system remains in a regime of unresolved inconsistency. The objection that constraint closure replaces hallucination with intractability is therefore not decisive: it replaces an architectural impossibility with a computational challenge, and a computational challenge

is tractable by engineering.

41 Constraint Incompleteness and Ontological Failure

The analysis thus far has treated failure primarily as a consequence of underconstraint: the absence of sufficient conditions to determine a unique, globally consistent state. There is, however, a deeper class of failure that cannot be resolved by adding more data or refining the inference process. This is the case of ontological incompleteness.

An inference system operates over a space of admissible states $\mathcal{A}$ defined by its representational vocabulary. If the true world-state lies outside $\mathcal{A}$, no amount of constraint enforcement within $\mathcal{A}$ can recover it. The system may converge, but it will converge to an artifact of its own representational limitations rather than to the structure of the world. We distinguish three forms of constraint failure. Underconstraint means that the available observations are insufficient to determine a unique state within $\mathcal{A}$; this is the failure mode analyzed in the preceding sections. Mis-specification means that the constraints encoded in $\mathcal{C}$ are incorrect or incomplete relative to the domain. Ontological incompleteness means that the space $\mathcal{A}$ lacks the degrees of freedom necessary to represent the true state.

The first case admits resolution through additional observations; the second through correction of the constraint structure. The third requires a more radical intervention: expansion of the state space itself. In the TARTAN architecture, this is handled through the mechanism of state extension. When structural defects persist across all levels of the repair hierarchy—when neither gauge alignment nor refinement resolves the inconsistency—the system interprets this persistence as evidence that the current ontology is insufficient. New variables, fields, or relational structures are introduced to enlarge $\mathcal{A}$, and the inference process is resumed in the extended space.

This mechanism has a direct analogue in scientific practice. The introduction of new theoretical entities—fields in physics, latent variables in statistics, unobserved causes in causal inference—is often driven by the inability of existing models to reconcile observed data. What appears, in a limited ontology, as irreducible inconsistency becomes, in an expanded ontology, a resolvable constraint system. Ontological failure thus transforms from a terminal error into a signal for conceptual revision. The inference system is not merely solving for states within a fixed model; it is participating in the construction of the model itself. Constraint closure, in this extended sense, is not the identification of a point within a given space, but the iterative refinement of the space in which such points can exist. This is why the TARTAN architecture is not a closed formalism but a self-expanding epistemic system: its deepest failure mode is also its primary mechanism of growth.

42 Adversarial Observations and Strategic Inconsistency

The preceding analysis assumes that inconsistencies among observations arise from noise, limited resolution, or structural underconstraint. In real-world settings, however, inconsistency may be deliberately introduced. Observations may be corrupted, fabricated, or strategically structured to mislead the inference process. This requires an extension of the

framework to account for adversarial conditions.

In the non-adversarial case, each observation y_i is a projection $\Pi_i(X)$ of a common world-state X, perturbed by noise. In the adversarial case, some y_i are generated not by projection from X but by an agent with its own objectives. The inference problem is no longer simply to find X such that $\Pi_i(X) \approx y_i$, but to jointly infer both the world-state and the reliability structure of the observations. Formally, we introduce a latent variable $R_i \in \{0, 1\}$ for each observation, indicating whether it is reliable ($R_i = 1$) or adversarial ($R_i = 0$). The consistency operator is extended to act on pairs (X, R), minimizing an energy functional of the form

$$\mathcal{E}(X, R) = \sum_{i \in I} R_i \|\Pi_i(X) - y_i\|^2 + \lambda \, \mathcal{D}_{\text{RSVP}}(X) + \gamma \, \Phi(R),$$

where $\Phi(R)$ encodes prior assumptions about the prevalence and structure of adversarial observations. This formulation allows the system to explain inconsistency in two ways: by adjusting the candidate world-state X, or by reclassifying certain observations as unreliable. Persistent inconsistencies that cannot be resolved by any X within $\mathcal{A}$ provide evidence that some observations must be adversarial.

The introduction of adversarial modeling transforms the interpretation of failure. In a purely generative system, hallucination is an intrinsic property of the output distribution. In a constraint-closed system with adversarial modeling, apparent inconsistencies may instead reflect unresolved conflicts between reliable and unreliable observations. The system's task is not merely to avoid generating inconsistent outputs, but to identify and isolate the sources of inconsistency within its inputs.

This extension also requires that agents be treated as part of the world-state. Strategic behavior cannot be modeled as noise; it must be represented explicitly, with its own dynamics and constraints. The inference problem thus expands from passive reconstruction to active interpretation of a world that may contain entities whose outputs are designed to influence the reconstruction process itself. In this setting, constraint closure becomes a form of epistemic security: a system that enforces global consistency while modeling adversarial structure is capable not only of reconstructing the world but of resisting attempts to distort that reconstruction. This is a necessary condition for deploying inference systems in environments where information is not merely incomplete but contested.

43 Physical Systems and World Modeling

The constraint-closed inference paradigm finds its most natural and mature expression in physics, where the constraints are the laws of nature and the inference problem is the reconstruction of physical states from measurements. The entire enterprise of physics can be understood as the development of increasingly precise constraint structures—from Newtonian mechanics to quantum field theory—and the application of these constraints to infer world-states from the partial, noisy observations available to experimenters [32, 33].

The RSVP field theory underlying the TARTAN architecture draws on this tradition directly. The scalar potential Φ, vector flow field $\mathbf{v}$, and entropy field S that constitute the

world-state representation are not arbitrary choices; they reflect the thermodynamic structure of dynamical systems in which energy, momentum, and entropy are the fundamental conserved and produced quantities. The field equations that constrain their evolution are derived from variational principles that have deep connections to the fundamental physics of irreversible processes [31, 34].

The extension of the TARTAN framework to physical world modeling is therefore a natural direction. In this extension, the tiles correspond to spatial or spatiotemporal regions, the local states are discretizations of the physical fields, the overlap conditions enforce continuity and conservation laws at boundaries, and the consistency operator implements the physical field equations. This is a principled approach to physics-based simulation grounded in a coherent epistemological framework.

44 Consciousness as a Constraint System

The most speculative extension of the constraint-closed paradigm concerns the relationship between constraint closure and consciousness. The suggestion is not that consciousness is a computation but that the phenomenal properties associated with consciousness—unity, coherence, self-referential structure, the sense of inhabiting a single consistent world—may be formal signatures of a particular kind of constraint closure.

The binding problem in consciousness research asks how diverse sensory streams are integrated into a single unified experience. The constraint-based perspective suggests that this integration is precisely global section construction: the brain is solving a sheaf gluing problem in real time, assembling locally processed sensory data into a globally consistent representation of the current world-state. Tononi's integrated information theory [29], which characterizes consciousness in terms of the integrated information generated by a system above and beyond its parts, can be interpreted in this framework as a measure of the degree to which the system enforces global consistency across its internal representations.

The RSVP field theory provides a speculative substrate for this interpretation. If the scalar, vector, and entropy fields of RSVP are understood not as physical fields but as information-theoretic fields representing the flow of constraint and the production of entropy in a cognitive system, then the constraint closure condition becomes a formal characterization of the coherence of experience. A system at rhetorical equilibrium—generating locally plausible but globally unrealizable outputs—would correspond, in this interpretation, to a dissociated cognitive state in which local processing occurs without global integration. This is speculative, and we claim no more than that the mathematical structures are suggestive; but the formal connection is genuine.

45 Moral Constraint and the Geometry of Obligation

The reconstruction of inference as constraint closure has, thus far, been developed primarily in epistemic terms. The framework specifies what it means for a representation to be globally realizable and how such realizability can be achieved through fixed-point convergence. It is natural to ask whether this structure extends beyond epistemology into the domain of

ethics. The argument of this section is that it does, and that certain moral principles can be understood as constraints on admissible world-states in precisely the same sense as physical or observational conditions.

Singer's argument in *Famine, Affluence and Morality* provides a particularly clear formulation of such a constraint [37]. Singer's central principle states that if an agent can prevent a significant harm without sacrificing anything of comparable moral importance, then the agent is morally obligated to do so. Within the constraint-based framework, this principle can be formalized as a condition on admissible world-states. Let X denote a candidate world-state, let $H(X)$ denote the distribution of harms across agents in that state, let A denote the set of actions available to an agent, and let X_a denote the world-state resulting from action $a \in A$. Singer's principle can then be expressed as the requirement that any world-state in which the inequality

$$\exists\, a \in A \text{ such that } H(X_a) < H(X) \quad \text{and} \quad C(a) \ll \Delta H$$

holds but is not acted upon is not fully admissible: it violates a moral constraint that is structurally analogous to a physical or observational inconsistency.

This reframing has several consequences. First, it dissolves the traditional distinction between moral duty and charity. In a constraint-based system, there is no category of supererogatory action in the sense of optional goodness that lies outside the space of constraint satisfaction. If a state violates a constraint, it is inadmissible; if it satisfies all constraints, it is admissible. The failure to prevent avoidable harm when the cost is negligible is therefore not a failure to exceed moral expectations but a failure to satisfy them.

Second, the irrelevance of distance and the number of potential actors follows directly from the structure of the constraint. The condition is defined in terms of the existence of an action that reduces harm at acceptable cost; it does not depend on spatial proximity or on whether other agents could perform the same action. A world-state in which a preventable harm occurs is inadmissible regardless of whether the responsible agent is physically near or one of many capable actors.

Third, the notion of marginal utility can be interpreted as a boundary condition on constraint satisfaction. An agent is required to act until the cost of further action becomes comparable to the harm prevented—the point at which the inequality $C(a) \ll \Delta H$ no longer holds. Beyond this point, additional actions do not reduce the energy functional associated with moral constraint violation and therefore do not contribute to the convergence toward a globally admissible state.

This formulation allows ethical reasoning to be integrated directly into the consistency operator. The energy functional becomes

$$\mathcal{E}(X) = \mathcal{E}_{\text{obs}}(X) + \mathcal{E}_{\text{dyn}}(X) + \mathcal{E}_{\text{struct}}(X) + \mathcal{E}_{\text{moral}}(X),$$

where $\mathcal{E}_{\text{moral}}(X)$ measures the degree to which the world-state fails to minimize avoidable harm subject to cost constraints. A morally improved state is one in which this term is

reduced; a morally optimal state, in the idealized limit, is one in which all avoidable harms have been eliminated consistent with the marginal utility condition.

The significance of this integration is not merely that ethical considerations can be formalized, but that they acquire the same status as other constraints in the inference process. Moral failure is no longer a separate category of judgment applied after the fact; it is a form of structural inconsistency within the space of possible world-states. Singer's argument, when viewed through this lens, is not merely a moral exhortation but a specification of a constraint on the geometry of admissible worlds. The ethical demand it articulates is therefore continuous with the broader project of this monograph: the identification of structures that distinguish realizable from unrealizable configurations, and the construction of systems that converge toward the former.

46 Toward a Unified Theory of Inference

The developments of this monograph converge on the outline of a unified theory of inference. The central claim of this theory is that inference, in all its manifestations—perceptual, scientific, linguistic, computational—is the process of constructing globally consistent representations of the world from locally available, partially inconsistent data. The mathematical structure of this process is sheaf-theoretic descent: the assembly of local sections into global sections through the iterative elimination of cohomological obstruction [24, 25]. The architectural realization of this process is the TARTAN descent engine, which implements the repair hierarchy of gauge alignment, refinement, and state extension.

This unified theory positions the TARTAN architecture not as a specialized system for a particular application domain but as a general-purpose inference platform applicable wherever global consistency is a criterion of success. Scientific reasoning, which must construct theories consistent with experimental data across all relevant scales, is a natural domain of application. Legal reasoning, which must construct judgments consistent with statutory law, precedent, and factual evidence, is another. Medical diagnosis, historical reconstruction, engineering design—in each of these domains, the inference task is the identification of a globally consistent world-state from partial, potentially inconsistent observations.

The theory also positions the relationship between rhetoric and truth in a new light. Rhetoric is not the enemy of truth-oriented discourse but a component of it: the theory of how globally consistent world-states are projected into observer-appropriate representations for communication. The rhetorical machine is not wrong in its orientation toward persuasion; it is incomplete in its indifference to the world-state that persuasive discourse is supposed to represent. The completion of the rhetorical machine—the addition of the constraint structure, the consistency operator, and the global section construction mechanism—is the transformation from a persuasive machine to a truth-tracking system.

Conclusion: From Persuasion to Structure

The transition described in this monograph is not incremental. It is not a matter of making current AI systems more accurate, more reliable, or more honest. It is a categorical shift in what AI systems are for and how they achieve it.

Persuasive machines are organized around the production of discourse that is locally admissible and globally unrestricted. Their success criterion is the response they elicit in human readers—fluency, coherence, apparent relevance—and they have been extraordinarily successful by this criterion. The crisis they have produced is not a failure to meet their own criterion but a collision between that criterion and what we actually need from intelligent systems: outputs that track the world, that can be relied upon in high-stakes domains, and that do not require external verification for every claim.

Constraint-closed systems are organized around a different criterion: global realizability. Their success criterion is the existence of a globally consistent world-state that their outputs correctly describe. This criterion is demanding; meeting it requires the full apparatus of sheaf-theoretic consistency, variational reconstruction, and iterative fixed-point convergence that this monograph has developed. But it is the right criterion, and the architecture required to meet it is now traceable.

The rhetoric of constraint closure is, in the end, an anti-rhetoric: a theory of discourse organized not around the production of persuasive effect but around the reconstruction of realizable worlds. In a cognitive ecology shaped by persuasive machines, this is a genuinely radical proposal. But it is the proposal implied by taking the epistemological situation seriously, and it is the proposal that the mathematics of sheaf theory, variational inference, and categorical descent provides the tools to realize.

From persuasion to realization: this is the direction of travel.

Part XI

Synthesis and Formal Closure

47 The Constraint-Closure Theorem

We are now in a position to formalize the central claim of this monograph: that the distinction between generative plausibility and reconstructive realizability is not merely empirical but structural. The failure of current systems arises from the absence of global consistency constraints, and conversely, the enforcement of such constraints yields a fundamentally different regime of inference.

47.1 Formal Setup

Let $\mathcal{X}$ denote a structured state space, and let $\{\Pi_i : \mathcal{X} \to \mathcal{Y}_i\}_{i \in I}$ be a family of projection operators corresponding to observational, semantic, or discursive constraints. Let $\mathcal{S}$ be a sheaf over a domain Ω such that local sections $s_i \in \mathcal{S}(U_i)$ represent locally admissible data over an open cover $\{U_i\}$. Define the defect cochain

$$\delta_{ij} = \rho_{ij}(s_i) - \rho_{ji}(s_j)$$

where ρ_{ij} are the restriction maps on overlaps $U_i \cap U_j$, and let the global energy functional be

$$\mathcal{E}(X) = \sum_{i \in I} \|\Pi_i(X) - y_i\|^2 + \lambda \, \mathcal{R}(X),$$

where $\mathcal{R}(X)$ encodes structural and entropic constraints.

47.2 Statement

Theorem 3 (Constraint-Closure Theorem)**.** *Let $\mathcal{X}$, $\{\Pi_i\}$, and $\mathcal{S}$ be defined as above. Suppose that the joint projection $\tilde{\Pi} : \mathcal{X} \to \prod_i \mathcal{Y}_i$ is injective on the feasible set $\mathcal{F}$, that $H^1(\Omega, \mathcal{S}) = 0$, and that $\mathcal{E}(X)$ is coercive and admits a unique minimizer X^*. Then the following hold simultaneously. First, there exists a unique $X^* \in \mathcal{X}$ such that $\Pi_i(X^*) = y_i$ for all $i \in I$. Second, a collection of local sections $\{s_i\}$ corresponds to a valid output if and only if it is induced by X^*. Third, no output can be generated that does not arise as a projection of X^*; in particular, hallucination—the nontrivial cohomology class—is structurally impossible. Fourth, X^* is the unique fixed point of the consistency operator $\mathcal{C}$, so that $\mathcal{C}(X^*) = X^*$.*

Proof. The vanishing of $H^1(\Omega, \mathcal{S})$ ensures that all compatible local sections glue to a global section. Injectivity of $\tilde{\Pi}$ guarantees that this global section corresponds to a unique state X^* in $\mathcal{X}$. Coercivity and regularity of $\mathcal{E}$ ensure existence and uniqueness of the minimizer. Since X^* minimizes $\mathcal{E}$, it satisfies all projection constraints and structural

regularization conditions. Any deviation from X^* would increase the energy, so no inconsistent configuration can be stable. The fixed-point characterization follows because the proximal map of a coercive strictly convex functional has its minimizer as its unique fixed point. □

Corollary 1 (Impossibility of Purely Generative Truth Guarantees)**.** *Any system that lacks an explicit mechanism enforcing global consistency—that is, any system that does not minimize a functional equivalent to $\mathcal{E}(X)$ over $\mathcal{X}$—cannot guarantee the elimination of hallucination, regardless of scale.*

Proof. Such a system operates solely within the space of local admissibility and does not enforce vanishing cohomology or projection consistency. Therefore, nontrivial defect classes may persist in its outputs. No improvement in local predictive fidelity can force those defect classes to zero, since they are not part of the objective. □

The theorem formalizes the paradigm shift. In generative systems, outputs are samples from a distribution over local admissibility; truth is external and subject to verification after the fact; hallucination is inevitable because the system has no access to global consistency conditions. In constraint-closed systems, outputs are projections of a single globally consistent state; truth is internal and enforced by construction; hallucination is not reduced but eliminated as a class of admissible outputs. The distinction is not one of degree but of kind.

48 The Relaxed Constraint-Closure Theorem

The Constraint-Closure Theorem establishes the ideal regime. Any practical system operates under finite computational resources, noisy observations, and incomplete coverage of the constraint space, so exact closure is generally unattainable. A theory of *approximate closure* is needed that preserves the structural guarantees of the framework while accommodating these limitations.

48.1 ε-Consistency

For $\varepsilon > 0$, define the ε-feasible set $\mathcal{F}_\varepsilon = \{X \in \mathcal{X} \mid \mathcal{E}(X) \leq \varepsilon\}$. A state $X_\varepsilon \in \mathcal{F}_\varepsilon$ satisfies *ε-closure* if all projection residuals and structural penalties are bounded by ε. We similarly define an ε-defect condition on overlaps: $\|\delta_{ij}\| \leq \varepsilon$ for all i, j.

Theorem 4 (Relaxed Constraint-Closure Theorem)**.** *Suppose the joint projection $\tilde{\Pi}$ is injective up to tolerance ε on $\mathcal{F}_\varepsilon$, the defect cochains satisfy $\|\delta_{ij}\| \leq \varepsilon$, and $\mathcal{E}(X)$ is locally strongly convex in a neighborhood of its minimizer. Then there exists a state $X_\varepsilon \in \mathcal{X}$ such that the local sections $\{s_i\}$ are ε-compatible and admit a global reconstruction up to tolerance ε; small perturbations in observations produce changes in X_ε bounded by $O(\varepsilon)$; and any deviation from realizability is bounded by ε, so no large-scale structural inconsistency can persist.*

The relaxed theorem replaces the binary distinction between truth and hallucination with a controlled continuum. In a generative model, there is no intrinsic bound on the magnitude of inconsistency; errors may be arbitrarily large even when local coherence is high. In contrast, an ε-closed system guarantees that all outputs lie within a controlled neighborhood of a realizable state.

Corollary 2. *Constraint-closure systems operate under bounded approximation error, while purely generative systems admit unbounded error even under scaling.*

Proof. In the constraint-closure framework, all admissible outputs satisfy $\mathcal{E}(X) \leq \varepsilon$, which bounds inconsistency. In generative systems, no such bound exists because inconsistency is not part of the objective. □

The relaxed framework also determines a principled refusal condition: if no X_ε exists within the computational budget such that $\mathcal{E}(X) \leq \varepsilon$, the system has detected an unresolved structural obstruction and the correct behavior is to signal inconsistency rather than produce an approximate answer. This provides a principled alternative to the current paradigm, in which systems generate confident outputs regardless of underlying structural failure.

The conclusion is that real systems need not achieve perfect consistency to be fundamentally different from current architectures. The key property is not exact closure but bounded deviation from closure. A system is reliable not because it never errs, but because the structure and magnitude of its errors are controlled, detectable, and reducible through additional computation.

49 Algorithmic Schema for Constraint Closure

We present the minimal computational structure required to enforce global consistency. The goal is not to prescribe a specific implementation but to define the abstract algorithmic skeleton that realizes the constraint-closure framework.

49.1 State and Core Loop

Let $\{y_i\}_{i \in I}$ denote observed data, $\{\Pi_i\}$ denote projection operators, $X_0 \in \mathcal{X}$ be an initial state estimate, and $\varepsilon > 0$ be the consistency tolerance. The system maintains a current state estimate X_k, a cover $\{U_i\}$ with local sections $s_i = \Pi_i(X_k)$, and defect measures δ_{ij} on overlaps. Reconstruction proceeds as follows: initialize $X \leftarrow X_0$, then repeat the following steps until termination.

In the projection step, compute $s_i \leftarrow \Pi_i(X)$ for each tile. In the defect evaluation step, compute $\delta_{ij} \leftarrow \rho_{ij}(s_i) - \rho_{ji}(s_j)$ on each overlap. In the energy evaluation step, compute $\mathcal{E}(X) \leftarrow \sum_i \|\Pi_i(X) - y_i\|^2 + \lambda \mathcal{R}(X)$. If $\mathcal{E}(X) \leq \varepsilon$, terminate with X_ε. Otherwise, identify the region $\mathcal{A}$ with the largest defect or residual, select the repair operator $\mathcal{C}_\mathcal{A}$, and apply $X \leftarrow \mathcal{C}_\mathcal{A}(X)$.

49.2 Three-Level Repair Hierarchy

The repair operator $\mathcal{C}_{\mathcal{A}}$ is drawn from a structured hierarchy. Gauge repair resolves defects arising from coordinate or representation mismatch: apply a reparametrization $X \mapsto \gamma(X)$ for $\gamma \in \mathcal{G}$ without altering underlying content. Resolution repair resolves defects arising from insufficient local detail: replace the cover $\{U_i\}$ with a finer cover $\{U_i'\}$. Structural repair resolves persistent defects by extending the state space: $\mathcal{X} \mapsto \mathcal{X}' = \mathcal{X} \times \mathcal{Z}$, introducing new latent variables or fields into the ontology.

To handle adversarial or inconsistent inputs, introduce reliability variables $R_i \in [0, 1]$ and modify the energy functional to

$$\mathcal{E}(X, R) = \sum_i R_i \|\Pi_i(X) - y_i\|^2 + \lambda \mathcal{R}(X) + \beta \|R - \mathbf{1}\|_1.$$

The system then jointly updates (X, R), down-weighting observations that cannot be reconciled with any consistent state.

49.3 Termination and Refusal

The algorithm terminates in one of two modes. If $\mathcal{E}(X) \leq \varepsilon$, return X_ε and its projections as the output. If no progress is made after K iterations and $\mathcal{E}(X) > \varepsilon$, the system certifies a structural obstruction and returns a failure signal rather than an approximate answer. This principled refusal condition is what separates the constraint-closure system from a generative one: where a generative model samples regardless of inconsistency, a constraint-closure system refuses to output what cannot be realized.

The schema translates the abstract requirement of global consistency into a concrete iterative procedure. A generative model implements $x_{t+1} \sim p(x_{t+1} \mid x_{\leq t})$, enforcing local sequential coherence with no mechanism for global validation. The TARTAN schema implements $X_{k+1} = \mathcal{C}(X_k)$, where $\mathcal{C}$ encodes global constraints; the output is not a trajectory in token space but a projection of a fixed point in state space. The system does not generate a response; it searches for a state in which the response is necessarily true.

50 A Minimal Counterexample: Local Coherence Without Global Closure

To establish that hallucination is a structural phenomenon rather than a statistical anomaly, we construct a minimal example in which all local constraints are satisfied yet no globally consistent state exists. This example realizes hallucination as a nontrivial cohomology class.

50.1 Construction

Let Ω be a domain covered by three overlapping regions $\Omega = U_1 \cup U_2 \cup U_3$, with nonempty pairwise overlaps $U_i \cap U_j \neq \emptyset$ but empty triple overlap $U_1 \cap U_2 \cap U_3 = \emptyset$. This is the simplest configuration capable of supporting a nontrivial Čech 1-cocycle.

Let $\mathcal{S}$ be a sheaf of scalar-valued assignments, and define local sections $s_1 \in \mathcal{S}(U_1)$, $s_2 \in \mathcal{S}(U_2)$, $s_3 \in \mathcal{S}(U_3)$ such that on pairwise overlaps

$$\rho_{12}(s_1) = \rho_{21}(s_2), \qquad \rho_{23}(s_2) = \rho_{32}(s_3), \qquad \rho_{31}(s_3) = \rho_{13}(s_1) + \Delta,$$

for some nonzero $\Delta \neq 0$. Thus each pair of sections agrees locally except for a cyclic inconsistency that is invisible to any pairwise check. The defect cochain satisfies $\delta_{12} = 0$, $\delta_{23} = 0$, $\delta_{31} = \Delta \neq 0$, defining a nontrivial element of $H^1(\Omega, \mathcal{S})$.

50.2 Absence of a Global Section

Suppose for contradiction that there exists a global section $X \in \mathcal{S}(\Omega)$ with $\rho_i(X) = s_i$ for all i. Then consistency on overlaps requires $\rho_{ij}(s_i) = \rho_{ji}(s_j)$ for all i, j, which contradicts $\delta_{31} = \Delta \neq 0$. Therefore no global section exists.

50.3 Interpretation

This construction satisfies all local admissibility conditions. Each s_i is valid on U_i; each pair (s_i, s_j) is locally consistent in isolation; no pairwise check detects the inconsistency. Yet globally the system is incoherent.

This is precisely the behavior of large language models. Individual statements are plausible, pairwise relations are coherent, but the full set cannot be jointly realized. A generative model operating locally accepts all s_i as valid and generates a discourse corresponding to $\{s_i\}$, which appears coherent but has no realizable underlying state. Since the model does not evaluate global cocycle conditions, it cannot detect the obstruction and produces the inconsistent configuration with full confidence.

In the constraint-closure framework, the defect Δ is explicitly computed. Since $\Delta \neq 0$, the system identifies a structural obstruction and initiates a repair step or declares failure if no resolution exists. The system cannot output an inconsistent configuration because inconsistent configurations are not admissible states.

The counterexample establishes that hallucination is not a consequence of insufficient data or imperfect optimization. It is the manifestation of a nontrivial cohomology class arising from the absence of global compatibility constraints. No amount of scaling can eliminate such obstructions unless the architecture enforces the vanishing of $H^1(\Omega, \mathcal{S})$. The transition from generative modeling to constraint closure is therefore not a matter of degree but of kind: it replaces a system that samples locally valid fragments with one that constructs globally consistent objects.

51 From Rhetoric to Reconstruction: A Paradigm Shift

The preceding analysis has established a precise diagnosis of the current limitations of artificial intelligence systems and a corresponding prescription for their resolution. What has often been described as a problem of hallucination is, in fact, the visible symptom of a deeper structural condition: the absence of global constraint closure.

Contemporary large language models operate within a regime of local admissibility. They generate outputs that are fluent, coherent, and contextually appropriate, navigating the statistical structure of discourse with remarkable precision. In this sense, they are highly effective rhetorical systems. As Gunkel observes [1], they do not fail because they are defective truth-tellers; they succeed because they are persuasive machines. Their outputs are optimized for plausibility within a linguistic manifold, not for realizability within a structured state space.

This distinction is not merely philosophical; it is formal. A system that operates purely within the space of local sections, without enforcing the existence of a global section, necessarily admits configurations that are locally consistent but globally incoherent. The minimal counterexample demonstrates that such configurations correspond to nontrivial cohomology classes. Hallucination is therefore not an accidental byproduct of insufficient scale or imperfect training; it is a structural feature of architectures that lack gluing conditions. The Constraint-Closure Theorem formalizes the ideal case in which global consistency is exact and hallucination is eliminated as a class of admissible outputs. The Relaxed Constraint-Closure Theorem extends this result to practical systems, demonstrating that even under finite computation inconsistency can be bounded, detected, and controlled. The algorithmic schema shows that these properties are computationally realizable through iterative repair and convergence.

Taken together, these results establish that the transition from generative to reconstructive intelligence is not incremental but categorical. Scaling improves performance within the generative regime but does not alter the underlying structure. Constraint closure, by contrast, introduces a new objective: not the maximization of plausibility but the enforcement of realizability.

This shift resolves the tension identified in the philosophical literature. The traditional opposition between rhetoric and truth is revealed to be an artifact of operating within discourse space. When systems are confined to this space, persuasion can outrun structure, and coherence can exist without correspondence. Constraint closure removes this asymmetry. A system cannot produce a statement that does not arise from a consistent state, because such a statement has no representation within the system's admissible set.

In this sense, the framework does not attempt to make machines care about truth in any anthropomorphic sense. It renders truth a structural property of the system's operation. Validity is not imposed externally through verification; it is enforced internally through constraint satisfaction. The measure of intelligence is not how well a system can speak, but whether what it says can exist.

References

[1] Gunkel, D. J. (2026). Persuasive machines: Large language models and the art of rhetoric. *AI & Society*. https://doi.org/10.1007/s00146-026-03022-9

[2] Plato. *Gorgias*. Trans. D. J. Zeyl. Hackett, 1987.

[3] Plato. *Phaedrus*. Trans. A. Nehamas and P. Woodruff. Hackett, 1995.

[4] Aristotle. *On Rhetoric: A Theory of Civic Discourse*. Trans. G. A. Kennedy. Oxford University Press, 2007.

[5] Austin, J. L. (1962). *How to Do Things with Words*. Harvard University Press.

[6] Searle, J. R. (1969). *Speech Acts*. Cambridge University Press.

[7] Perelman, C. and Olbrechts-Tyteca, L. (1969). *The New Rhetoric: A Treatise on Argumentation*. University of Notre Dame Press.

[8] Habermas, J. (1984). *The Theory of Communicative Action, Vol. 1*. Beacon Press.

[9] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460.

[10] Weizenbaum, J. (1976). *Computer Power and Human Reason*. W. H. Freeman.

[11] Dennett, D. C. (1987). *The Intentional Stance*. MIT Press.

[12] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of FAccT 2021*.

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

[14] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33.

[15] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

[16] Cheng, N., Broadbent, G., and Chappell, W. (2025). Cognitive loop via in-situ optimization: Self-adaptive reasoning for science. *arXiv preprint arXiv:2508.02789*. https://arxiv.org/abs/2508.02789

[17] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423.

[18] Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

[19] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*, 2nd edition. Wiley-Interscience.

[20] Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1), 1–7.

[21] Solomonoff, R. J. (1964). A formal theory of inductive inference. *Information and Control*, 7(1), 1–22.

[22] Bredon, G. E. (1997). *Sheaf Theory*, 2nd edition. Springer, New York.

[23] Bott, R. and Tu, L. W. (1982). *Differential Forms in Algebraic Topology*. Springer, New York.

[24] Mac Lane, S. (1998). *Categories for the Working Mathematician*, 2nd edition. Springer, New York.

[25] Riehl, E. (2017). *Category Theory in Context*. Dover, New York.

[26] Spivak, D. I. (2014). *Category Theory for the Sciences*. MIT Press.

[27] Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd edition. Cambridge University Press.

[28] Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

[29] Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *Biological Bulletin*, 215(3), 216–242.

[30] Evans, L. C. (2010). *Partial Differential Equations*, 2nd edition. American Mathematical Society, Providence.

[31] Prigogine, I. and Stengers, I. (1984). *Order Out of Chaos*. Bantam Books, New York.

[32] Misner, C. W., Thorne, K. S., and Wheeler, J. A. (1973). *Gravitation*. W. H. Freeman, San Francisco.

[33] Penrose, R. (2004). *The Road to Reality*. Jonathan Cape, London.

[34] Jacobson, T. (1995). Thermodynamics of spacetime: The Einstein equation of state. *Physical Review Letters*, 75(7), 1260–1263.

[35] Davidson, D. (1984). *Inquiries into Truth and Interpretation*. Oxford University Press.

[36] Barwise, J. and Perry, J. (1983). *Situations and Attitudes*. MIT Press.

[37] Singer, P. (1972). Famine, affluence, and morality. *Philosophy & Public Affairs*, 1(3), 229–243.