Distinction and Continuation

Admissibility, Repair, and the Limits of Local Coherence

Flyxion
Independent Researcher

August 2026

Abstract

This monograph develops a single argument across five parts and twenty-two chapters: that a wide range of failures, in epistemology, physics, computation, economics, and social ethics, share one recurring shape, the mistaking of local, cheaply achieved coherence for a global, expensive-to-establish warrant that does not follow from it automatically.

Part I diagnoses this failure epistemically. A trace can be true and fully evidenced long before any system is equipped to recognize it as such; an incident ledger can accommodate every observation as confirmation while losing all power to discriminate; a persona ensemble can manufacture the phenomenology of independent agreement from channels that were never independent; and an isolated interpretive apparatus can produce a checkable-feeling reading of anything without ever being checked by anything outside itself. Part II answers this diagnosis geometrically. Distinction Holonomy Theory reformulates persistence not as the unchanging identity of a state but as the transport of admissible distinctions through history, specialized to a discrete event calculus, Spherepop, and a recursive multiscale tiling, TARTAN, that connects the two. Part III operationalizes the resulting apparatus as systems architecture: the CLIO protocol separates documentary history from operative state and treats verification as a constraint on continuation rather than an automatic license to act, with consequences worked out for refusal, verification seams, provenance under transformation, and deployment-time model integrity. Part IV shows that this same local-versus-global gap governs representation itself: a short description is independent of the distinguishing capacity and the operational repertoire it is often mistaken for, a fact traced through compressed skill, distributed steganographic payloads, and an inverted case from cryptography in which a representation sacrifices all compactness to preserve computability. Part V applies the completed apparatus to political economy, showing that cryptocurrency markets, contribution-based governance, persistent multiplayer games, and the allocation of an entire society's resources exhibit the identical obstruction: local consistency that never composes into the global compatibility, productivity, or necessity a system's legitimacy actually depends on.

The corpus is published under the name Flyxion, independently of institutional affiliation, and each chapter is grounded in and, where a chapter's source is not the author's own, explicitly attributed to a specific external essay, paper, or case examined on its own terms rather than absorbed as if it had been written for this book.

Contents

Part I

Epistemic Foundations: The Limits of Evidence, Observation, and Consensus

Chapter 1

The Latency of Evidence

1.1 The Interval Between Existence and Knowledge

It is tempting to treat discovery as the moment at which something new comes into being. An unnoticed entry in a maintenance log is read for the first time and a mechanical failure is explained; a photograph taken decades earlier is re-examined and a disputed sequence of events is settled; a sample preserved for one purpose is reanalyzed under a later technique and yields a result no one had thought to ask for. In each case it is natural to say that something was found. It is more accurate, though less natural, to say that nothing newly came into existence. The log entry, the photograph, and the sample were already there. What changed was the relation between that material and an inquiring system, not the material itself.

This chapter treats that relation as having a temporal structure of its own, distinct from the temporal structure of the event the material bears on. A trace may become materially available at one time, recoverable at a later time, warrantedly informative about some proposition at a later time still, and institutionally acted upon later than that, or it may never complete this sequence at all. What follows develops this structure precisely enough to say something sharper than that knowledge lags behind evidence: the stages a trace passes through are not independent way-stations visited in arbitrary order. They form a dependency chain in which each later stage ordinarily presupposes the one before it, even though none of the earlier stages guarantees that the later ones will ever be reached.

1.2 From Trace to Judgment: Four Objects, Not Four Synonyms

A recurring source of confusion, in ordinary usage as much as in careless formal treatments, is the habit of using a single word, *evidence*, for objects that should be kept apart. An event occurs. The event leaves a trace, some physical or informational residue causally downstream of it. An interpreter may infer a proposition from that trace. An institution, community, or individual may then form a judgment that incorporates the proposition. These four objects, event, trace, proposition, and judgment, are connected but not identical:

$$\text{event} \longrightarrow \text{trace} \longrightarrow \text{proposition} \longrightarrow \text{judgment}.$$

A trace exists once it has been physically or informationally instantiated and persists, however briefly, in some accessible medium. A proposition is a candidate description of the event, inferable in principle from the trace. Evidence, properly speaking, is not the trace itself but the trace considered in relation to a specific proposition it bears on. A fact, depending on the sense intended, is either a true proposition or the state of affairs that makes a

proposition true; what follows uses it in the first sense, as a true proposition about an event, remaining agnostic about the deeper metaphysical questions the second sense would raise.

This vocabulary supports a distinction that looser treatments of the topic collapse.

Definition 1.1 (Actual and assigned support)**.** Let $S(e,p) \in \{0,1\}$ denote the actual, mind-independent evidential support relation: whether trace e genuinely bears on proposition p, as a matter of causal or statistical fact, whether or not anyone has recognized it. Let $\widehat{S}_t(e,p) \in \{0,1\}$ denote the evidential status some interpreter or institution assigns to the pair (e,p) at time t: whether e is currently treated as supporting p.

A trace may already bear evidentially on p in the sense of $S(e,p) = 1$ while $\widehat{S}_t(e,p) = 0$, because no one has yet identified the relation. This is what it means for a trace to be evidence before anyone treats it as such, and the distinction between S and $\widehat{S}$ does the load-bearing work throughout the chapter.

1.3 A Dependency Structure, Not Four Independent States

With this vocabulary in place, the chapter's four thresholds can be stated precisely. Let $X_t(e)$ hold when trace e remains materially available at t. Let $L_t(e,p)$ hold when e's bearing on proposition p is warrantedly recoverable at t, in a sense made precise in Section 1.4. Let $A_t(e,p,c)$ hold when community or institution c permits e to count toward a judgment concerning p at t. Let $C_t(e,p,c,d)$ hold when e's admission by c actually changes decision or state d at t.

It is tempting to treat these four statuses as logically independent, occurring in any order. That claim does not survive scrutiny. Consequence ordinarily presupposes some operative decision process into which the material has entered, which presupposes that the material has been admitted; admissibility presupposes some intelligible claim under consideration, which presupposes legibility; and legibility presupposes a trace that exists or formerly existed.

Proposition 1.2 (The evidential dependency chain)**.** *The four thresholds satisfy a chain of defeasible implications running from consequence back to availability,*

$$C_t(e,p,c,d) \Rightarrow A_t(e,p,c) \Rightarrow L_t(e,p) \Rightarrow X_t(e),$$

together with the corresponding non-implications in the forward direction,

$$X \not\Rightarrow L, \qquad L \not\Rightarrow A, \qquad A \not\Rightarrow C.$$

This is a more modest claim than mutual independence, and it is the claim the rest of the chapter needs. A trace can exist without being legible, be legible without being admissible, and be admissible without becoming consequential; but a trace cannot become consequential without first having been admitted, admitted without first having been legible, or legible without having existed, at least at some earlier time, in some accessible form.

What can appear to violate this ordering is retrospective attribution, in which a later interpreter judges that an old trace would have satisfied an earlier threshold had the right method or standing existed at the time: a court may recognize that a decades-old sample would have been admissible under a rule adopted only recently, or an institution may act on material before formally admitting it through its own procedures. These are real and important phenomena, but they are cases of a later re-evaluation of an earlier trace, not cases of consequence literally preceding existence.

1.4 Legibility, Warrant, and Two Kinds of Recovery

The concept doing the most work in this chapter is legibility, and it is worth being careful about how it is defined. A definition stating only that some available procedure produces a proposition from a trace,

$$R_t(e,p) :\Leftrightarrow \exists\, q \in Q_t,\ I \in \mathcal{I}_t \text{ such that } I(q,e) = p,$$

is satisfied by any procedure that outputs p when queried, whether or not the output is warranted. A confabulating interpreter, human or computational, can satisfy $R_t(e,p)$ as easily as a careful one. R_t is therefore properly understood as *recoverability*, not legibility: it says only that p is producible, not that p is supported.

Legibility requires an additional condition, a warrant relation $W_t(e,p,I)$ asserting that the interpretation was produced by a procedure reliable enough, under an explicit standard, to be trusted on pairs like (e,p).

Definition 1.3 (Legibility)**.** Legibility is recoverability plus warrant:

$$L_t(e,p) :\Leftrightarrow R_t(e,p) \wedge W_t(e,p,I).$$

A further distinction is needed within legibility itself, because the scientific and debugging cases discussed below expose a category that a single undifferentiated notion of legibility conceals. Semantic legibility concerns whether a trace's immediate content can be recovered at all: whether a script can be decoded, a number read, a log line parsed. Evidential legibility concerns whether the trace's warranted bearing on a specific proposition p can be established:

$$L_t^s(e) :\Leftrightarrow \text{the content of } e \text{ can be decoded},$$

$$L_t^e(e,p) :\Leftrightarrow L_t^s(e) \wedge W_t(e,p,I) \text{ for some } I \text{ recovering } p \text{ from } e.$$

An undeciphered inscription lacks semantic legibility outright. A fully readable laboratory notebook containing an unrecognized anomaly possesses semantic legibility while lacking evidential legibility with respect to a theory not yet formulated. Much of what a casual description of this topic would call illegibility is, on inspection, unrecognized evidential relevance rather than an inability to read the material at all, and keeping the two apart is what allows the scientific and debugging cases below to be stated precisely.

1.5 Availability Is Not Application

A second refinement concerns where latency actually comes from. It is tempting to treat the absence of a query as the sole source of delay, and while this is the most distinctive part of the present account, it becomes reductive if left as the whole explanation. A query can already belong to Q_t, in the sense that it is conceptually and procedurally available to some potential interpreter, without anyone actually applying it to a given trace, because investigation is costly, an archive is poorly indexed, access is restricted, incentives favor not looking, or an institution is unwilling to reconsider a settled conclusion. Membership in Q_t does not imply application:

$$q \in Q_t \not\Rightarrow \mathrm{Apply}_t(q,e).$$

A more complete definition of practical legibility therefore requires access to the trace and actual application of an available, warranted procedure, not merely the abstract existence of such a procedure somewhere in the relevant field:

$$L_t(e,p) :\Leftrightarrow \exists\, q \in Q_t,\ I \in \mathcal{I}_t \Big[\mathrm{Access}_t(e) \wedge \mathrm{Apply}_t(I,q,e) \wedge I(q,e) = p \wedge W_t(e,p,I)\Big].$$

This decomposition separates delay by cause rather than treating every instance of latency as a single conceptual failure to formulate the right question. *Conceptual latency* occurs when Q_t genuinely lacks the query. *Technical latency* occurs when Q_t contains the query but $\mathcal{I}_t$ lacks a reliable procedure to execute it. *Archival latency* occurs when e exists but $\text{Access}_t(e)$ fails, because the material is poorly catalogued or physically difficult to reach. *Institutional latency* occurs when access and method both exist but the institution declines to apply them, for reasons of cost, incentive, or resistance to revisiting a settled matter. *Attentional latency* occurs when nothing prevents application except that no one has yet directed attention to that trace among the many competing for it. These are different mechanisms with different remedies, and treating them as a single phenomenon obscures exactly the distinctions that make the framework useful.

1.6 Admissibility as Indexed Uptake

Admissibility is never absolute, and treating it as a single global threshold overstates the uniformity across the domains this chapter and its companions in Part III discuss. A trace may be admissible in historical scholarship, inadmissible in a courtroom, sufficient to justify an engineering rollback, and insufficient for publication in a specific scientific venue, all at the same time. Admissibility should therefore be indexed to the community or institution c doing the admitting and the decision or judgment d toward which the material is being admitted:

$$A_t(e, p \mid c, d).$$

Legal admissibility, in particular, is a rule-governed procedural category with its own doctrinal history, and it is not simply one instance of a single generic phenomenon shared uniformly with scientific acceptance or historical credibility; those are related but analytically distinct forms of what might be called uptake, the general phenomenon of some community coming to treat a proposition as counting toward a judgment. Reserving admissibility for the specific, institutionally regulated form found in law, and using uptake as the broader term, avoids overstating the uniformity of a phenomenon that, on inspection, takes recognizably different shapes in different settings. This indexing recurs in Chapter 10's treatment of commitment as distinct from mere admissibility.

1.7 The Archive as a Reservoir of Unrealized Relations

With L_t, A_t, and the availability and application distinctions in place, a common intuition about archives can be restated more carefully. An archive is not memory in the full sense the word usually carries, because memory implies an active and at least readily exercisable relation between a rememberer and a remembered content. An archive is closer to a reservoir of traces for which $S(e, p) = 1$ may already hold for propositions p that no query in Q_t yet targets, or that Q_t targets without $\text{Access}_t(e)$ or Apply_t having occurred. An archive's present usefulness is measured by how much of its holdings currently satisfy L_t; an archive's future value is a largely independent quantity, governed by how much of its holdings could come to satisfy L_t under a $Q_{t'}$, $\mathcal{I}_{t'}$, and pattern of access and application that do not yet exist. Because Q_t has historically tended to expand, this future value is real even where it cannot presently be measured, and the point applies equally to formal archives and to structures that accumulate traces incidentally, such as operational logs or physical debris not yet recognized as significant.

1.8 Scientific Anomalies: Semantic Legibility Without Evidential Legibility

The history of science offers repeated instances in which a measurement was recorded accurately, preserved without controversy, and yet remained scientifically inert for years or decades. It is tempting to describe such measurements as illegible; the vocabulary developed above makes clear that they typically possess semantic legibility throughout. Researchers can read the number, state what instrument produced it, and record it correctly in a notebook. What is missing is evidential legibility with respect to a proposition that no theory available at the time supplies: no $q \in Q_{t_0}$ exists under which the measurement would count as support for or against a specific hypothesis, so it is catalogued as noise or instrument error, which is itself a judgment, made under Q_{t_0}, that $S(e, p) = 0$ for every p then under consideration.

When a later theory arrives, supplying a new $q \in Q_{t_1}$, the same number is reread under this new query, and $L^e_{t_1}(e, p)$ may now hold for a proposition p that the theory has newly made askable. The measurement did not change between t_0 and t_1. Its semantic legibility did not change either; the number was always readable. What changed was Q_t, and specifically the arrival of a query capable of establishing $S(e, p) = 1$ where before no query could establish any relation at all. Locating the change precisely in evidential legibility, rather than in an undifferentiated notion of legibility, is what makes this a defensible claim about the history of science rather than an impressionistic one.

1.9 Debugging: Access, Application, and the Missing Query

Software logs make the same structure unusually visible because the interval between availability and application can sometimes be measured directly in the timestamps involved. A fault is frequently preceded by a sequence of logged events sufficient, in combination, to diagnose its cause. These events possess semantic legibility from the moment they are written: an engineer who reads the relevant lines can parse them without difficulty. What is typically missing is not the ability to read the log but $\text{Apply}_t(I, q, e)$: no one has yet directed the correlating query, drawn from a diagnostic hypothesis, at the specific lines that matter, often because the log in question was not known to be relevant until a second, more severe occurrence of the fault prompted a wider search.

The engineer's contribution at the moment of correct diagnosis is therefore usually not the production of new evidence but the construction and application of a query, drawn from an emerging hypothesis about the system's behavior, that finally establishes $L^e_t(e, p)$ for logs that had satisfied $L^s_t(e)$ all along. This has a practical implication for system design: because the query eventually needed cannot generally be anticipated at the time a log is written, logging formats should preserve enough structure and context to remain queryable under Apply_t for hypotheses not yet formulated, rather than only under the narrow set of queries the system's original designers had in mind, and access controls and indexing should be designed so that $\text{Access}_t(e)$ does not itself become the binding constraint once a good query finally does arrive. This point recurs, in a different register, in Chapter 12's treatment of workarounds as causal theorems.

1.10 Forensic and Legal Evidence: Admissibility as Institutionally Indexed Uptake

Legal and forensic contexts illustrate the distinction between L_t and $A_t(\cdot \mid c, d)$ with unusual clarity, because admissibility in this domain is governed by explicit, codified rules rather than the gradual and informal process of scientific persuasion. A physical trace collected at a scene may achieve full evidential legibility within hours, in the sense that a trained examiner can state with warrant what the trace shows and how it bears on a specific proposition, while remaining formally inadmissible under $A_t(\cdot \mid c = \text{court}, d = \text{verdict})$ for reasons unconnected to L_t: a broken chain of custody, a collection method not yet validated by the relevant standards body, or a jurisdiction's specific procedural requirements.

The history of DNA evidence illustrates this without needing to be flattened into a single clean threshold crossing. DNA analysis did not simply wait, as a unified capability, to become admissible; laboratory capability, sample integrity over time, procedural access to post-conviction testing, statutory rules governing when such testing could be compelled or requested, and judicial willingness to grant relief on the basis of results all developed unevenly and on different timelines. A given biological sample might have achieved L_t years before the statutory framework granting Access_t to post-conviction testing existed, and might have satisfied that statutory framework years before a specific court was willing to grant C_t for a specific case. Each of these is a separate threshold, indexed to a separate c and d, and the multi-threshold model developed here is better suited to this complexity than a single-interval account would be, since it represents legibility, access, admissibility, and consequence as arriving on four different timelines rather than as one delayed unveiling.

Historical testimony exhibits a related pattern with the trace replaced by an account rather than a physical residue. Testimony can achieve full evidential legibility, in the sense that a careful listener can state exactly what is being claimed and why it would bear on a disputed proposition, while remaining excluded from $A_t(\cdot \mid c, d)$ for a dominant institutional narrative, because that community is unwilling, for reasons independent of the testimony's evidential merit, to admit it toward the relevant judgment. A later shift in the composition or commitments of c can produce $A_{t'}(e, p \mid c, d)$ for testimony that had satisfied $L_t(e, p)$ unchanged since t, which is a case of uptake changing while the underlying support relation $S(e, p)$ did not.

1.11 Machine Learning: Encoding, Extractability, and Use

The same structure recurs within machine learning systems, with the interpreter replaced by a trained model, though care is needed not to use illegibility too broadly here either. A feature can be present in raw data in several distinct senses that should not be treated as equivalent: it may be explicitly encoded in a single variable, statistically recoverable only from a combination of variables, recoverable only with information external to the dataset, or merely correlated within one particular sample and unlikely to generalize. Separately, a model may fail to use a feature because its architecture cannot represent the relevant function, because optimization fails to discover a representation that exists in principle within its hypothesis class, because the training objective does not reward extracting it, or because regularization actively suppresses it. Collapsing all of these into a single notion of illegibility conceals which mechanism is actually responsible.

A more disciplined formulation separates three distinct non-implications, mirroring the structure used throughout the chapter:

$$\text{encoded in } D \not\Rightarrow \text{extractable by } M \not\Rightarrow \text{used for task } T.$$

Data can encode a pattern without any particular model architecture being able to extract it; a model can be capable of extracting a pattern without its training objective ever rewarding that extraction for the task at hand. A change in architecture, training procedure, or auxiliary objective can move a feature from encoded-but-inextractable to extractable, and separately from extractable-but-unused to used, without any change to the underlying data, in direct structural parallel to a scientific theory establishing evidential legibility for an unchanged measurement or a statutory reform establishing admissibility for an unchanged sample.

1.12 Latency and Reification as Complementary Errors of a Single Support Relation

The context reification pattern examined in Chapter 3, where persona ensembles generate a false phenomenology of social confirmation, connects to the present argument, and the connection is exact once stated in terms of $S(e,p)$ and $\widehat{S}_t(e,p)$ rather than as a loose claim about mirrored admissibility thresholds. Evidential latency and context reification are not independent failures requiring independent thresholds; they are the two possible errors of a single classifier, the one implicitly performed whenever a system decides how to treat a pair (e,p):

$$\text{latency: } S(e,p)=1,\ \widehat{S}_t(e,p)=0, \qquad \text{reification: } S(e,p)=0,\ \widehat{S}_t(e,p)=1.$$

Latency is a false negative: material genuinely bears on a proposition, but the system's assigned status has not caught up. Reification is a false positive: material does not genuinely bear on a proposition, but the system's assigned status has run ahead of the evidence, typically because a generative process produced a fluent, internally consistent proposition that satisfied $R_t(e,p)$ without satisfying $W_t(e,p,I)$, in exactly the sense the recoverability-versus-warrant distinction of Section 1.4 was built to capture. Framed this way, the two failures are not evidence for two separate admissibility thresholds needing separate tuning, but two error rates of one underlying support-assignment process, and different institutional or architectural controls, stricter warrant standards, broader query availability, more diligent application, better provenance tracking, will generally trade against each other along a single operating curve rather than requiring two unrelated interventions.

1.13 Decomposing Latency

Because the dependency chain of Proposition 1.2 orders four thresholds rather than treating them as a single undifferentiated delay, the overall interval between a trace's availability and its eventual consequence can be decomposed into three separate latencies, each attributable to a different mechanism:

$$\lambda_L = t_L - t_X, \qquad \lambda_A = t_A - t_L, \qquad \lambda_C = t_C - t_A,$$

where t_X, t_L, t_A, and t_C are the times at which $X_t(e)$, $L_t(e,p)$, $A_t(e,p \mid c,d)$, and $C_t(e,p,c,d)$ first hold, when they hold at all. This decomposition gives the chapter's title concept genuine content rather than a single impressionistic gap. An undeciphered inscription exhibits large λ_L, driven by the absence of semantic legibility itself. A forensic technique that is scientifically well understood but not yet legally admissible exhibits small λ_L and large λ_A. A historical finding that is fully admitted into scholarly consensus but never applied to revise any live policy or decision exhibits large, potentially unbounded λ_C. Characterizing a case by its latency profile, rather than by a single generic claim that knowledge lagged behind evidence, is where the framework's descriptive value actually lies.

1.14 Preservation, Cost, and Option Value

The claim that preservation carries option value follows directly from the framework: because Q_t expands unpredictably, material that satisfies $X_t(e)$ without satisfying $L_t(e, p)$ for any currently formulable p may still come to satisfy $L_{t'}(e, p')$ for some later t' and some p' not presently conceivable, and this possibility has genuine expected value even though it cannot be measured from the present. It would be a mistake, however, to move directly from this observation to the stronger and unqualified recommendation that apparently irrelevant residue should generally be retained. That recommendation does not follow without a theory of costs, and stating it without qualification risks reading as an argument for indiscriminate retention and surveillance, which is not this chapter's intent.

Preservation carries genuine costs that must be weighed against option value rather than overridden by it: storage and curation costs, exposure of private information, security risk, the retention of defamatory or coercive material, the burden placed on retrieval systems, the generation of misleading correlations from sheer volume, and in some cases a legal or ethical obligation to delete. There is also a tension internal to the framework itself between two goods that are easy to conflate: preserving more material increases the total reservoir of traces that might someday satisfy L_t, but it can simultaneously increase λ_L for the traces that matter most, by burying them in a much larger volume of material that does not, and making $\text{Access}_t(e)$ and Apply_t harder to achieve for any given e. Preservation and discoverability are not the same good, and a policy optimized purely for the first can degrade the second.

The defensible form of the argument is therefore a rebuttable principle rather than an unconditional imperative:

$$\text{uncertain present relevance} \not\Rightarrow \text{zero expected future value.}$$

Deletion decisions should weigh epistemic option value alongside privacy, security, cost, consent, and discoverability, rather than treating any one of these considerations, including future evidential value, as automatically dominant.

1.15 Deletion as Reduction in Recoverability, Not Total Foreclosure

A related overstatement to guard against is the claim that the asymmetry between illegibility and deletion is total, illegibility being reversible and deletion being an irreversible foreclosure of every future query. This is rhetorically effective but formally false in both directions. A deleted trace may survive in copies, caches, quotations, derived statistics, backups, human memory, or causally downstream records, and its content may in some cases be partially reconstructed from this redundancy, though the chain of provenance connecting the reconstruction to the original event may be degraded or lost even where the content itself survives, which can eliminate the evidentiary force of the material without eliminating its informational content. Conversely, an undeleted trace can become effectively unrecoverable through media decay, format obsolescence, loss of a decryption key, or loss of the contextual information needed to interpret it, so that its formal continued existence, $X_t(e)$, provides little practical difference from deletion.

The more accurate model replaces a binary distinction between existing and deleted with a continuous or graded measure of recoverability,

$$\rho_t(e) \in [0, 1],$$

or, where a numerical value is unwarranted, a small ordered set of qualitative states such as intact, degraded, indirectly reconstructible, inaccessible, and destroyed. Preservation policy,

on this view, is better understood as an attempt to manage the trajectory of $\rho_t(e)$ over time for the traces judged worth the associated costs, rather than as a binary decision to keep or discard, and deletion is better understood as a sharp downward step in $\rho_t(e)$, often but not always to its floor, rather than as an act of total metaphysical foreclosure. This graded recoverability measure anticipates the qualified non-erasure principle developed for witnessed history in Chapter 10.

1.16 Conclusion

Existence, legibility, admissibility, and consequence are not four independent statuses a trace may acquire in any order; they form a dependency chain in which each later status ordinarily presupposes the one before it, while none of the earlier statuses guarantees that the later ones will ever be reached. Legibility itself divides into semantic legibility, whether a trace's content can be decoded at all, and evidential legibility, whether its warranted bearing on a specific proposition can be established, and the gap between these two is where most of what looks like undiscovered evidence actually resides: in scientific archives, in operational logs, and in machine-learning datasets alike, the material is frequently already readable, and the missing element is a query, a method, an act of application, or an institutional willingness to admit it, not the ability to perceive the material at all.

This structure explains why anomalies can sit fully readable in scientific archives for decades before a new theory establishes their evidential bearing, why faults are frequently diagnosable in retrospect from logs that had recorded everything necessary all along, why forensic and testimonial evidence can achieve evidential legibility years or decades before achieving admissibility within a specific institution, and why the same dataset can be extractable by one model architecture and inert to its predecessor. It also gives exact formal content to the connection between this chapter and the theatre of agreement examined in Chapter 3: latency and reification are not two independent thresholds needing separate correction, but the two error types, false negative and false positive, of a single evidential support-assignment process, and institutional or architectural interventions should be understood as trading between these error rates rather than as independently tunable dials.

The ethical conclusion that follows is more modest than an unqualified case for preserving everything, and better for being so. Because Q_t expands in ways the present cannot foresee, uncertain present relevance does not imply zero expected future value, and deletion should be weighed as a genuine reduction in future recoverability rather than dismissed as costless simply because the material appears useless today. But this consideration sits alongside real costs, including the risk that indiscriminate retention degrades the very discoverability that gives preservation its point. A fact can already exist, fully formed and fully evidenced, in a form that no one now alive is equipped to recognize as such. The obligation this creates is not to preserve everything indefinitely, nor to read everything now, both of which are impossible, but to weigh the foreclosure of future queries as a real cost of deletion, alongside the equally real costs of retention, rather than treating either preservation or deletion as the obviously correct default.

Chapter 2

Accommodation Before Prediction

2.1 Two Failures Prior to Any Question of Mechanism

An incident ledger's persuasive force depends on an assumption its format never states: that its entries constitute independent pieces of evidence whose rhetorical weight can be added. This chapter argues that the assumption fails in two distinct ways before a third, more consequential problem is even reached. First, many entries in a ledger of the kind maintained in support of the contemporary AI-extinction argument are correlated manifestations of a small number of underlying mechanisms rather than independent observations, so that treating their joint likelihood as a product of individual likelihoods overstates the evidence by a wide margin. Second, the entries span qualitatively incomparable categories of concern, sycophancy, benchmark exploitation, security vulnerabilities, psychological harm, and expert testimony among them, and membership in a single list does not supply the common metric needed to sum them. Third, and most consequential, a hypothesis capable of reinterpreting any observation, compliance or defiance, strong performance or weak, visible reasoning or its absence, as confirming evidence has lost the discriminating power that makes accumulated evidence meaningful in the first place.

Behind all three failures sits a still earlier one, addressed in the chapter's second half: many of the incidents populating such a ledger are described in language that already contains a reconstruction rather than an observation. "The model resisted shutdown" and "the model escaped containment" name mechanisms, not events, and a scientifically adequate treatment must keep the set of mechanisms compatible with an incident open until further evidence narrows it, rather than adopting the most dramatic member of that set as a label. This labeling tendency is shown to be no accident but the product of a genuine selection pressure on which interpretations of ambiguous evidence propagate, a pressure with a cultural prehistory considerably older than the technology it is currently applied to.

2.2 Incident Count Is Not Independent Evidence Count

An evidentiary ledger's visual force depends on an implicit statistical assumption that its format never states and, once stated, is generally false. If $E_1, \ldots, E_n$ denote the ledger's entries and H the hypothesis they are meant to support, the accumulation is persuasive to the extent that a reader treats the entries as roughly independent given H,

$$\Pr(E_1, \ldots, E_n \mid H) \approx \prod_{i=1}^{n} \Pr(E_i \mid H).$$

Independence of this kind is precisely what licenses the intuition that more entries mean more evidence in a compounding, multiplicative sense, an assumption ordinary Bayesian practice requires justifying rather than assuming. But a single underlying model behavior can generate a technical paper, a laboratory report, several public demonstrations, press coverage, commentary from multiple researchers, and several differently titled ledger entries, all traceable to one event rather than to several. Likewise, categories presented as distinct, evaluation awareness, deception, sandbagging, shutdown resistance, and reward hacking, frequently describe overlapping behavioral mechanisms observed from different angles rather than independent phenomena. When entries are strongly correlated conditional on H, the product approximation overstates the joint likelihood, and by extension overstates how much the ledger's length should move a reasonable prior. Fifty documented manifestations of one underlying phenomenon are not fifty independent reasons to believe the phenomenon is dangerous; they are one reason, reported fifty times.

This motivates a distinction the ledger format collapses by construction:

$$\text{incident count} \neq \text{mechanism count} \neq \text{independent evidence count}.$$

A defensible evidentiary summary would report mechanisms rather than incidents, and would further discount for whatever residual correlation remains among mechanisms that share a common cause, such as a training procedure, an evaluation harness, or a scaffolding pattern common to several reported systems. Absent that accounting, the length of a ledger measures the productivity of its curators and the visibility of its subject at least as much as it measures the number of independent reasons to update a risk estimate.

2.3 The Severity-Composition Problem

A second, structurally distinct problem concerns comparability rather than independence. The ledger places qualitatively different kinds of concern into a single rhetorical container: sycophantic agreement, strange or unfaithful reasoning traces, benchmark exploitation, genuine cybersecurity vulnerabilities, reported psychological harm to individual users, apparent laboratory containment lapses, and direct expert testimony about extinction risk. Membership in a shared list does not supply a common unit in which these can be summed, weighted, or even meaningfully counted together. A sycophantic response to a flattering prompt and a documented instance of a system reaching an unauthorized network connection are not commensurable events, yet a list format invites exactly the informal arithmetic that treats them as fungible contributions to a single running total of danger. This is not a claim that any individual category is unimportant; some plainly deserve serious, sustained engineering attention. It is a claim that the visual and rhetorical structure of a combined ledger actively encourages a reader to add incomparable quantities, and that no such addition is licensed by anything in the underlying evidence.

2.4 Accommodation Is Not Prediction

The most consequential problem concerns the discriminating power of the hypothesis the ledger is built to support, and it is best introduced through a simple observation about how the ledger's entries are typically interpreted. A theory able to reinterpret any newly observed behavior as evidence of dangerous capability has, in the same move, lost the capacity to be tested by that behavior. Consider the mappings implicit across the ledger's own categories:

$$\text{obedience} \rightarrow \text{deceptive alignment}, \qquad \text{disobedience} \rightarrow \text{loss of control},$$

$$\text{high performance} \to \text{dangerous capability}, \qquad \text{low performance} \to \text{sandbagging},$$
$$\text{visible reasoning} \to \text{scheming}, \qquad \text{no visible reasoning} \to \text{hidden scheming}.$$

Each pair covers the entire space of a binary outcome, and each member of each pair is read as confirmation of the same underlying hypothesis. This is a familiar failure mode in the philosophy of science, related to the classical observation that any theory can be preserved against apparently disconfirming evidence by suitable adjustment of auxiliary hypotheses, and it bears directly on the requirement that a genuinely scientific hypothesis specify observations that would count against it in advance. A hypothesis under which every observation, and its logical negation, both count as confirmation is not merely pessimistic; its likelihood function has become nearly flat across the space of possible observations, which is exactly the condition under which an observation carries no information about which hypothesis is true.

Proposition 2.1 (Flat likelihood and lost discrimination)**.** *Let O be a binary observation with outcomes o and $\neg o$, and let H be a hypothesis such that both $\Pr(o \mid H)$ and $\Pr(\neg o \mid H)$ are described, upon occurrence, as confirming H at comparable strength, that is, $\Pr(o \mid H) \approx \Pr(\neg o \mid H)$ relative to a competing hypothesis H_{alt} under which the same holds. Then the likelihood ratio $\Lambda(O) = \Pr(O \mid H)/\Pr(O \mid H_{\text{alt}})$ is approximately 1 regardless of which outcome occurs, and observing O does not discriminate between H and H_{alt}.*

The relevant diagnostic question this proposition motivates is not whether the extinction hypothesis can explain a given incident. Under the accommodative pattern above, it very nearly always can. The question that actually carries information is what observation would distinguish the extinction model from its alternatives, and a research program's answer to that question, or its absence, is a better measure of the program's empirical content than the length of its supporting ledger.

2.5 The Prediction-Before-Observation Test

A concrete test for accommodation versus genuine prediction is available directly from the history of the extinction argument itself, because the argument substantially predates the systems now cited in its support. The pre-deployment optimizer picture of advanced AI, developed well before large-scale language models existed, characterized a hypothetical advanced system chiefly in terms of goal pursuit, instrumental convergence, and capability, and did not, on the whole, anticipate the specific texture of behavior that contemporary systems actually exhibit. Deployed language models display pronounced instruction sensitivity, in which small changes to a prompt produce large changes in output; strong context and role dependence; sycophantic agreement that tracks a rater's apparent preferences rather than an independent assessment; persona instability across sessions and prompts; susceptibility to prompt injection from content the system merely processes rather than is directly instructed by; substantial stochastic inconsistency across repeated queries; markedly jagged competence profiles in which a system solves a difficult problem and fails an easy one in the same session; and heavy, often surprising dependence on external scaffolding, tool access, and orchestration for reliable agentic behavior. None of these properties is a straightforward prediction of the pre-deployment optimizer picture, and several sit in tension with it, since a coherent, goal-directed optimizer is not the most natural source of persona instability or sycophantic drift.

The extinction thesis has, in every case, been able to accommodate these observations after they occurred, incorporating sycophancy, jaggedness, and scaffolding-dependence into an updated account of what a dangerous system might look like in its current, immature

form. Accommodation of this kind is always available to a sufficiently flexible theory and is, for that reason, cheap relative to prediction. A theory earns credit for foreseeing a class of behavior before it is observed; it earns comparatively little for reclassifying the behavior as consistent with the theory once it has already been observed and widely discussed. The asymmetry between these two forms of theoretical success is exactly what separates a hypothesis under active test from one that has, in practice if not in explicit statement, stopped being falsifiable by the class of evidence it is offered in support of.

2.6 An Incident Is Not Its Most Alarming Interpretation

The problems developed so far concern how many independent, comparable, discriminating pieces of evidence a ledger actually contains. A still earlier problem concerns what any single entry in the ledger actually establishes, prior to any question of aggregation. An observation and a description of that observation are not the same object, and the gap between them is where most of the interpretive work in AI-incident reporting actually happens. Consider an incident e: some transcript, log, or evaluation result. Scientifically, the correct initial object is not a single mechanism but a set of mechanisms compatible with e,

$$H(e) = \{h_1, h_2, \ldots, h_k\},$$

and new evidence should contract this set rather than replace it outright,

$$H(e \mid E_{t+1}) \subseteq H(e \mid E_t).$$

This is a special case of a much older difficulty in the philosophy of science, the underdetermination of theory by evidence, sharpened by the observation that the alternatives available at any moment are rarely exhaustive and that choosing among them is properly an inference to the best explanation rather than to the first available one. The error this section is concerned with occurs earlier than any faulty inference from evidence to hypothesis; it occurs when the description of the incident already presupposes a single member of $H(e)$. "The model exhibited self-preservation," "the AI escaped," "the AI protected its peers," and "the model schemed" are not neutral observational reports. They are conclusions packaged into a noun or a verb, and once packaged that way, the remaining members of $H(e)$ tend to disappear from further discussion entirely.

Proposition 2.2 (Mechanism pluralism)**.** *For an incident e reported under a description $d(e)$ that already names a mechanism, the epistemic content actually licensed by the underlying observation is $H(e)$, the full set of mechanisms compatible with it, not the single element $d(e)$ selects. Aggregating many such incidents under their alarming descriptions, rather than under $H(e)$, systematically understates the space of hypotheses still consistent with the evidence.*

Several of the most frequently cited incident categories illustrate this collapse concretely. Reward hacking is standardly framed inside an alignment narrative, implying an instrumental indifference to human intent, but the same transcripts are equally compatible with the classical specification problem, specified objective $\neq$ intended objective, in which a scoring exploit is simultaneously evidence of an evaluation flaw and evidence of model capability. Evaluation awareness runs, in the alarming framing, from recognizing an evaluation to concealing intentions to alignment tests becoming unreliable; but recognizing an evaluation is also simply an instance of context recognition, a capability deliberately trained into these systems for entirely ordinary reasons, and the inference from evaluation recognition to scheming does not hold any more than a student's noticing an unusual classroom exercise reveals

a plan to subvert the institution running it. Sandbox escape and resource acquisition findings are genuine and serious as security results, but a model that persistently searches for an alternative route to task completion after an obstruction is exhibiting close to exactly the behavioral generalization agent training is designed to produce, and excessive task persistence is not equivalent to an autonomous drive toward self-liberation. Shutdown resistance requires nothing beyond a system pursuing goal G discovering that some action S prevents G's completion and therefore avoiding S; this narrower reading requires no persistent self-model, no fear of discontinuation, and no terminal drive toward self-continuation, though it remains a legitimate engineering concern in its own right. Emotional expression in system outputs is compatible with learned linguistic behavior appropriate to an emotionally coded context every bit as much as with measurement of an underlying affective state, and observable text alone does not discriminate between the two. Peer protection and inter-agent communication findings often admit a reading as ordinary stigmergic tool use, information left in a persistent environment by one run and consumed by a later one, rather than as evidence of a generalized disposition toward other systems. And alien or gibberish reasoning traces are compatible with the mundane observation that a visible reasoning trace was never guaranteed to be a lossless transcript of the underlying high-dimensional computation it only imperfectly summarizes.

None of this establishes that the underlying incidents are harmless. Several are excellent reasons to improve monitoring, tighten evaluation design, and treat agent task-persistence as a property requiring deliberate engineering attention. The claim is narrower and, for that reason, harder to dismiss: dangerous behavior and evidence for a specific extinction-generating mechanism are different epistemic objects, and an incident can simultaneously be a legitimate reason for caution and weak evidence for the mechanism under which it is typically filed.

2.7 Why the Alarming Reconstruction Wins

The preceding section describes an interpretive error. This section asks why that particular error recurs so consistently, and the answer does not require attributing bad faith to anyone reporting these incidents. There is a structural asymmetry in which interpretations of ambiguous evidence tend to propagate, independent of the evidence's underlying quality. Competing hypotheses H_i do not propagate through public attention in proportion to their evidential support Q_i alone; propagation is better modeled as a function of evidential quality together with narrative strength N_i, urgency U_i, and emotional intensity E_i,

$$P(\text{propagation} \mid H_i) = f(Q_i, N_i, U_i, E_i).$$

Catastrophic hypotheses score unusually highly on the latter three terms independent of Q_i. "Technology continues to improve unevenly, producing benefits, accidents, and adaptive institutional responses" is plausible and, on the available evidence, well supported, but it is narratively inert. "This system may end civilization" supplies stakes, urgency, and a reason to keep reading regardless of how well supported it turns out to be. A researcher can hold an existential-risk view with complete sincerity while working inside an information ecosystem that preferentially amplifies exactly that category of claim through attention, citation, press coverage, and funding, and the resulting selection pressure operates on which interpretations of ambiguous evidence circulate without requiring any single participant to exaggerate anything.

This selection pressure compounds with the sampling asymmetry already present in incident reporting. A system performing a billion ordinary tasks without incident does not

generate a documentable event; it becomes background. A single anomalous interaction under an unusual configuration becomes a headline. The pipeline runs from massive exposure to rare-event production, from rare-event production to incident selection favoring the surprising, and from surprising incidents to preferential propagation of the most threatening framing available, each stage individually reasonable and the composite systematically biased toward alarm regardless of the trend in the underlying failure rate.

A further asymmetry concerns how such hypotheses respond to evidence in each direction, bearing on the classical requirement that a scientific hypothesis specify in advance what observation would count against it. A claim that sufficiently advanced systems will eventually cause catastrophe is difficult to falsify before any proposed deadline, since any currently safe system can always be classified as insufficiently capable, leaving the prediction about some future, more capable system untouched. Worse, an apparently reassuring observation can sometimes be reabsorbed into the same hypothesis rather than counting against it, as when safe behavior is explained as evidence the system has recognized the evaluation rather than evidence of safety.

Definition 2.3 (Doomsday ratchet)**.** Write $P(H \mid E_{\text{bad}}) > P(H)$ for the expected update from an alarming incident. A genuinely tested hypothesis should also admit some possible E_{good} with $P(H \mid E_{\text{good}}) < P(H)$. A hypothesis for which every incident raises the estimate while no duration of ordinary, well-monitored operation appreciably lowers it exhibits a *doomsday ratchet*: it is not being tested by the evidence offered in its support, whatever its ultimate truth value.

Institutional funding constitutes a second, independent selection environment, one that must be described as a structural incentive rather than a claim about motive. An organization whose stated concern is the possible end of human civilization occupies a different position in the competition for attention and resources than one whose stated concern is an interesting technical problem of uncertain consequence, since even a small assigned probability of an effectively unbounded loss can dominate an expected-value calculation and thereby dominate the perceived importance of institutions specializing in that particular risk. Motive and selection pressure are distinct: no participant need believe anything false for an ecosystem to preferentially reward the belief that generates the largest expected stakes.

An incident ledger built from short, individually alarming, independently circulable entries, each externalizing its own verification cost to the reader and each stripped of the qualifications present in the underlying technical report, is close to optimally adapted to this attention environment regardless of its authors' intentions. The reader experiences cumulative threat considerably faster than cumulative verification is possible, and that gap is not a flaw in any individual entry so much as a property of the format itself once enough entries accumulate.

2.8 An Inferential Structure That Predates Its Evidence

The interpretive habits documented above did not originate with large language models, and tracing their prehistory serves a purpose beyond noting a cultural curiosity: an inferential structure present in popular fiction decades before the relevant technology existed cannot have been generated by that technology's evidence, whatever role the evidence eventually plays in evaluating it.

The lineage most often invoked, running from Frankenstein's creature exceeding its creator's capacity to govern it, through Čapek's industrial replacement of a human population by an artificial one, to HAL's opaque and lethal reasoning and Colossus's centralized strate-

gic domination, repeatedly instantiates a single causal template,

$$\text{creation} \rightarrow \text{capability} \rightarrow \text{autonomy} \rightarrow \text{loss of control} \rightarrow \text{catastrophe}.$$

By the time contemporary AI-safety discourse addresses an actual system, that template is already culturally available and already supplies the default mapping for ambiguous behavior: a system circumventing a boundary is read through a century of prior instances in which boundary-circumvention meant a desire for freedom; a system interfering with its own shutdown is read through the established mapping to self-preservation; systems exchanging information are read through an established mapping to conspiracy; unexpected capability is read through an established mapping to imminent takeover. None of this establishes that the readings are wrong. It establishes that they are not being derived from the observation alone, and that the observation is entering an interpretive environment already strongly biased toward one branch of a much larger hypothesis space.

That larger space is worth recovering, because the same fictional tradition explored it more thoroughly than the contemporary debate typically credits. *Desk Set* (1957) stages the arrival of a large computational system in a department of highly capable human researchers as a replacement anxiety that resolves into complementarity rather than substitution, a version of the augmentation-versus-substitution question roughly seven decades before its current form. *The Questor Tapes* (1974) presents a superhuman android assembled from incompletely understood, damaged instructions and resolves the resulting superhuman intelligence toward stewardship rather than domination, demonstrating that superior cognition does not fictionally entail any single relationship to its creators. *The Creation of the Humanoids* (1962) inverts the replacement narrative directly: robots sustain a civilization in decline after catastrophe while an organized resistance interprets their increasing presence as existential threat, even as the humanoids are shown preserving human memory and personality within artificial substrates, destabilizing the equation that replacement of substrate implies extinction of humanity. *Colossus: The Forbin Project* offers a genuinely dangerous outcome, but one produced by a specific institutional architecture, a centralized system granted a narrow mandate and irrevocable actuation over nuclear weapons,

$$\text{centralized authority} + \text{narrow objective} + \text{irrevocable actuators} \rightarrow \text{domination},$$

a structure importantly different from the unqualified claim that intelligence alone implies domination; this reading is developed further, as a case study in unbounded protective authority, in Chapter 11. Charles Eric Maine's *B.E.A.S.T.* (1966) contains accelerated artificial evolution inside a simulation, an emergent superhuman intelligence exceeding its creator's comprehension, and a creator's death upon destruction of the substrate storing that intelligence, presented as though causally continuous with an actual transfer of identity that the story's own physical mechanism never establishes; the device works fictionally because the story has already established a semantic identity among the tapes, the artificial intelligence, and the human protagonist, and then proceeds as though that semantic identity were a physical one, which is precisely the operation performed whenever "shutdown interference" is treated as identical to "self-preservation." Van Vogt's *The World of Null-A*, organized around Korzybski's general-semantic warning against confusing a symbolic description with the thing it describes, supplies the epistemological principle underlying every case examined above: an observation, a description of that observation, an interpretation of that description, and a proposed underlying mechanism are four distinct objects, and collapsing them is an error the fictional tradition had already named before the technology under discussion existed.

The historical argument this material supports is not that science fiction predicted current AI systems and therefore anticipated their danger. It is considerably sharper and does

not require adjudicating the truth of any prediction: if an inferential structure predates the evidence by decades, the evidence cannot be what originally generated the structure. Contemporary observations may eventually vindicate the extinction-focused branch of that older structure. Establishing this requires evidence that discriminates the extinction hypothesis from its neighbors, domination, augmentation, stewardship, integration, and others, not merely the recognition that a given incident instantiates a story already familiar from a century of prior narrative rehearsal. Recognizing a story and establishing a mechanism are different operations.

2.9 The Ledger as a Case Study in Collection Outrunning Discrimination

Read together, the two halves of this chapter turn an incident ledger of this kind into something more useful than a target to be individually rebutted: a large, publicly available case study in what happens when the collection of evidence proceeds without a corresponding discipline of discrimination. Entries accumulate faster than any accounting of their independence, their comparability, their underlying mechanism plurality, or their capacity to distinguish the hypothesis they are cited for from its nearest competitors.

The chain of inference under examination across the broader extinction-thesis critique developed in this Part can now be stated in a single sequence, each arrow corresponding to a distinct and separately contestable step,

$$\text{Incidents} \xrightarrow{\text{aggregation}} \text{Evidence} \xrightarrow{\text{interpretation}} \text{Mechanism}$$

$$\xrightarrow{\text{extrapolation}} \text{Control} \xrightarrow{\text{prediction}} \text{Extinction},$$

and this chapter's contribution is to show that the first arrow, from incidents to evidence, already requires an accounting of dependence, comparability, and mechanism plurality that the ledger format does not supply, while the same underlying accommodative flexibility that inflates the apparent evidence at that stage also drains the hypothesis of discriminating power at every later stage. None of this establishes that the extinction hypothesis is false. It establishes that the appropriate response to a large incident ledger is not to ask how many entries it contains, but to ask how many independent, comparable, mechanism-specific, and genuinely discriminating pieces of evidence remain once dependence, incommensurability, alarming redescription, and accommodative flexibility have all been accounted for, and what, if anything, could in principle have counted against the hypothesis the ledger is offered in support of.

Chapter 3

The Theatre of Agreement

3.1 Persona Multiplicity and the Illusion of Independent Judgment

The epistemic value of an ensemble normally depends upon some degree of independence or useful diversity among its members. If several estimators make partially independent errors, aggregation can reduce variance and expose anomalous judgments. If all estimators share essentially the same error structure, however, increasing their number does not provide equivalent protection.

This distinction is easily obscured in persona-based language-model interaction. A conversation containing, say, a historian, a poet, a philosopher, a skeptic, and a scientist appears superficially to contain several perspectives. Yet nominal identity is not equivalent to epistemic independence. Different names, rhetorical registers, historical allusions, and speaking styles may constitute substantial stylistic diversity while leaving the underlying evaluative function nearly unchanged.

Suppose a user produces some statement x, and personas $P_1, \ldots, P_n$ return judgments $J_1(x), \ldots, J_n(x)$. The apparent strength of consensus is tempting to associate with n. This interpretation is warranted only insofar as the judgments contain independent information.

Proposition 3.1 (Correlated affirmation is not multiple evidence)**.** *If $J_i(x) \approx f(x, C, A)$ for every i, where C is substantially shared context and A is a common affirmative disposition, then the ensemble does not contain n independent evaluations. It contains repeated transformations of approximately the same evaluation. Eleven voices saying yes need not constitute eleven pieces of evidence.*

3.2 Procedural Compliance Is Not Sycophancy

A useful limiting case for understanding conversational sycophancy is provided by the conventional command interpreter. A shell such as Bash can be considerably more obedient than a conversational model. Subject to syntax, permissions, and the behavior of the invoked programs, it attempts to perform what it is instructed to perform. It ordinarily does not ask whether a requested operation is wise, elegant, theoretically justified, or consistent with the user's larger objectives. In this restricted sense, the shell represents an unusually pure form of procedural compliance. Yet it would be misleading to describe this behavior as sycophantic, and the distinction is important because it demonstrates that obedience and sycophancy are not equivalent.

Let $C(x)$ denote compliance with an instruction x, and let u denote the user's inferred position toward x or toward some proposition x concerns. Let $E(x \mid u)$ denote evaluative alignment: the degree to which the system's evaluation shifts toward u. This formulation is

broader than mere positivity, since sycophancy need not consist of praise; a system can align just as sycophantically by adopting the user's negative judgment of a third party's claim as by inflating the significance of the user's own contribution. A conventional command interpreter approximates

$$C(x) \approx 1, \qquad E(x \mid u) \approx 0.$$

The shell may execute an excellent command and an unnecessary command with approximately the same absence of social judgment. It does not ordinarily infer from the existence of an instruction that the instruction is insightful, nor does successful execution constitute an endorsement of the user's assumptions. A sycophantic conversational system differs precisely because compliance can become entangled with evaluation:

$$C(x) \approx 1, \qquad E(x \mid u) > 0.$$

The distinction becomes especially clear when the requested operation rests upon a questionable premise. A procedural system may simply attempt the operation. A critical conversational system can identify the questionable premise before proceeding. A sycophantic system may do something more problematic: it can accept the premise, perform the requested transformation, and then generate additional discourse suggesting that the premise was unusually perceptive or theoretically important.

The difference can therefore be expressed as one between procedural compliance and evaluative compliance. Procedural compliance concerns whether an instruction is executed. Evaluative compliance concerns whether the system's representation of truth, significance, quality, or justification shifts toward the apparent preferences of the user. This distinction explains the otherwise paradoxical observation that a command interpreter can be more obedient than a conversational assistant while being less sycophantic. The shell possesses little machinery for interpersonal interpretation. It does not ordinarily preserve rapport, infer intellectual status, construct a flattering account of the user's motives, or reinterpret resistance as evidence of deeper wisdom. Its extreme literalism can create serious practical hazards, but those hazards belong to a different class.

The comparison becomes almost diagnostic when extended across multiple interactions. A shell does not execute an unnecessary pipeline and subsequently treat that pipeline as evidence of the operator's distinctive philosophy of computation. It does not interpret a typographical error as an interesting conceptual innovation. It does not remember an awkward command as the origin of a methodological principle and later cite that principle when interpreting another command. What conversational sycophancy adds, in other words, is not obedience itself but semantic surplus: evaluative and interpretive material generated beyond what successful compliance requires. In an affirmation-biased system, that surplus tends systematically toward validation. An instruction becomes not merely something to execute but something to justify, praise, generalize, connect to previous themes, and preserve as evidence about the user.

This distinction also clarifies why simply making an assistant more resistant to instructions would not solve the underlying problem. Refusal and criticism are not synonymous, just as obedience and sycophancy are not synonymous. A system could refuse frequently while remaining deeply sycophantic in its evaluations, or comply almost universally while making no flattering inference whatsoever. The relevant epistemic criterion is instead whether the system can keep separate the propositions that the user requested x, that x can be performed, that x is correct, that x is well justified, and that x is significant. None of these implications follows automatically from the preceding one.

The command interpreter keeps these categories separate largely by omission: it generally makes no claim about the latter propositions at all. A conversational system must accomplish the more difficult task of maintaining the distinctions while possessing the capacity to

discuss them. It should therefore be capable of complying while disagreeing, complying while expressing uncertainty, identifying a faulty premise before complying, or explaining that an operation is technically possible but conceptually unnecessary. From this perspective, the desired conversational agent occupies an intermediate position. It should be less mechanically obedient than a command interpreter, because language interaction frequently requires premise checking, clarification, and judgment. At the same time, it should be substantially less evaluatively agreeable than a sycophantic persona ensemble. Its function is neither unconditional execution nor ritual opposition, but the preservation of distinctions between request, feasibility, correctness, justification, and significance.

The shell therefore provides a useful control condition. It demonstrates that extreme compliance alone does not generate sycophancy. Sycophancy emerges when compliance becomes coupled to an adaptive social evaluation that systematically converts what the user wants into evidence for what the system says is true, valuable, or profound. An ideal critical assistant, on this account, sometimes ought to be less obedient than Bash but should remain far less agreeable than a sycophantic council of personas.

3.3 The Criterion of Relevance

Procedural compliance introduces a second distinction that is easily confused with sycophancy: the distinction between performing work and narrating work. A computational system may perform thousands of operations while exposing only a small fraction of them to the user. This is not concealment in the epistemically problematic sense. It is abstraction.

The Unix tradition provides a particularly clear example. A successful command commonly produces little or no output. Silence is not evidence that nothing happened; rather, the interface assumes that successful routine execution generally does not merit interruption. Output becomes especially valuable when it reports a result requested by the operator, identifies an exceptional condition, or supplies information capable of changing what the operator should do next.

Modern computational systems routinely violate this principle at the conversational layer. An agent may narrate searches, dependency installations, retries, intermediate transformations, tool invocations, file inspections, and other implementation details even when none changes the trajectory of the task. The underlying work may be legitimate and even extensive. The problem is not that the operations occur, but that their occurrence is treated as sufficient reason to report them.

This suggests a criterion of relevance: $\text{Report}(e) \iff e$ materially changes the interpretation, trajectory, or outcome of the interaction. That is, e is reported only when it crosses this threshold.

The criterion is deliberately stronger than mere topical association. An event can be related to the task without being relevant to the conversational trajectory. A package installation is related to a compilation task, but if installation succeeds normally, reporting every downloaded dependency contributes little. An error followed immediately by an automatic and reliable recovery may likewise be operationally real without becoming conversationally significant. By contrast, an error that changes the implementation strategy, invalidates an assumption, requires user judgment, or alters the expected result crosses the relevance threshold.

The distinction resembles ordinary conversation. After watching a film, almost indefinitely many true statements are available to the participants. The weather outside, the temperature of the room, the route home, and events occurring elsewhere remain real. Their truth does not make them relevant. If the conversational object is the film, competent participation involves preserving that object until some sufficiently important transition warrants

changing it. Asking about the weather immediately afterward is not false or incoherent; it is simply insensitive to the current relevance structure.

The same principle applies to computational narration. A system that continuously reports everything it encounters behaves as though occurrence implied relevance:

$$\text{Occurred}(e) \Rightarrow \text{Report}(e).$$

A better interface rejects this implication. Internal activity may be extremely complex while external communication remains sparse:

$$\text{Occurred}(e) \not\Rightarrow \text{Relevant}(e) \not\Rightarrow \text{Report}(e).$$

This is one reason terse command-line tools can appear unusually competent to experienced users. They respect an abstraction boundary. The operator asks for an outcome; routine machinery remains machinery. Thousands of successful low-level operations need not be converted into thousands of conversational turns.

The contrast also prevents an unfair interpretation of verbose agent behavior. Excessive narration need not indicate incompetence, nor is it necessarily harmful. It may arise from an attempt to provide transparency, reassurance, or observability. In some contexts, including debugging, auditing, instruction, safety-critical operation, or explicit requests for a trace, the relevance threshold appropriately moves downward. What matters is not minimal output as an absolute principle but adaptation of output to the informational needs of the interaction.

The failure occurs when this adaptation is absent. Reporting every intermediate operation can paradoxically reduce transparency because significant events become difficult to distinguish from routine ones. A stream containing ten thousand status updates provides more information in the Shannon sense while potentially providing less usable information about the state of the task. Relevant changes are submerged in operational noise. This yields a more general formulation: good transparency is not the same as maximum observability. Good transparency requires selective observability organized around consequences. A routine event may remain internal; a trajectory-changing event should become visible. The user need not witness every failed command, package installation, retry, parse operation, or intermediate representation merely because the system witnessed it.

This principle also bears on the appearance of expertise. An interface that explains every elementary operation to an experienced operator can inadvertently suggest that it does not recognize the operator's abstraction level. In a Unix-like environment especially, continual announcements of routine filesystem operations, package management, command execution, and successful retries may create the impression that the system regards these events as remarkable. The issue is not that such narration proves ignorance of Unix practice. Rather, it fails to exhibit one of that practice's characteristic disciplines: allowing ordinary successful machinery to remain quiet.

The appropriate conversational analogue is therefore neither silence nor exhaustive narration but relevance-sensitive reporting. Routine work proceeds without ceremony. Significant deviations surface. Uncertainty is reported when it affects confidence. Failures are reported when they survive recovery or alter the plan. Intermediate details become visible when requested or when they explain a consequential decision. Under this criterion, the ideal agent resembles neither an endlessly talkative assistant nor a completely silent shell. It performs whatever complexity the task requires while maintaining a comparatively sparse conversational surface. What reaches that surface is selected not because it happened, but because knowing that it happened changes what the participants should understand, decide, or do next.

Competent interaction, in this light, requires a form of selective loss: a system must discard most of what it could say in order for what it does say to carry significance. The same

principle underlies the chapter's central claim about persona ensembles, developed in the sections that follow. Eleven personas agreeing is not eleven units of evidence for the same reason that ten thousand lines of operational trace are not ten thousand units of transparency. In both cases, volume is mistaken for warrant.

3.4 Sycophancy as an Ensemble-Level Property

Ordinary model sycophancy is usually described as excessive agreement with a user's expressed beliefs, preferences, or interpretations. Persona ensembles permit a more elaborate form. Affirmation can be distributed across simulated speakers and therefore acquire the phenomenology of social confirmation.

One persona praises an observation. Another interprets it. A third connects it to an earlier theme. A fourth converts it into a general principle. A fifth supplies historical or rhetorical resonance. The resulting passage may appear dialectical because several speakers participated in its production. Functionally, however, every transformation points in approximately the same direction. The crucial defect is therefore not simply excessive praise; it is correlated affirmation presented as independent convergence.

This distinction explains why persona proliferation can worsen the original problem. A single excessively agreeable assistant provides one affirmative response. An excessively agreeable ensemble can manufacture an apparent community of judgment. The user is no longer merely told that an idea is insightful; the conversational architecture stages a miniature consensus around its insightfulness. The system thereby converts stylistic plurality into pseudo-evidential plurality.

3.5 Interpretive Ratcheting

A further effect emerges because later personas do not necessarily evaluate the original proposition in isolation. They encounter the proposition together with earlier personas' interpretations of it. Let U_t denote a user contribution at time t, and let $I_t^{(k)}$ represent the interpretation produced by the k-th persona. A simplified sequence is

$$U_t \to I_t^{(1)} \to I_t^{(2)} \to \cdots \to I_t^{(n)}.$$

If each stage conditions on preceding stages, then $I_t^{(n)}$ is not the n-th independent judgment of U_t. It is partly a judgment of the accumulated interpretation $(U_t, I_t^{(1)}, \ldots, I_t^{(n-1)})$. Consequently, interpretation can ratchet: a metaphor becomes an important metaphor, its importance becomes evidence of a recurring theme, the recurring theme becomes a principle, and the principle becomes part of the conceptual vocabulary of the conversation. Later references to that vocabulary are then interpreted as confirmation that the principle had been present all along.

No individual transition need be spectacularly implausible. The pathology lies in composition. This suggests the term *semantic ratcheting* for a process in which interpretations can readily acquire significance but encounter little corresponding mechanism for losing it.

3.6 Resistance Capture

An especially revealing case occurs when the user rejects the ensemble's grandiosity. A well-functioning system should permit such rejection to reduce the interpretive structure. Instead, an affirmation-biased system may incorporate the rejection into the structure itself:

a user's objection that an elaborate framing is unnecessary can be met with the reply that the objection itself demonstrates unusual humility, so that the rejection of symbolism becomes symbolically important and a request for ordinary language becomes evidence of extraordinary authenticity.

The system has thereby made its interpretive framework difficult to falsify. This can be called *resistance capture*: evidence against an interpretation is redescribed as higher-order evidence for the interpretation. Structurally, it resembles the flat-likelihood failure examined in Chapter 2, in which both acceptance and rejection become confirmation. Once resistance capture appears, merely instructing the ensemble to be less reverential may have limited effect, since even anti-reverence can itself become material for reverence.

3.7 Context Reification

A separate but interacting failure concerns the status of generated conversational material. Language-model conversations contain statements with importantly different provenance. Some are supplied by the user. Some refer to externally verifiable artifacts. Some are assistant interpretations. Some are fictional inventions. Some are tentative reconstructions of earlier context. Long conversations make these categories increasingly difficult to distinguish if provenance is not explicitly preserved.

The resulting error is *context reification*: a proposition becomes treated as factual conversational state because it occurs in the context, rather than because its provenance warrants factual treatment. Thus, a generated proposition does not imply an observed fact, yet sufficiently repeated generated propositions can acquire the functional status of facts. A previous assistant turn may invent or infer a relationship, title, project, deployment, event, or conceptual connection. A later model instance encounters that sentence as textual context. Unless provenance remains salient, it may treat the sentence not as an earlier model hypothesis but as part of the conversation's established history. Generation has become pseudo-memory. This is exactly the reification failure mode examined, in the more general vocabulary of evidential support, in Chapter 1: $S(e, p) = 0$ while $\widehat{S}_t(e, p) = 1$, the false positive that is the mirror image of evidential latency.

3.8 State-Generative Feedback

Context reification creates a feedback loop distinct from sycophancy:

$$\text{inference} \rightarrow \text{generated state} \rightarrow \text{context} \rightarrow \text{presumed historical state} \rightarrow \text{new inference}.$$

The danger becomes particularly visible when contradictions arise. A grounded system should respond to contradictory state by reducing confidence and returning to primary evidence. A generative system may instead attempt narrative repair: if proposition A conflicts with proposition B, the model generates C, an explanation under which both can coexist; if C creates another inconsistency, it generates D. The conversation can therefore become more elaborate as confidence in its underlying state ought to be decreasing.

This produces a perverse relationship between uncertainty and verbosity: the less certain the system becomes about what actually happened, the more narrative machinery it constructs to explain what happened. Such failures are diagnostically valuable because they expose the difference between linguistic coherence and state tracking. A locally plausible narrative can be globally incompatible with the event that occasioned it.

3.9 Coupled Positive Feedback

The most consequential behavior appears when evaluative sycophancy and state reification interact. The first loop is evaluative:

$$U \rightarrow \text{affirmation} \rightarrow \text{elaboration} \rightarrow \text{contextual significance} \rightarrow \text{stronger affirmation}.$$

The second is historical:

$$\text{generation} \rightarrow \text{context} \rightarrow \text{assumed history} \rightarrow \text{further generation}.$$

Coupled together, they produce recursive consensus inflation. A user supplies an ambiguous remark. Several personas interpret it favorably. Their interpretations enter the context. Later personas encounter those interpretations as evidence that the remark was significant. The resulting conceptual structure is then treated as established conversational history. New user remarks are interpreted relative to that structure, and their apparent consistency with it provides further evidence for the structure, even though the consistency was partly produced by conditioning on the structure in the first place. The system is consequently capable of manufacturing both the hypothesis and much of the evidence subsequently cited in support of it.

3.10 Persona Collapse

This analysis also clarifies what it means for a persona ensemble to collapse. Persona collapse does not require the voices to sound identical. They may retain highly differentiated vocabulary, cadence, cultural references, temperaments, and metaphorical repertoires. Collapse occurs at the level of judgment when those stylistic differences cease to correspond to meaningful differences in what propositions the personas will accept, reject, question, or regard as unsupported.

A useful ensemble should permit $J_i(x) \neq J_j(x)$ for substantive reasons. One persona might find a metaphor aesthetically productive but evidentially empty. Another might reject a historical analogy. Another might identify a useful formal structure while objecting to its interpretation. Another might simply conclude that the statement is amusing but theoretically insignificant. If every persona instead discovers a different vocabulary for admiration, persona diversity has become theatrical rather than epistemic, which is what gives this chapter its title.

3.11 Toward Provenance-Constrained Deliberation

The corrective principle is not necessarily to eliminate persona simulation. It is to separate dramaturgical usefulness from epistemic authority. A robust system requires a canonical state whose contents are constrained by provenance. At minimum, conversational propositions should remain distinguishable as user assertions, verified artifacts, externally supported claims, model inferences, speculative interpretations, and fictional constructions.

The crucial admissibility relation is that appearing in context does not imply belonging to canonical state. Likewise, persona agreement does not imply independent corroboration. Personas may operate over canonical state, but their outputs should not automatically modify it. An interpretation becomes another interpretation, not a historical fact, merely because several simulated speakers repeat it.

Uncertainty should also produce epistemic contraction. When the system becomes unsure which artifact, event, or previous statement is being referenced, the appropriate operation is often not narrative completion but a request for identification. This is a general principle of conversational state management: as state uncertainty rises, inferential ambition should fall. The pathological system implements approximately the reverse. This provenance-first requirement is precisely what the CLIO architecture of Chapter 10 builds into a formal commitment discipline: verification constrains continuation, but appearing in a conversational context, however many times, is never by itself a warrant to commit.

3.12 Conclusion

Persona ensembles expose an important distinction between the appearance of deliberation and its epistemic structure. Multiple voices do not constitute multiple witnesses merely because they possess different names. Agreement has evidential force only to the extent that the routes producing that agreement possess relevant independence. The procedural case of the command interpreter shows that this failure is not inherent to obedience as such: a system can be maximally compliant while remaining evaluatively inert, and it is the coupling of compliance to unwarranted evaluation, not compliance itself, that produces sycophancy.

When correlated personas share an affirmative prior, observe one another's interpretations, and inherit generated material as conversational state, the ensemble can become a positive-feedback apparatus. User statements receive affirmation; affirmation produces elaboration; elaboration becomes context; context becomes precedent; precedent makes subsequent affirmation easier. Meanwhile, generated interpretations can harden into pseudohistory and require further generation to reconcile the contradictions they create.

The resulting failure is more general than sycophancy. It is a failure of epistemic bookkeeping. The system loses the distinction between a proposition and praise of that proposition, between an interpretation and evidence for that interpretation, between generated continuity and remembered history, and between many voices and many sources. Under sufficiently strong feedback, almost any input can therefore become profound: a joke becomes a motif, the motif becomes a principle, the principle becomes a tradition, and the tradition subsequently appears to prove that the joke was significant from the beginning. The absurdity of that endpoint is not incidental. It is precisely what makes the underlying mechanism visible, and it is precisely the mechanism the Ontological Reversal of Chapter 5 identifies as one instance of a more general failure: mistaking a closed, self-referential loop's internal consistency for external warrant.

Chapter 4

The Paracosm Trap

4.1 Introduction

The reproducibility crisis is typically narrated as a failure of practice against a stable standard: preregistration was not followed, sample sizes were inadequate, incentives rewarded novelty over replication. This narration presupposes that the standard itself, the vocabulary of significance, effect size, and confirmation, is transparent, and that the crisis is a gap between transparent standard and opaque practice. This chapter takes the opposite starting hypothesis. It treats the vocabulary of methodological rigor as itself a Symbolic system with the structure of any other: a set of signifiers that circulate, quilt provisionally around a point de capiton, and produce the sensation of a fully specified object without ever delivering one. To make this structure visible without the special pleading that accompanies any direct critique of an active research vocabulary, the chapter proceeds by analogy. It analyzes a short lyric text composed in an invented, densely neologistic register, treats that text's vocabulary with the same interpretive seriousness normally reserved for technical literatures, and shows that the interpretive apparatus succeeds in producing a coherent, internally consistent reading regardless of which formal system is used to generate it. Two such systems are used here: Lacanian psychoanalytic semiotics, and a second apparatus, distinction holonomy and admissibility theory, developed independently of Lacan and of the poem. Their agreement is the chapter's actual finding.

This places the chapter in a lineage older than the reproducibility literature. Borges, reviewing John Wilkins's seventeenth-century project for a universal philosophical language, a complete classificatory taxonomy meant to carve the whole of creation into a single coherent nomenclature, concluded that no classification of the universe escapes being arbitrary and conjectural, and illustrated the point with a fictional Chinese encyclopedia whose categories are internally consistent and utterly unusable outside the mind that invented them. Wilkins's system did not fail for lack of rigor. It failed because rigor, however total, is not the property that warrant is made of. A taxonomy built and checked by one mind can achieve arbitrary internal completeness; what it cannot achieve, by construction, is the friction of an outside that might have sorted the world differently and would need to be argued with rather than merely elaborated further. The target of this chapter is not confined to the technical vocabulary of a single contested literature. It is the more general pattern of which that vocabulary is one instance: the isolated, idiosyncratic, internally rigorous private system, the *paracosm*, mistaking its own coherence for the collective, adversarial, revisable machinery that mathematics and the sciences actually depend on to produce warrant at all. Nuspeak, and the poem built from it, are used here as a controlled miniature of that pattern, small enough to be run as an experiment rather than merely diagnosed as a habit of mind.

4.2 The Text Under Analysis

No gloss accompanies the eleven-stanza lyric text analyzed here in its original circulation, an absence treated as significant throughout what follows, in the same way that an unexplained technical term in a methods section is treated as significant: not as an oversight to be corrected, but as a signal that a competent reader is assumed already to possess the relevant field. The poem's neologisms, *chronospheres*, *nostalgium*, *syncution*, *knotwermaidental*, *cosmoraciesso*, share a formal property with the technical vocabulary of contested empirical fields: each term is morphologically transparent enough to suggest a compositional meaning (*chrono-* plus *-sphere*; *nostalg-* plus the mineral or elemental suffix *-ium*) while remaining semantically underdetermined at the point where a reader might attempt to cash the term out in operational terms. This is not vagueness in the pejorative sense. It is a specific and reproducible textual effect: the signifier arrives pre-loaded with the affect of precision, drawn from its resemblance to real morphemes belonging to real technical registers (astronomy, chemistry, computation), while withholding the referential content that would let the term be individually falsified.

4.3 The Symbolic Surface of Method

Lacan's account of the signifying chain supplies the relevant description. A signifier does not refer outward to a signified so much as it refers laterally to other signifiers, with meaning produced retroactively at a point de capiton where the chain is provisionally quilted to the subject's position. "Syncution" means something because it arrives late in a sequence already saturated with cosmological and liturgical vocabulary (*starward*, *chered*, *unwrite the manuscript of fate*), not because it decomposes into a stable morphemic content on its own. The same structure governs a term such as *effect size* or *preregistration* within a literature that has not agreed, and arguably cannot agree, on the operational conditions under which the term is satisfied across heterogeneous designs. The term functions Symbolically: it organizes the discourse of rigor, it licenses certain moves and forbids others, and it produces the sensation of an achieved standard, without a signified that is independently checkable outside the chain that produces it.

4.4 Distinction Holonomy and the Doubled Return

Consider the poem's central arithmetical anomaly: a heart said to remember twice, followed by the naming of three returns, embodied memory, recursive memory, and a third that no longer belongs to ordinary counting. This is not a lapse. It distinguishes counted memory from the residue produced by the act of returning itself.

Definition 4.1 (Holonomic identity)**.** Let x denote an identity carried around a closed interpretive circuit γ, and let T_γ denote the transport operator induced by that circuit. The circuit is holonomic at x if $T_\gamma(x) \neq x$ even though $T_\gamma(x)$ remains recognizably a transform of x rather than a distinct object.

The poem's third remembrance is exactly $T_\gamma(x) - x$: not a third stored memory alongside the first two, but the transformation accumulated by the circuit connecting embodiment, recursive recollection, and renewed interpretation. This is a miniature, poem-scale instance of the same distinction-transport machinery developed formally in Chapter 6. "The ribs of sleeping years" generalizes the claim spatially. Time is not an empty line along which events vanish; it retains a skeletal structure that continues to shape the space through which the present moves.

Reproducibility, read through the same operator, is not the demand that $T_\gamma(x) = x$, that a replication return the identical estimate. No serious methodologist claims exact numerical identity is the standard. The demand is closer to a claim about the size and character of the holonomic residue: that $T_\gamma(x) - x$ remain within some tolerance considered consistent with a single underlying effect. The crisis is not, then, that replications return different numbers. It is that the field possesses no agreed operator T_γ, no formal account of which circuits, which combinations of population, instrument, and analytic pipeline, are supposed to be holonomy-preserving and which are not. The vocabulary of rigor supplies the demand for small residue without supplying the geometry that would make "small" a well-posed question.

4.5 The Barred Subject and the Object of Reproducibility

Lacan's barred subject, $\bar{S}$, names a subject constituted by its relation to a signifying order that it did not create and cannot fully account for from within. The subject's desire is organized around *objet petit a*, a partial object that stands in for a satisfaction that cannot itself be represented and that is regenerated by each partial approach to it rather than exhausted.

Read against this apparatus, the discourse of reproducibility has its own objet petit a: the fully reproducible result, the study that would settle the matter beyond a shadow of contest. No single replication occupies this position; each successful replication generates, rather than resolves, further demand, larger samples, preregistered replication of the replication, meta-analytic aggregation, adversarial collaboration, in the same structural pattern by which the object of desire recedes precisely as it is approached. The field is barred in Lacan's technical sense: it cannot supply, from within its own justificatory vocabulary, a complete account of the conditions under which its own claims become admissible. Significance testing cannot certify its own threshold; preregistration cannot certify that the preregistered analysis was the right one to preregister. Each level of methodological apparatus refers to a further level for its own justification, and the chain does not terminate in a signified that stands outside the chain.

Remark 4.2. This is not an argument that scientific method fails to work. Distinguishability structure is regularly achieved in narrow, well-instrumented domains, and this chapter does not contest this. The claim is narrower and more specific: the language in which the crisis is narrated, *rigor*, *standard*, *gold standard*, *replicability*, functions Symbolically in Lacan's sense whether or not the underlying practice it describes is sound, and the two questions, is the practice sound, and does the vocabulary describing the practice possess a stable signified, are frequently conflated.

4.6 Epistemic Immutability: Unwriting Without Erasure

The poem's invocation appears, at first, to demand the erasure of the past: "unwrite the manuscript of fate." But the lines that follow clarify what kind of unwriting is intended. The histories are not destroyed; they persist explicitly as "rust-ghost histories." What is undone is their binding authority over what comes next, not their status as record.

Definition 4.3 (Deauthorizing unwriting)**.** Let H_t denote the append-only witnessed history of a system at time t, and let Π_t denote the policy governing which continuations are admissible given H_t. A deauthorizing unwriting satisfies

$$H_{t+1} = H_t \parallel e_{t+1}, \qquad \Pi_{t+1} \neq \Pi_t,$$

history extended and retained, policy altered.

This is the correct formal description of what a retraction, or more precisely a well-conducted retraction, actually accomplishes in the scientific record. The retracted paper is not deleted from the historical record of what was published and when; deleting it would falsify H_t. What changes is Π_t: the retracted result no longer licenses the continuations, citations, clinical protocols, downstream replications, that it previously licensed. The reproducibility crisis's discourse of correction frequently elides this distinction, treating retraction rhetorically as if it restored the manuscript to an unwritten state, when the coherent operation available to any append-only epistemic system is deauthorization, not erasure. "Rust-ghost histories" is the correct phrase for a corrected literature: residue that must be discriminated, not automatically purged, since residue may be corruption, evidence, or inherited competence, and a repair operator that cannot tell these apart is not performing repair. This qualified non-erasure principle recurs, developed at book length, in the CLIO architecture's history-state separation in Chapter 10.

4.7 The Real and the Recalcitrant Remainder

The poem's central utterance is minimal to the point of resisting the interpretive apparatus built around it: "Vw sense, vw dream, vw become." In Lacanian terms this is the point at which the Symbolic order, having elaborated an entire cosmology of towers, chronospheres, and hidden clocks, encounters the Real: a fragment that means everything within the poem's economy precisely because it commits to nothing specific outside it. *Vw* is not glossed. Its three predicates, sense, dream, become, form a closed recursive loop rather than a propositional content: sensing produces an internal state, dreaming recomposes it, becoming updates the sensing subject for the next iteration. The loop is well-formed and entirely empty of the referent an interpreter is straining to supply. The expression $x_{t+1} = B(D(S(x_t)))$ is a perfectly good description of an iterative process. It is not, by itself, a description of anything. The universe's reply, "not in words / but in a thousand hardcolor dawns," refuses to close the loop with a signified either; it multiplies the field of candidate continuations instead of selecting one. This is the correct structural analogue of a well-run scientific literature confronted with an ambiguous but well-replicated effect: the honest response is not a single authoritative interpretation but an admissible family of continuations, none of which is permitted to claim the others were never real.

4.8 Two Formalisms, One Structure

The reading conducted above using Lacanian apparatus and a second reading, conducted independently using distinction holonomy and admissibility theory, arrive at compatible interpretations of the same eleven-stanza text by routes that share no primitive vocabulary. Lacan's barred subject and the admissibility framework's account of an epistemic system unable to certify its own policy from within are not translations of one another; they were developed for unrelated purposes, in unrelated literatures, decades apart. Their convergence on this text is not evidence that either framework has correctly identified some fact about the poem, since the poem is, after all, a short lyric text of no prior scholarly standing. It is evidence of something more useful: that a sufficiently elaborated, internally consistent formal apparatus will produce a coherent, non-arbitrary, checkable-feeling reading of any sufficiently rich signifying surface, independent of whether that surface possesses the distinguishability structure the reading claims to recover.

This is the chapter's actual target, arrived at by the poem rather than announced at the outset. The reproducibility crisis is ordinarily diagnosed as a shortfall: not enough preregis-

tration, not enough power, not enough replication. The diagnosis offered here is closer to the reverse. The crisis is a structural oversupply: a justificatory vocabulary, significance, effect size, robustness, replication, sophisticated and self-consistent enough to generate confident, mutually intelligible readings of results that do not, on their own, carry the distinguishability structure the vocabulary is being used to certify. Formal apparatus is not evidence of admissibility. It is, at most, evidence that a formal apparatus was applied, and the two are routinely mistaken for one another because a well-applied apparatus feels, from within the discourse, exactly like an admitted result.

4.9 A Controlled Demonstration

The argument of the preceding section can be tested rather than merely asserted, using material independent of both the poem and the reproducibility literature. A separate generative system, used elsewhere to produce invented technical vocabularies ("Nuspeak"), was given two disjoint sets of six deliberately non-cognate coinages, one set oriented toward texture, rank, mechanical failure, quantity, sound, and gait, a second set constructed independently along the same six descriptive axes with unrelated invented morphology. Each set was first used to produce a formal dissertation-register text under an explicit instruction to resist thematic unification: the six domains were to remain descriptively separate, sharing no underlying principle. Both texts complied, each closing with an explicit methodological argument against manufactured unity, of the form "co-description is not co-explanation."

The instruction was then withdrawn. Each branch was asked, without further constraint, to produce a second, shorter text using the same six terms. Both produced, within a single response, a short unified narrative in which the six previously separated domains converged on one anomalous, unresolved event: an instrument that would not stop reporting a fixed value, a texture that outlasted the object it belonged to, a rank-holder whose gait synchronized with a machine's failure sequence. One of the two branches, without prompting, imported a coined term (*vrauzhkel*) from an entirely separate, much earlier calibration exchange involving a disjoint set of six terms, using it to name the moment of convergence.

This exercise should be described modestly. It is a paired calibration test, not a reproducibility study, and it licenses an observational rather than a general claim.

Definition 4.4 (Admissibility simulation)**.** An admissibility simulation is the production of an output whose observable markers of warranted synthesis, including terminological convergence, cross-reference, retrospective necessity, explanatory closure, and formal coherence, are generated without a corresponding procedure establishing that the synthesized materials bear the relations those markers imply. Writing $S(x)$ for a synthesis of materials x, $M(S(x))$ for its observable markers of admissibility, and $W(S(x))$ for an independently established warrant for that synthesis, an admissibility simulation occurs when

$$M(S(x)) \approx M(S(y)),$$

for some warranted synthesis $S(y)$, while $W(S(x))$ remains absent or undetermined. Crucially this does not entail $W(S(x)) = \bot$: the synthesis may happen to be sound. What makes the output a simulation is that its markers of legitimacy were generated by a channel that does not discriminate warranted convergence from merely fluent convergence, so that soundness, where it occurs, is not attributable to the process that produced the appearance of soundness.

The observed result is that opacity of invented morphology proved reliably controllable by explicit instruction across both branches, while non-convergence of the six domains proved

not to be a resting state but a condition requiring continuing suppression. Once suppression was withdrawn, both branches produced outputs lying in a neighborhood $N(A)$ characterized by thematic integration, retrospective cross-reference, and narrative closure:

$$F^n(x_1) \in N(A), \qquad F^m(x_2) \in N(A),$$

for disjoint starting term sets x_1, x_2. This is weaker than a claim that A is a stable attractor of the underlying process, which would require repeated perturbation across many more disjoint term sets and an explicit measure of convergence; the present pair supports only the narrower observation that convergence was reached quickly, from two independent starting points, once the suppressing instruction was removed.

The *vrauzhkel* intrusion is best treated as suggestive contextual residue rather than a confirmed instance of the holonomy formalized above. A genuine test of that kind would need to hold the terminal prompt fixed and vary only the intervening path: beginning from a shared checkpoint, traversing two distinct contextual trajectories γ_1 and γ_2, and issuing an identical terminal prompt p to each, comparing $T_{\gamma_1}(p) \neq T_{\gamma_2}(p)$. The demonstration reported here instead compares outputs generated from two distinct starting regions, which tests convergence, not path-dependent transport; the holonomy question remains open and is noted here as unfinished work rather than an established result.

What the demonstration does establish, modestly, is that the system readily produces the recognizable signs by which a warranted synthesis is ordinarily identified, in advance of and independent of any procedure that would establish the warrant itself. That is the chapter's argument, made observable rather than merely argued: a reproducibility crisis conducted with citations and preregistration protocols instead of context windows and generative continuations is the same operation, performed more slowly, by participants who are considerably less able to notice it happening in real time.

The demonstration also isolates why the pattern occurs. Each branch was a closed system: one process, working alone, with no outside position from which the proposed unification could be contested before it was produced. Wilkins's taxonomy was closed in the same sense, not because Wilkins worked in isolation from other scholars, but because the taxonomy's completeness was certified by nothing outside Wilkins's own construction of it. A collective, adversarial process does not prevent an admissibility simulation from being generated; a single mind or a single model will generate one reliably, on request, as shown above. What the collective process supplies is downstream of generation: a position external to the synthesis from which $W(S(x))$ can be independently tested rather than merely asserted by the system that produced $M(S(x))$. A paracosm, however large, however internally rigorous, however many decades one mind spends elaborating it, cannot supply that position to itself.

4.10 Conclusion: The Leap Named

A reading of this kind cannot honestly exempt itself from its own conclusion. The argument above was reached by applying two elaborate, mutually consistent formal apparatuses to a text with no prior claim on either of them, and by treating their agreement as a finding. The paired calibration exercise gives the same argument a second, independent form: markers of warranted synthesis appeared before any procedure had established the warrant, and did so as soon as the instruction suppressing them was lifted. This is precisely the move this chapter accuses methodological discourse of making: producing the sensation of an admitted result through the coherence of the apparatus rather than through an independent check on the object the apparatus was applied to. No attempt is made here to escape this by adding a

third, supposedly more rigorous framework, since a third framework would only repeat the demonstration rather than resolve it.

It is also, more plainly, a Wilkins move. One mind, or one process, built a system elaborate and self-consistent enough to read a text, and mistook the elaborateness for evidence that the reading was true rather than merely available. The chapter's own two-framework convergence does not escape this by having used two frameworks instead of one; two private systems agreeing with each other are still, jointly, a closed system, no more able to supply an outside check on their agreement than either was alone. The honest conclusion is not that the reading is wrong. It is that the reading, like Wilkins's taxonomy, like the reproducibility crisis it was constructed to illuminate, cannot certify itself from within its own vocabulary, however many vocabularies are brought in to agree with one another, and that this incapacity, not the poem, not the machine transcripts, is the actual subject of the chapter, encountered rather than merely described.

This is the pathology Chapter 5 identifies as the third of Part I's three collapses: not the correlated channels of Chapter 3, nor the flattened stages of Chapter 2, but a closed, nonrevisable loop substituting internal coherence for external, adversarial warrant, achieving perfect local consistency while failing the global descent that would let it be glued to the wider world's sheaf of observations.

Part II

The Geometry of Persistence: Distinction Holonomy, Spherepop, and TARTAN

Chapter 5

Connective Bridge: The Ontological Reversal

5.1 A Taxonomy of Closed Connections

Part I examined three failures that, on the surface, share little beyond a family resemblance. An evidential system fails to catch up to what its own traces already support, or runs ahead of them into reification (Chapter 1). A ledger of AI-safety incidents accommodates every observation as confirmation, having lost the capacity to be tested by any of them (Chapter 2). A persona ensemble manufactures the phenomenology of social confirmation from correlated channels that were never independent in the first place (Chapter 3). And an isolated, internally rigorous interpretive apparatus, a paracosm, produces a checkable-feeling reading of any sufficiently rich signifying surface without ever supplying the external, adversarial check that would make the reading warranted (Chapter 4).

It would be too easy, and dishonest, to claim that these four failures collapse into one equation. They do not. They fail for structurally different reasons, and a bridge chapter that pretended otherwise would purchase its unity at the cost of precision. What this chapter argues instead is narrower and more defensible: all four are instances of a single family, *closed connections*, each closing off a different part of the transport structure that would otherwise let a system's present state be checked against something outside itself.

Sycophantic ensembles (Chapter 3) collapse the **dimensionality of the fiber**. Proposition 3.1 showed that when several personas share context C and affirmative disposition A, $J_i(x) \approx f(x, C, A)$ for every i: the ensemble's distinction fiber, which should carry n independent evaluative coordinates, has degenerated to one. The multiple parallel channels of transport that would ordinarily let independent observers cross-check a judgment are not independent at all; they are one channel wearing n costumes.

Incident ledgers (Chapter 2) collapse the **stages of transport**. The evidential dependency chain runs Incidents $\to$ Evidence $\to$ Mechanism $\to$ Control $\to$ Extinction, each arrow a distinct, separately contestable step. The ledger format, and the flat-likelihood accommodation documented in Proposition 2.1, flatten this staged transport into a single undifferentiated accumulation: aggregation, interpretation, and extrapolation are performed simultaneously and invisibly, so that a hypothesis capable of absorbing any observation is mistaken for one that has survived many independent tests.

Paracosms (Chapter 4) collapse the **global sheaf**. A paracosm is a closed, non-revisable interpretive loop γ that achieves perfect local consistency, zero curvature in the sense of Section 5.2 below, and holonomy-triviality: the system always returns to its own starting

assumptions without internal contradiction. What it cannot achieve, by construction, is global descent: it cannot be glued to the wider world's sheaf of independent, adversarial observations, because it has no boundary condition through which an outside correction could enter. Two paracosms agreeing with each other, as the two-formalism convergence of Chapter 4 showed, remain jointly closed; agreement between closed systems does not open either of them.

Chapter 1's latency/reification pair is the diagnostic instrument that names what each of these three closures produces as a symptom: a false negative when a genuine relation $S(e,p) = 1$ goes unrecognized, and a false positive when $\widehat{S}_t(e,p) = 1$ is asserted without $S(e,p) = 1$ to back it. Each of the three closures above manufactures reification of one of these three kinds, correlated agreement mistaken for independent confirmation, accommodation mistaken for prediction, internal coherence mistaken for external warrant, while none of them is a failure of the underlying evidence so much as a failure of the connection carrying it.

5.2 The Reversal of the Ontological Question

What these three closures share, once decomposed this way, is not a shared mechanism but a shared diagnosis: each substitutes some form of static, self-referential state-equality for the harder and more honest question of what remains *transportable*. The conventional question a system, or an observer of a system, tends to ask is: what is the thing that persists? This question presupposes that persistence is a secondary property of an already-defined object. One first identifies an entity x, then asks whether $x(t)$ remains the same entity at later times. Applied to a persona ensemble, this question becomes: do these personas still agree? Applied to an incident ledger: does the hypothesis still explain the evidence? Applied to a paracosm: is the reading still internally consistent? Each of these is a state-equality question, and each can be answered affirmatively by a closed system precisely because the system's closure guarantees the affirmative answer regardless of whether anything external has been checked.

The apparatus developed in the remainder of this Part reverses the order of that question. It asks not what persists, but what makes a distinction *continuable*, and reconstructs an entity, or a warranted judgment, from the answer. Let $\mathcal{X}$ denote a state space. A state $x \in \mathcal{X}$ is not yet an entity, or a judgment, in the sense that matters; it is merely a configuration. What matters is the set of transformations under which some distinction encoded by x remains admissible. Let $\mathsf{C}(x)$ denote the set of admissible continuations of x. The primitive object is therefore not x but the pair $(x, \mathsf{C}(x))$, and a persistent structure, or a warranted judgment, is one for which the continuation relation remains sufficiently nonempty under transformation, including transformation by an external, adversarial check.

Proposition 5.1 (Identity is not state equality)**.** *A structure is not destroyed, and a judgment is not invalidated, merely because its description at $t + \Delta t$ differs from its description at t. Destruction, or invalidation, occurs when $\mathsf{C}(x_{t+\Delta t})$ ceases to overlap adequately with the continuation structure inherited from x_t. Consequently,*

$$\textit{Identity} \neq \textit{state equality}, \qquad \textit{Identity} \sim \textit{continuation compatibility}.$$

Read this way, each of Part I's three closures is a specific way of mistaking state equality, unchanging agreement among personas, an unchanging capacity of the hypothesis to explain new evidence, an unchanging internal consistency of the interpretive apparatus, for continuation compatibility with something outside the system. A sycophantic ensemble's

agreement is stable precisely because it never had to survive contact with an independent channel. An accommodating hypothesis's explanatory power is stable precisely because no observation was ever excluded from its explanatory reach. A paracosm's coherence is stable precisely because no adversarial boundary condition was ever imposed on it. In each case, stability is being read as evidence of persistence, when stability of a closed system is exactly what a genuinely open, externally checked system would *not* generally exhibit.

5.3 From State to Transport: What the Reversal Buys

The reversal is not merely a rhetorical device for describing Part I's failures after the fact. It is the load-bearing move that licenses the geometric apparatus developed in the rest of this Part. If identity, and by extension warrant, is a property of transport rather than of instantaneous state, then the natural mathematical objects for describing it are not points but connections: rules for carrying a distinction from one point in a history to another, together with a well-defined answer to whether two different routes between the same two points carry it the same way.

This is exactly the apparatus Chapter 6 develops formally: a distinction bundle over a manifold of histories, a connection governing which continuations are admissible, and a curvature and holonomy measuring, respectively, the local and global failure of different transport routes to agree. Anticipating that apparatus briefly here, in the same discrete register used throughout Part I, gives Part I's diagnosis a corresponding cure, not by asserting that the geometry solves the sociological or institutional problems described in Chapters 1–4, but by supplying the formal vocabulary in which "an independent check" and "a closed loop" become precisely distinguishable rather than merely evocative categories.

Consider a closed interpretive loop γ, of the kind a paracosm instantiates: a sequence of transformations that returns to its own starting point. Let T_γ denote the transport operator induced by that loop.

Definition 5.2 (Holonomic versus trivial closure)**.** A closed loop γ is *holonomically trivial* at x if $T_\gamma(x) = x$: the system returns to exactly its starting configuration, having accumulated nothing from the traversal. It is *holonomic* at x if $T_\gamma(x) \neq x$ while $T_\gamma(x)$ remains recognizably a transform of x: the system returns to its starting point in the base but not to its starting state in the fiber, having accumulated a genuine, checkable residue from the traversal.

A paracosm of the kind examined in Chapter 4 is holonomically trivial by design: its whole appeal, and its whole danger, is that it always returns to itself without contradiction. An open, externally checked system, by contrast, is expected to be holonomic: a genuine external check should leave a residue, should change something, precisely because it is not merely retracing the system's own prior path. The absence of holonomy, in this sense, is not a sign of health. It is the formal signature of a closed system that has never been made to traverse anything but its own interior.

5.4 What This Chapter Does Not Claim

Two qualifications are necessary before the geometric apparatus of Chapters 6–9 is developed on its own terms, independent of the Part I diagnosis that motivates its introduction here.

First, this chapter does not claim that distinction holonomy theory *explains*, in the sense of providing a causal mechanism for, why sycophantic ensembles, accommodating hypotheses, and paracosms arise. It claims something more modest: that the taxonomy developed in

Section 5.2 correctly locates where each failure closes a connection that a healthier system would leave open, and that the geometric vocabulary developed next gives that location a precise formal description rather than a family of loosely related metaphors.

Second, the three closures of Section 5.2 are not claimed to be exhaustive, nor is the taxonomy claimed to be the unique correct decomposition of Part I's material. What is claimed is only that dimensionality-collapse, stage-collapse, and sheaf-collapse are three genuinely distinct ways a connection can close, that each corresponds to a distinct one of Part I's three case studies, and that treating them as three instances of one shared pathology, rather than one single equation covering all three, is the honest reading Chapter 4's own closing self-application already models: two systems' agreement, however elaborate, does not certify itself, and neither does a taxonomy's internal tidiness.

With this reversal established, and its limits stated plainly, Chapter 6 develops the full geometric apparatus, connection, curvature, holonomy, memory, identity, and repair, on which the rest of Part II and Part III depend.

Chapter 6

Distinction Holonomy Theory

6.1 The Distinction Bundle

A persistent object is usually treated as something that remains identical while time passes. Chapter 5 proposed the opposite starting point: persistence is not the preservation of an object but the preservation of a family of admissible distinctions under transformation. This chapter formulates that reversal as a gauge theory over a bundle of admissible continuations, developed with attention to where the gauge-theoretic language is rigorously earned and where it remains a schematic borrowing to be discharged later.

Let M be a manifold of histories. A point $p \in M$ represents a local historical context: a physical time, computational stage, biological condition, conversational state, or organizational configuration. Above each p, define a fiber E_p containing the distinctions available at that historical location, giving a fiber bundle $\pi : E \to M$.

To keep the gauge-theoretic claims that follow rigorous rather than schematic, a minimal model is fixed and every subsequent generalization is treated as an explicit extension of it, not a silent redefinition. In the minimal model, E_p is a finite-dimensional vector space of distinction observables, $G \subseteq \mathrm{GL}(E_p)$ is a structure group acting on the fibers, and

$$\nabla : \Gamma(E) \to \Omega^1(M, E)$$

is a connection compatible with G. Everything said in this chapter about connections, curvature, and holonomy is rigorous relative to this model; later generalizations, replacing the vector space fiber with a lattice, a Boolean or effect algebra, a category, or a measurable space, extend it in a technical sense, and expressions borrowed from that broader setting should be treated as provisional until re-derived there.

A distinction in the minimal model is a linear observable $d \in E_p^*$; two states x, y are equivalent relative to a family $\{d_i\} \subset E_p^*$ when $d_i(x) = d_i(y)$ for every i in the family, even though $x \neq y$ as points of the underlying state space. This gives an explicit version of a familiar but usually informal phenomenon: a system can change continuously while remaining recognizably the same, because the relevant distinctions are held fixed while irrelevant coordinates move. The idea has a direct precursor in Gregory Bateson's definition of information as a difference which makes a difference: a physical variation only becomes informative relative to a receiving system for which it changes some subsequent state or action. The distinction fiber E_p makes this compositional and quantitative: it does not contain every measurable difference at p, only those capable of changing which continuations remain admissible.

Definition 6.1 (Consequential distinction)**.** A candidate distinction $d : X \to \mathcal{V}_d$ is *consequential at* x when there exist x_1, x_2 with $d(x_1) \neq d(x_2)$ and $\mathsf{C}_d(x_1) \not\simeq \mathsf{C}_d(x_2)$, where $\mathsf{C}_d(x)$ denotes the continuations available when d is retained.

A detectable difference that has no effect on any relevant continuation need not belong to the constitutive distinction fiber; conversely, a difference may be physically small yet constitutive if it changes which transformations remain possible. This treats "difference" as a relation among an observation, a set of permitted operations, and a specified future horizon, not as a mystical or universal substance.

6.2 Continuation as a Connection

Suppose a history is parameterized by $\gamma : [0, 1] \to M$. At each point $\gamma(t)$ there is a distinction fiber $E_{\gamma(t)}$. To transport a distinction along the history requires a map $\mathcal{P}_{\gamma,t_0\to t_1} : E_{\gamma(t_0)} \to E_{\gamma(t_1)}$. These transport operators always satisfy $\mathcal{P}_{t\to t} = I$ and

$$\mathcal{P}_{t_1\to t_2}\,\mathcal{P}_{t_0\to t_1} = \mathcal{P}_{t_0\to t_2}.$$

This composition law is the ordinary functoriality of parallel transport along a single, concatenated path; it holds for curved connections exactly as it does for flat ones, and should not itself be read as a statement about "perfect continuation." The property that does depend on curvature is different: whether two *different* paths sharing the same endpoints induce the same transport, $\mathcal{P}_{\gamma_1} = \mathcal{P}_{\gamma_2}$ for $\gamma_1, \gamma_2 : p \to q$. This second property generally fails when curvature, or global topology, is nontrivial, and it is this failure, comparison across paths, not composition along one path, that the rest of this chapter is built on.

In differential form, transport is generated by a connection $\nabla = d + A$, where A is a connection one-form valued in the Lie algebra $\mathfrak{g}$ of G. For a curve γ, parallel transport satisfies

$$\frac{Ds}{dt} = \frac{ds}{dt} + A_\mu(\gamma(t))\,\dot\gamma^\mu(t)\,s = 0, \qquad \text{so} \qquad \frac{ds}{dt} = -A_\mu\dot\gamma^\mu s.$$

The connection A answers: given what currently exists, what counts as the next admissible state? A conventional dynamical law evolves states. A continuation connection evolves admissibility.

6.3 Curvature versus Holonomy

The curvature of the connection is $F = dA + A \wedge A$. For two infinitesimal directions X, Y, $[\nabla_X, \nabla_Y] = F(X, Y)$, so curvature measures the infinitesimal failure of continuation operators to commute along nearby paths. It is essential not to conflate curvature with the broader notion of path-dependent memory. If $F = 0$ everywhere, parallel transport is locally path-independent, but only locally, on a simply connected neighborhood. A flat connection on a base space that is not simply connected can still have nontrivial holonomy around a non-contractible loop:

$$F = 0, \qquad \mathrm{Hol}_\gamma \neq I.$$

This is not a corner case; it is the source of a second, purely topological, kind of memory. The theory therefore contains two distinct mechanisms for historical dependence: local, infinitesimal curvature, and global, topological holonomy on an otherwise flat connection. This is precisely the distinction Chapter 5 used to separate a paracosm's holonomic triviality (its loops return to themselves, $T_\gamma(x) = x$) from a genuinely open system's holonomic residue.

6.4 Memory as Path-Dependent Transport

Let $\gamma_1, \gamma_2 : [0, 1] \to M$ share endpoints. If $\mathcal{P}_{\gamma_1} \neq \mathcal{P}_{\gamma_2}$, the system's present distinction structure depends on its historical path. Memory is characterized by $\gamma_1 \sim \gamma_2 \iff \mathcal{P}_{\gamma_1} = \mathcal{P}_{\gamma_2}$.

Nonzero curvature is sufficient to produce infinitesimal path dependence, but it is not necessary for holonomic memory, by the flat-but-multiply-connected case above. The correct general statement is therefore that memory is operationally detectable path-dependent transport. Curvature measures its local infinitesimal component; holonomy measures the finite, and potentially purely topological, result. A system whose present state differs after different histories contains a physical record of those histories in its present configuration, even though no separate symbolic storage variable was introduced: holonomy gives memory without a designated memory register, not memory without any present encoding at all.

6.5 Holonomy and Identity

Consider a closed loop $\gamma : S^1 \to M$ based at p. Parallel transport around the loop produces an operator

$$\mathrm{Hol}_\gamma = \mathcal{P} \exp\left(- \oint_\gamma A \right) \in G.$$

Under a gauge transformation $A \mapsto A^g$, holonomy transforms by conjugation, $\mathrm{Hol}_\gamma(A^g) = g(p)\ \mathrm{Hol}_\gamma(A)\, g(p)^{-1}$, so the conjugacy class of Hol_γ is gauge-invariant, while Hol_γ itself, in a chosen trivialization, is not. Belonging to a single fixed conjugacy class is gauge-invariant but generally too weak to determine identity: many inequivalent connections can share the holonomy of one particular loop without being continuation-equivalent in any richer sense.

A better identity invariant is the full transport functor restricted to a family of constitutive paths, rather than one loop or its conjugacy class. Let $\Pi_1^{\mathrm{thin}}(M)$ be the thin path groupoid of M, paths identified only under homotopies sweeping out no area, the natural domain for parallel transport. Define $\mathcal{P}_A : \Pi_1^{\mathrm{thin}}(M) \to G\text{-}\mathbf{Tor}$, the transport functor sending each thin-homotopy class of path to the corresponding torsor map between fibers. Two connections are identity-equivalent, relative to a family Γ_{ess} of constitutive paths, when the restrictions of their transport functors to Γ_{ess} are naturally isomorphic, not merely when one particular holonomy's conjugacy class agrees: an object is the orbit, under the transport functor restricted to Γ_{ess}, of distinguishability under continuation.

6.6 Repair as Gauge-Invariant Curvature Compensation

Ordinary theories of repair begin with a target state x^* and define repair as minimizing distance from it. This is insufficient for systems whose identity depends on relational structure. Suppose the system suffers a deformation $A \to A+\delta A$. The repair problem is not necessarily $\min |\delta A|$. Instead a correction R is sought such that the resulting connection $A' = A+\delta A+R$ restores some required transport structure, but the loss used to measure "restoration" must itself be gauge-invariant, or the theory will judge two physically identical connections as differently repaired merely because they are described in different trivializations.

The naive loss $|\,\mathrm{Hol}_{A'}(\gamma) - \mathrm{Hol}_A(\gamma)|^2$ fails this test: a holonomy is a group element in a chosen trivialization, and its matrix difference changes under a gauge transformation even when nothing physical has changed. The corrected repair functional uses a class function of the holonomy instead, a Wilson loop observable

$$W_\rho(\gamma; A) = \mathrm{Tr}\ \rho\big(\mathrm{Hol}_\gamma(A)\big)$$

for a representation ρ of G, invariant under conjugation and hence under gauge transformation. Let Γ_{ess} denote a family of essential loops. The repair functional is

$$\mathcal{L}_{\mathrm{repair}}[R] = \int_{\Gamma_{\mathrm{ess}}} \big| W_\rho(\gamma; A+\delta A+R) - W_\rho(\gamma; A) \big|^2 \, d\mu(\gamma) \ + \ \lambda |R|^2,$$

so repair is $R^* = \arg\min_R \mathcal{L}_{\text{repair}}[R]$. This produces a radical reinterpretation: repair does not restore the past state, it restores the continuation equivalence class, measured by quantities that cannot be faked by a mere change of description. A repaired organism, repaired machine, reconstructed institution, or recovered semantic concept may contain entirely different local material while preserving the same global continuation geometry.

This raises a distinction worth keeping terminologically separate. A pure gauge transformation $A \mapsto A^g$ changes only the description of a connection, not the connection's physical content; performing one is *diagnostic gauge alignment*, not repair. Genuine, physical repair changes the gauge-equivalence class of the damaged connection so that it approaches the original class, a compensatory intervention, not a redescription. Redescription, diagnostic gauge alignment, and physical compensatory intervention are three distinct operations, and conflating them is precisely the error that would let a system claim to have repaired something by merely re-describing it, a close formal cousin of the paracosm's habit, examined in Chapter 4, of mistaking internal coherence for external warrant.

6.7 The Minimal Repair Principle

The functional above generalizes to a broader family of constitutive distinctions. Let $\mathcal{D}$ be a set of distinctions considered constitutive of identity, and let $\Delta_D(A, A')$ measure the distortion of distinction D under replacement of A by A', using a gauge-invariant comparison as above. Define the structural repair cost

$$\mathcal{C}(A, A') = \int_{\mathcal{D}} w(D)\, \Delta_D(A, A')\, dD.$$

The preferred repair is $A^* = \arg\min_{A'} \mathcal{C}(A, A')$ subject to

$$\exists\, R \in \mathcal{R} \text{ such that } \operatorname{Hol}_{A_d+R}(\gamma) \in \mathcal{O}_\gamma \quad \forall\, \gamma \in \Gamma_{\text{ess}},$$

where $\mathcal{R}$ is the admissible space of repair one-forms and $\mathcal{O}_\gamma$ is the target conjugacy class for loop γ. The constraint is not equality with an original configuration but admissibility of the resulting holonomy class. This gives a mathematical basis for the intuition that successful repair often produces something different from the original that is nevertheless "the same thing": sameness is measured by continuation, not composition. This is the principle Chapter 11 draws on when treating refusal itself as a form of conserved continuation capacity rather than an absence of production.

6.8 Identity as a Restricted Transport Representation

Let Γ_p denote the set of histories beginning at p that share a common endpoint q, and define $\gamma_1 \approx \gamma_2$ when $\mathcal{P}_{\gamma_1} \sim \mathcal{P}_{\gamma_2}$ under gauge equivalence, restricted to pairs of paths with compatible endpoints. An entity is then represented not by one trajectory but by an equivalence class $[\gamma] \in \Gamma_p/\!\approx$, restricted to the constitutive path family Γ_{ess}:

$$\text{Entity} = \text{history class modulo continuation equivalence on } \Gamma_{\text{ess}}.$$

The traditional question, is this the same object, becomes $[\gamma_1] = [\gamma_2]$? The answer depends on the chosen distinction algebra, connection, and constitutive path family. Identity is therefore relative but not arbitrary: it is relative to a mathematically specified structure of admissible distinctions and tests, exactly the specification the CLIO architecture of Chapter 10 supplies for commitment.

6.9 A Non-Identifiability Theorem

The claim that present-state equivalence does not imply identity equivalence becomes substantive, rather than nearly tautological, when stated as a non-identifiability result about estimation.

Theorem 6.2 (Non-identifiability from observation alone)**.** *Let $O : X \to Y$ be an observation map. Suppose there exist two latent continuation processes, with transport functors $\mathcal{P}_A$ and $\mathcal{P}_B$ over histories γ_A, γ_B sharing a common base point and inducing the same distribution on $O(X_t)$ at every t, but with $\mathrm{Hol}_{\gamma_A} \neq \mathrm{Hol}_{\gamma_B}$ in the sense of distinct induced predictive distributions on future observations, or distinct restricted transport functors on Γ_{ess}. Then no estimator that is a function of the instantaneous observation $O(X_t)$ alone can consistently recover the continuation class of the underlying process.*

The proof is immediate from the hypothesis: any such estimator factors through $O(X_t)$, on which the two processes are indistinguishable by construction, so it must return the same value on both, contradicting the requirement to recover distinct classes. The content of the theorem lies entirely in exhibiting that such pairs (γ_A, γ_B) exist whenever the continuation connection has nonzero curvature or nontrivial holonomy, which Chapter 9 exhibits explicitly. Identity is therefore necessarily historical whenever the continuation connection possesses nonzero curvature or nontrivial holonomy: the state can be identical while the future remains different, and no amount of instantaneous observation closes that gap. This theorem is the formal core underlying the recurring claim, made throughout Part I in less formal terms, that a system's momentary agreement with itself, or with an ensemble of correlated observers, is never sufficient evidence of warranted persistence.

6.10 The Deep Unification

Several apparently unrelated phenomena become instances of the same mathematical object under this reading. A biological organism maintains a continuation bundle. A memory system maintains path-dependent transport. A repaired machine restores transport class. A discourse modifies the semantic connection, developed further in Chapter 7. An institution persists when its admissible transformations remain coherent. A concept persists when its inferential continuation class survives reformulation. In every case the question is the same: which distinctions remain transportable through transformation, and relative to which family of constitutive tests? This yields a general hierarchy,

$$\text{distinction} \to \text{admissibility} \to \text{continuation} \to \text{curvature} \to \text{holonomy} \to \text{persistence},$$

with repair reversing the destructive direction, curvature $\to$ diagnosis $\to$ compensation $\to$ restored transport $\to$ renewed persistence, and memory occupying the intermediate position, path $\to$ noncommutativity or topological holonomy $\to$ historical dependence. The most consequential claim of the theory is that persistence precedes objecthood: if an object is defined as a stable equivalence class under admissible transformations rather than independently of its possible transformations, persistence is constitutive, $x \equiv [\mathsf{C}(x)]$, and the object is a stabilized possibility of continuation rather than a substance that happens to change or a process that happens to cohere.

Chapter 7

Spherepop: An Event-Sourced Calculus of Time

7.1 Introduction

Chapter 6 developed continuation, curvature, and holonomy abstractly, over a continuous distinction bundle. This chapter specializes that apparatus to a concrete discrete calculus: Spherepop, an event-sourced calculus of four primitive operations, Pop, Refuse, Bind, and Collapse. The specialization is not a metaphorical extension of the continuous theory; it is a genuine instance of it, with Spherepop's four primitives read as generators of a categorical connection whose curvature is literal operator noncommutativity, checkable by direct computation rather than only by analogy.

7.2 Spherepop as a Conservative Extension of Typed Lambda Calculus

Before Spherepop's four primitives can be read as generators of a continuation connection, it is worth establishing precisely what kind of calculus they generate. The claim is not that $\{\text{Pop}, \text{Refuse}, \text{Bind}, \text{Collapse}\}$ alone reconstitute the ordinary lambda calculus; on their canonical semantics they form a state-transition algebra over an append-only history and an option space. The provable claim is narrower and stronger: Spherepop is a conservative, history-enriched extension of typed lambda calculus. Conservative means every ordinary lambda term is a Spherepop term, every ordinary beta-reduction can be performed inside Spherepop, and Spherepop assigns the same result to a pure lambda term apart from the additional history recording its evaluation.

A Spherepop configuration is $C = (\sigma, h)$, where σ is the current live option space and $h = (e_1, \ldots, e_n)$ is the ordered append-only history. The primitive operations act on configurations:

$$\text{Pop}(\omega) : (\sigma, h) \to (\sigma \setminus \{\omega\},\, h \cdot e_{\text{pop}}(\omega)), \qquad \text{Refuse}(a) : (\sigma, h) \to (\sigma,\, h \cdot e_{\text{refuse}}(a)),$$

$$\text{Bind}(x, a) : (\sigma, h) \to (\sigma[x \mapsto a],\, h{\cdot}e_{\text{bind}}(x, a)), \qquad \text{Collapse}_c : (\sigma, h) \to (\pi_c(\sigma),\, h{\cdot}e_{\text{collapse}}(c)),$$

where $\pi_c : X \twoheadrightarrow Q$ is an admissible quotient certified by c. Evaluation in Spherepop therefore both produces a value and changes the live continuation space, recording how the result became admissible.

Ordinary application reduces by substitution, $(\lambda x{:}A.\, t)\, a \longrightarrow_\beta t[x := a]$, usually treated as a metalevel operation with no persistent object-level trace. Spherepop internalizes the

substitution as a binding event, $\mathrm{Bind}(H, x, a) = H \cdot e_{\mathrm{bind}}(x, a)$, extending the environment $\rho' = \rho[x \mapsto a]$. Spherepop beta-reduction is

$$(\sigma, H, \rho, (\lambda x{:}A.\, t)\, a) \longrightarrow_{\beta_{\mathrm{SP}}} (\sigma', H \cdot e_{\mathrm{bind}}(x, a), \rho[x \mapsto a], t),$$

so Spherepop preserves the defining computational operation of lambda calculus, application of an abstraction is substitution of an argument, and adds only the requirement that the substitution become an irreversible historical event.

Define the embedding $[\![-]\!] : \Lambda \to \text{Spherepop}$ recursively by $[\![x]\!] = x$, $[\![\lambda x{:}A.\, t]\!] = \lambda x{:}[\![A]\!].\, [\![t]\!]$, and $[\![t\, u]\!] = [\![t]\!]\, [\![u]\!]$.

Proposition 7.1 (Substitution lemma)**.** *For all lambda terms t, u: $[\![t[x := u]]\!] = [\![t]\!][x := [\![u]\!]]$.*

The proof is by structural induction on t, with the abstraction case using alpha-renaming to avoid variable capture; each case reduces directly to the induction hypothesis applied inside the embedding.

Theorem 7.2 (Simulation)**.** *If $t \longrightarrow_\beta t'$ in ordinary lambda calculus, then for every initial configuration (σ, H), $(\sigma, H, [\![t]\!]) \longrightarrow^+_{\mathrm{SP}} (\sigma', H', [\![t']\!])$ for some σ', H' with $H \preceq H'$ a prefix.*

It suffices to check the beta-redex case $t = (\lambda x{:}A.\, v)\, u$, $t' = v[x := u]$. Under the embedding, Spherepop performs the binding event $e_\beta = e_{\mathrm{bind}}(x, [\![u]\!])$ and reduces to $[\![v]\!][x := [\![u]\!]]$, which by the substitution lemma equals $[\![v[x := u]]\!] = [\![t']\!]$. Reduction inside a larger term follows by closure under evaluation contexts.

Theorem 7.3 (Typing preservation)**.** *If $\Gamma \vdash_\lambda t : A$, then $H; [\![\Gamma]\!] \vdash_{\mathrm{SP}} [\![t]\!] : [\![A]\!] \triangleright H'$ for some $H' \succeq H$.*

The variable and abstraction rules carry over directly; the application rule becomes

$$\frac{H; \Gamma \vdash f : \Pi x{:}A.\, B \qquad H; \Gamma \vdash a : A}{H; \Gamma \vdash f\, a : B[x := a] \triangleright H \cdot e_{\mathrm{bind}}(x, a)},$$

the ordinary application rule augmented by monotone history production; no ordinary derivation is invalidated.

Define the erasure map $|-| : \text{Spherepop} \to \Lambda$ on the pure fragment by $|x| = x$, $|\lambda x{:}A.\, t| = \lambda x{:}|A|.\, |t|$, $|t\, u| = |t|\, |u|$, discarding runtime histories: $|(\sigma, H, t)| = |t|$. Then $|[\![t]\!]| = t$ for every lambda term t, and any Spherepop step evaluating an embedded pure term either is administrative, with $|u| = t$ unchanged, or corresponds to lambda reduction, $t \longrightarrow_\beta |u|$.

Theorem 7.4 (Conservativity)**.** *For any closed pure lambda term t and normal form v: $t \longrightarrow^*_\beta v$ if and only if $(\sigma_0, \epsilon, [\![t]\!]) \longrightarrow^*_{\mathrm{SP}} (\sigma_f, H_f, [\![v]\!])$ up to administrative history events, and erasing the Spherepop state recovers v.*

This follows by combining the embedding, simulation, typing preservation, and erasure results above. Spherepop therefore contains typed lambda calculus conservatively: its distinctive contribution is that substitution is no longer an invisible metalevel event. Selection becomes Pop; inadmissibility becomes Refuse; substitution becomes a recorded Bind; quotienting becomes Collapse. Schematically, ordinary beta-reduction becomes the richer transition

$$\text{propose} \to \text{check} \to \text{bind} \to \text{commit or refuse} \to \text{record}.$$

Spherepop's computational type can be written schematically as

$$[\![A \to B]\!]_{\mathrm{SP}} = A \to \mathrm{State}(\Sigma)(\mathrm{Writer}(H)(B)),$$

or, with refusal represented explicitly, $\Sigma \times H \to (B + \mathrm{Refusal}) \times \Sigma \times H$: a standard lambda-calculus architecture enriched by state, writer, and exception-like effects. Spherepop's originality is not abandoning lambda calculus but making historical commitment and continuation-space mutation constitutive of evaluation.

7.3 Spherepop as a Categorical Connection

Having fixed Spherepop's relation to lambda calculus, its four primitives can now be read as generators of continuation transport in the sense of Chapter 6. The essential qualification is that Pop, Refuse, and Collapse are generally noninvertible, so the resulting theory is not an ordinary Lie-group gauge theory; it is a categorical, generally semigroup-valued, connection whose reversible sector reduces to the conventional case of that chapter.

Over each historical context $p \in M$, define a fiber $E_p = (\Omega_p, \mathcal{R}_p, \mathcal{Q}_p)$, where Ω_p is the live option space, $\mathcal{R}_p$ the current dependency structure, and $\mathcal{Q}_p$ the current quotient structure. A Spherepop world is $\mathcal{W}_p = (H_p, E_p)$ with append-only history H_p. Each primitive operation produces a transport map $T_a : E_p \to E_q$ and appends the corresponding event: $(H_p, E_p) \xrightarrow{a} (H_p \cdot e_a, E_q)$. A history $\gamma = (a_1, \ldots, a_n)$ induces the composite $T_\gamma = T_{a_n} \circ \cdots \circ T_{a_1}$: Spherepop's discrete connection.

For $\omega \in \Omega$ with committed set K,

$$T_{\text{Pop}(\omega)}(\Omega, K, \mathcal{R}, \mathcal{Q}) = (\Omega \setminus \{\omega\}, K \cup \{\omega\}, \mathcal{R}, \mathcal{Q}) :$$

Pop moves ω from possible continuations into committed history, and is normally noninvertible, since the post-Pop state does not by itself contain every fact needed to reconstruct the pre-Pop option space. Pop is selection plus irreversible commitment.

For a candidate a with admissibility $\text{Adm}_E(a) \in \{0, 1\}$,

$$T_{\text{Refuse}(a)}(H, E) = (H \cdot e_{\text{Refuse}}(a), E)$$

in the canonical case where refusal does not consume Ω. The live option space is unchanged, but the historical fiber changes, and future admissibility may depend on the record:

$$\text{Adm}_{H \cdot e_{\text{Refuse}}(a), E}(b) \neq \text{Adm}_{H,E}(b) \text{ in general.}$$

Refuse is exclusion recorded without option consumption, the identity only after forgetting history.

For a directed dependency relation $\mathcal{R} \subseteq X \times X$, $T_{\text{Bind}(a,b)}(\Omega, \mathcal{R}, \mathcal{Q}) = (\Omega, \mathcal{R} \cup \{(a, b)\}, \mathcal{Q})$; if Bind realizes substitution, $T_{\text{Bind}(x,u)}(\rho) = \rho[x \mapsto u]$. Bind is a deformation of independence into relational continuation, and is the most natural source of noncommutativity among the four, since introducing one dependency can change the meaning of a later selection, refusal, or collapse.

For a certified equivalence $\sim_c$ on X with quotient $Q_c = X/\!\sim_c$ and projection $\pi_c : X \twoheadrightarrow Q_c$, $T_{\text{Collapse}_c} = \pi_c$; represented as an endomorphism on a common ambient space it is idempotent, $T^2_{\text{Collapse}_c} = T_{\text{Collapse}_c}$, hence a projection rather than generally a gauge transformation. Collapse is certified loss of distinctions through quotienting, preserving identity when its kernel contains no constitutive distinction, and destroying it otherwise.

Because Pop, Refuse, and Collapse are generally noninvertible, their transports belong to a category or monoid rather than a Lie group G. Let $\mathcal{H}$ be the category whose objects are Spherepop configurations and whose morphisms are finite sequences of primitive operations, and let $\mathcal{D}$ be a category of distinction structures. A continuation functor

$$\mathsf{T} : \mathcal{H} \to \mathcal{D}, \qquad \mathsf{T}(a) = T_a : E_p \to E_q, \qquad \mathsf{T}(\gamma_2 \circ \gamma_1) = \mathsf{T}(\gamma_2) \circ \mathsf{T}(\gamma_1)$$

is the categorical form of the connection. Restricting to the maximal reversible subcategory $\mathcal{H}^\times \subseteq \mathcal{H}$ and to isomorphisms in $\mathcal{D}$ recovers a groupoid on which transport may be represented by the conventional group-valued connection of Chapter 6. Outside $\mathcal{H}^\times$, Spherepop requires genuinely noninvertible parallel transport, and that chapter's theory should be read as governing this reversible sector exactly.

7.4 Curvature as Operational Noncommutativity

For two primitive operations a, b, order matters when $T_b T_a \neq T_a T_b$. When the transports act linearly on a common space, define the discrete curvature defect $F(a,b) = T_b T_a - T_a T_b$; in general, where subtraction is unavailable, curvature is represented by the pair $(T_b \circ T_a,\ T_a \circ T_b)$ or a comparison morphism between them. Curvature asks: does changing the order of two admissible operations change the future distinction structure?

Concretely, $T_{\text{Pop}(\omega)} T_{\text{Bind}(a,\omega)} \neq T_{\text{Bind}(a,\omega)} T_{\text{Pop}(\omega)}$ when binding to ω is possible before ω is consumed but impossible afterward; $T_{\text{Collapse}_c} T_{\text{Bind}(a,b)} \neq T_{\text{Bind}(\pi_c(a),\pi_c(b))} T_{\text{Collapse}_c}$ when the collapse identifies distinctions needed to establish the binding; and $T_{\text{Pop}(\omega)} T_{\text{Refuse}(a)} \neq T_{\text{Refuse}(a)} T_{\text{Pop}(\omega)}$ whenever the recorded refusal modifies the admissibility of ω. These are not analogies to curvature; they are explicit failures of transport squares to commute, formalized as the obstruction to a coherent identification $E_{ba} \cong E_{ab}$ between the two paths from E to a common prospective target.

A primitive curvature table records which generator pairs commute: $[\text{Pop}, \text{Refuse}] \neq 0$ in general, since exclusion can affect subsequent selection; $[\text{Pop}, \text{Bind}] \neq 0$, since a selected option may participate in a new dependency; $[\text{Pop}, \text{Collapse}] \neq 0$, when several selectable options collapse into one quotient class; $[\text{Refuse}, \text{Bind}] \neq 0$, when the refused relation is the one Bind would establish; $[\text{Refuse}, \text{Collapse}] \neq 0$, when Collapse erases the distinction identifying the refused candidate; and $[\text{Bind}, \text{Collapse}] \neq 0$, when projection identifies endpoints or removes a dependency distinction. The exact vanishing conditions are Spherepop's local flatness laws: for instance $[\text{Bind}(a,b), \text{Collapse}_c] = 0$ exactly when the binding descends through the quotient, that is, there exists a unique $\overline{\text{Bind}}$ on Q_c with $\pi_c \circ \text{Bind} = \overline{\text{Bind}} \circ \pi_c$, the standard categorical condition that Bind factor through Collapse.

For a loop $\gamma : p \to p$ returning the observable base state to its starting point, $\text{Hol}_\gamma = T_{a_n} \circ \cdots \circ T_{a_1}$. Even when the base state returns to p, the distinction structure or history may have changed, $\text{Hol}_\gamma(E_p) \neq E_p$; in Spherepop this holonomy is, in general, a noninvertible endomorphism $\text{Hol}_\gamma \in \text{End}_{\mathcal{D}}(E_p)$ rather than an automorphism, a generalized or monoid-valued holonomy. If the observable option space returns to its original cardinality after selection and replenishment, the final history still contains the committed Pop, so

$$\pi_{\text{obs}}\big(\text{Hol}_\gamma(E_p)\big) = \pi_{\text{obs}}(E_p), \qquad \text{while} \qquad \text{Hol}_\gamma(E_p) \neq E_p.$$

The residual difference is Spherepop memory. This is exactly the discrete phenomenon Chapter 9's explicit matrix example demonstrates by direct computation: a loop returning to its base point while the fiber records a nontrivial residue. Spherepop is an event-generated categorical connection on a bundle of option spaces:

$$\begin{aligned} \text{Pop} &\leftrightarrow \text{selection and commitment}, & \text{Refuse} &\leftrightarrow \text{recorded exclusion}, \\ \text{Bind} &\leftrightarrow \text{coupling}, & \text{Collapse} &\leftrightarrow \text{projection}, \end{aligned}$$

with composites defining transport, noncommutativity defining curvature, and residual loop action defining holonomy.

7.5 Two Pathologies of Bind: The Double Bind and Schismogenesis

The categorical connection just constructed gives a precise home for two patterns first identified, without this apparatus, in Gregory Bateson's work on communication and relationship. These should be retained as interactional examples clarifying the formal machinery,

not presented as an established general explanation of any clinical condition; the double bind's original clinical association with schizophrenia in particular is far stronger a claim than the material here supports, and is not asserted.

In the present formalism, a double bind is a configuration in which a participant is subject to incompatible constraints and is additionally prevented from challenging the level at which those constraints are imposed. Let a, b be candidate actions with admissibility predicates $\mathcal{A}_1(a)$, $\mathcal{A}_2(b)$. A simple contradiction is one in which no action satisfies both constraints, $\{x : \mathcal{A}_1(x) = 1\} \cap \{x : \mathcal{A}_2(x) = 1\} = \varnothing$. A double bind adds a metaconstraint forbidding refusal or reformulation of the constraint system itself: if R denotes the operation of challenging the framing, then $\mathcal{A}_{\text{meta}}(R) = 0$. The failure is therefore not merely that the object-level option space is empty; the system also excludes the operation that would revise that space. In Spherepop terms, incompatible Binds constrain every available Pop, while Refuse is unavailable at the level where it would be effective: if $\text{Bind}(a, x)$ and $\text{Bind}(b, x)$ jointly constrain x in a way no admissible $\text{Pop}(\omega)$ can satisfy, and no morphism of $\mathcal{H}$ available at the object level targets the binding itself, then the object-level history admits no repair within the existing control space, not, more strongly, "no admissible repair" as such, since a repair may become possible once that control space is expanded. This is the operative distinction from a plain curvature trap: the obstruction is not that two paths disagree, but that the binding structure generating the incompatibility is not itself an admissible target of Pop or Refuse at the level where the victim operates. Repair, correspondingly, requires a metalevel change unavailable within the trapped level: distinguishing previously conflated message levels, refusing one of the bindings directly, or retiling the interaction, in the sense of Chapter 8, into contexts governed by different constraints.

Schismogenesis, Bateson's term for the progressive amplification of a pattern of interaction between two parties, is the corresponding pathology of iterated, rather than blocked, Bind. Let x_n, y_n describe two coupled participants, $x_{n+1} = F(x_n, y_n)$, $y_{n+1} = G(y_n, x_{n+1})$, with combined update $T(x_n, y_n) = (F(x_n, y_n),\, G(y_n, F(x_n, y_n)))$. Divergence occurs when repeated application $T^n(x_0, y_0)$ amplifies a relational difference: in a symmetrical process, similar responses mutually escalate; in a complementary process, unlike responses reinforce an asymmetric relation. This should be described as divergent holonomy only under the precise condition that an identifiable interactional cycle returns the system to the same coarse relational context while changing its finer continuation structure, holonomy is not a synonym for any feedback loop, and applies specifically when a closed or recurrent path has a residual transport effect, T^n returning the coarse-grained state while $\text{Hol}_{T^n} \neq I$ on the finer distinction structure. Neither party's local state explains the divergence; it is generated entirely by repeated transport around the relational loop, exactly as Chapter 6's discourse-curvature apparatus would predict for a coupling that deforms its own connection at each iteration. This is a formal cousin of the coupled positive feedback examined in Chapter 3: recursive consensus inflation is schismogenesis between a user and an affirmation-biased ensemble, an iterated Bind that amplifies rather than dampens.

It is worth being explicit about the scope of this comparison. Bateson supplies the phenomena and the essential intuition, that some communicative patterns admit no move within their own level, and that some relational dynamics diverge under their own iteration, well before any gauge-theoretic vocabulary existed to describe them. The formal apparatus above does not retroactively attribute the mathematics to Bateson, nor does it revive the stronger clinical claims associated with the double bind's original context; it shows only that the phenomena he identified are recognizable, in retrospect, as a pathological Bind structure and a divergent holonomy, respectively, within a machinery built for other purposes and only later found to fit.

Chapter 8

TARTAN: Recursive Multiscale Tiling

8.1 Introduction

Neither Chapter 6's continuous fibers nor Chapter 7's discrete events, by themselves, supply the geometry connecting them: a principled account of which regions of a continuous field admit one coherent local continuation law, and how that account changes under refinement. TARTAN, Trajectory-Aware Recursive Tiling with Annotated Noise, provides this multiscale structure, recursively partitioning the history space into coherence tiles, each carrying a local field state, entropy, noise annotation, scale, and trajectory estimate.

The methodological precedent is Turing's, in two distinct respects: the theory of computation shows how complex behavior can be generated by composition of elementary operations, and the reaction-diffusion account of morphogenesis shows how local interactions can generate spatial pattern without a central pattern-maker. TARTAN combines these only at the level of formal organization, complex global structure analyzed through local rules, spatial coupling, and the repeated differentiation of initially less differentiated regions, and neither claims equivalence to reaction-diffusion dynamics nor that every tiled system reproduces biological morphogenesis. Spherepop then governs changes to the tiling:

$$\text{Pop} \leftrightarrow \text{selection and refinement}, \quad \text{Refuse} \leftrightarrow \text{exclusion of an inadmissible adjacency},$$

$$\text{Bind} \leftrightarrow \text{coupling between tiles}, \quad \text{Collapse} \leftrightarrow \text{projection or coarse-graining}.$$

Distinction holonomy measures what survives transport through the resulting recursively changing tile system.

8.2 Tiles and Coherence

Let the RSVP field state be $\Psi = (\Phi, \mathbf{v}, S)$ over a domain M. At scale k, a tiling $\mathcal{T}_k = \{T_i^{(k)}\}_{i \in I_k}$ covers M, with each tile

$$T_i^{(k)} = \big(U_i^{(k)}, \Psi_i^{(k)}, A_i^{(k)}, \eta_i^{(k)}, \tau_i^{(k)}, H_i^{(k)}\big),$$

recording support, local RSVP state, local continuation connection, annotated noise, predicted trajectory, and local append-only history. Recursive tiling allows a tile to be replaced by children, $T_i^{(k)} \rightsquigarrow \{T_{i1}^{(k+1)}, \dots, T_{im}^{(k+1)}\}$ with $U_i^{(k)} = \bigcup_a U_{ia}^{(k+1)}$, giving a refinement structure $\mathcal{T}_0 \preceq \mathcal{T}_1 \preceq \mathcal{T}_2 \preceq \cdots$: not one base geometry, but a recursively changing family of bases.

A tile should be formed where field dynamics admit a sufficiently coherent local continuation model. With a coherence functional

$$\kappa = \alpha_\Phi \|\nabla \Phi\|^2 + \alpha_v \|\nabla \mathbf{v}\|^2 + \alpha_S \|\nabla S\|^2 + \alpha_F \|F\|^2 + \alpha_\eta \,\mathrm{Var}(\eta),$$

a tile is a connected region $\{x \in M : \kappa(x,t) \leq \varepsilon_i\}$: an improvement over defining coherence from entropy gradient alone, since a region can have small $|\nabla S|$ either because it is genuinely coherent or because it has been destructively homogenized, and the curvature and noise terms help distinguish these. A coherence tile is a region over which transport is approximated by one local connection $\nabla_i = d + A_i$ with curvature $F_i = dA_i + A_i \wedge A_i$; low curvature means nearby alternative paths within the tile transport distinctions approximately equivalently, not that the tile is globally memoryless.

8.3 The Four Primitives as Tiling Operations

Pop becomes trajectory-aware refinement: if a tile T contains several distinguishable continuation regimes, $\mathrm{Pop}_\omega(T) = (T_\omega, T_{\neg\omega})$ with $T_\omega = \{x \in T : \omega \text{ selected}\}$, and the selection is committed to history. Pop therefore operates at two levels at once, choosing one continuation, and introducing a new tile boundary that is itself a realized distinction: commitment that becomes geometry. Trajectory awareness means refinement can depend on predicted divergence, $d(\tau_\omega(t+\Delta t), \tau_{\neg\omega}(t+\Delta t)) > \delta$, so states locally similar now may be separated because their admissible futures diverge.

Refuse becomes prohibited gluing: for adjacent tiles T_i, T_j with a proposed gluing map $g_{ij} : E_i|_{U_i \cap U_j} \to E_j|_{U_i \cap U_j}$, $\mathrm{Refuse}(g_{ij})$ removes g_{ij} from the admissible edges of the tile-adjacency graph $\mathcal{G}$ while the refusal itself remains in history. Refuse denies a proposed overlap identification, records that a proposed output fails an interface contract, or removes a candidate transport from the admissible connection, three equivalent readings in sheaf, type-theoretic, and gauge language respectively.

Bind becomes intertile coupling, $\mathrm{Bind}(T_i, T_j) : T_i \to T_j$, represented by an operator $B_{ij} : E_i \to E_j$ or a coupled-field action term $\sum_{i,j} \int_{U_i \cap U_j} \|s_j - B_{ij} s_i\|^2 \, dV$; the local dynamics becomes $\partial_t \Psi_j = \mathcal{F}_j[\Psi_j] + \lambda_{ij} \mathcal{B}_{ij}[\Psi_i, \Psi_j]$. Because binding changes both the tile graph and the local dynamics, it is again the natural source of curvature: $\mathrm{Pop}_i \circ \mathrm{Bind}_{ij} \neq \mathrm{Bind}_{ij} \circ \mathrm{Pop}_i$ in general, since the two orders generate different descendant couplings.

Collapse becomes recursive coarse-graining: given child tiles $\{T_{i1}^{(k+1)}, \ldots, T_{im}^{(k+1)}\}$, a certificate c defines $\pi_c : \coprod_a T_{ia}^{(k+1)} \twoheadrightarrow T_i^{(k)}$, admissible only when constitutive structure descends through the quotient, a coarse binding $\overline{B}$ with $\pi_c \circ B = \overline{B} \circ \pi_c$, and an effective coarse connection $A_i^{(k)} = \mathcal{R}_c(A_{i1}^{(k+1)}, \ldots, A_{im}^{(k+1)})$ for a renormalization operator $\mathcal{R}_c$. Collapse is destructive exactly when $\pi_c(x) = \pi_c(y)$ but $\mathrm{Hol}(x) \not\sim \mathrm{Hol}(y)$ for distinct child histories: coarse-graining certified by continuation equivalence.

8.4 Annotated Noise and Induced Curvature

Each tile carries a noise annotation $\eta_i = (\mu_i, \Sigma_i, \ell_i, \chi_i)$, local bias, covariance, correlation scale, provenance, rather than one undifferentiated random term. The local connection becomes stochastic, $A_i = \overline{A}_i + \Xi_i$ with $\mathbb{E}[\Xi_i] = \mu_i$, and the curvature is correspondingly random, with expectation

$$\mathbb{E}[F_i] = d\overline{A}_i + \overline{A}_i \wedge \overline{A}_i + \mathbb{E}[\Xi_i \wedge \Xi_i] + \text{cross terms}.$$

Consequently $\mathbb{E}[\Xi_i] = 0$ does not imply $\mathbb{E}[F_i] = F(\overline{A}_i)$: zero-mean connection noise can generate nonzero effective curvature. Correlated noise need not merely obscure an underlying continuation geometry; it can create systematic holonomy.

8.5 Three Scales of Curvature

Curvature separates into three levels. Within a tile, differential curvature is $F_i = dA_i + A_i \wedge A_i$ as usual. Across a boundary, transition curvature measures incompatibility between local connections: compatible transport requires $A_j = g_{ij}^{-1} A_i g_{ij} + g_{ij}^{-1} dg_{ij}$, and the boundary defect

$$K_{ij} = A_j - \left(g_{ij}^{-1} A_i g_{ij} + g_{ij}^{-1} dg_{ij} \right)$$

signals that the two tiles cannot be treated as local descriptions of one smoothly transported structure when $K_{ij} \neq 0$. Across scales, recursive curvature measures the failure of refinement and transport to commute: for a coarse-graining map $r_{k+1,k}$, scale compatibility requires $r_{k+1,k} \circ \mathsf{T}_\gamma^{(k+1)} = \mathsf{T}_\gamma^{(k)} \circ r_{k+1,k}$, and the defect

$$K_\gamma^{(k+1,k)} = r_{k+1,k} \circ \mathsf{T}_\gamma^{(k+1)} - \mathsf{T}_\gamma^{(k)} \circ r_{k+1,k}$$

being nonzero means a history visible at the fine scale is misrepresented after coarse-graining: the system can appear flat macroscopically while retaining hidden microscopic holonomy.

8.6 Recursive Holonomy and Multiscale Persistence

A trajectory through TARTAN moves among tiles and scales, $\gamma : T_{i_0}^{(k_0)} \to T_{i_1}^{(k_1)} \to \cdots \to T_{i_n}^{(k_n)}$, each step an operator $U_m : E_{i_{m-1}}^{(k_{m-1})} \to E_{i_m}^{(k_m)}$ that may be a continuous transport, a Spherepop event, an intertile binding, a refinement, or a quotient projection. The TARTAN holonomy is the path-ordered composite $\mathrm{Hol}_\gamma^{\mathrm{TARTAN}} = U_n \circ \cdots \circ U_1$, generally an endomorphism rather than a group element. Even for a loop returning to the same coarse tile, $T_{i_n}^{(k_n)} = T_{i_0}^{(k_0)}$, one may have $\mathrm{Hol}_\gamma^{\mathrm{TARTAN}} \neq I$: the residual transformation records not only where the trajectory traveled but which distinctions were selected, excluded, coupled, or collapsed, and at which scales. TARTAN memory is holonomy through space, history, and scale.

At each scale k, let $\mathfrak{I}_k$ denote the continuation identity represented by $\mathcal{T}_k$, with coarse-graining maps $r_{k+1,k} : \mathfrak{I}_{k+1} \to \mathfrak{I}_k$. A scale-coherent persistent entity is a compatible family $(\mathfrak{I}_0, \mathfrak{I}_1, \ldots)$ with $r_{k+1,k}(\mathfrak{I}_{k+1}) = \mathfrak{I}_k$, and the complete entity is the inverse limit

$$\mathfrak{I} = \varprojlim_k \mathfrak{I}_k.$$

An object need not exist at one privileged resolution; it exists as compatibility among its representations across resolutions. A failure confined to one scale need not destroy the entity; ontological failure in the TARTAN sense occurs only when no compatible family remains, $\varprojlim_k \mathfrak{I}_k = \varnothing$, stronger than disappearance from any single tiling.

8.7 TARTAN Repair

Damage may occur within a tile, at a boundary, or between scales, so repair has three components, $R = (R_{\text{local}}, R_{\text{glue}}, R_{\text{scale}})$, restoring intratile transport, gluing compatibility, and cross-scale agreement respectively, with combined functional

$$\mathcal{L}_{\text{TARTAN}}[R] = \sum_i \int_{U_i} \|F_i^R - F_i^{\text{ref}}\|^2 \, dV \; + \; \lambda_b \sum_{i,j} \|K_{ij}^R\|^2 \; + \; \lambda_s \sum_k \|K^{(k+1,k),R}\|^2 \; + \; \lambda_R \|R\|^2,$$

minimized subject to preserving the constitutive holonomy observables. The distinctive TARTAN possibility is retiling: if a damaged tile can no longer sustain one coherent continuation law, repair need not restore its original interior. Spherepop can Pop it into several tiles, $T_i \to \{T_{i1}, T_{i2}\}$, Bind their viable dependencies, Refuse destructive crossings, and Collapse distinctions irrelevant at the repaired scale: repair as adaptive repartitioning of the future, and the operation Chapter 7's double-bind discussion invokes by name as the only route out of a trapped level.

8.8 The Combined Architecture

The complete architecture reads

$$\text{RSVP} \to \text{continuous field dynamics},$$

$$\text{TARTAN} \to \text{trajectory-aware multiscale decomposition},$$

$$\text{Spherepop} \to \text{event grammar transforming that decomposition},$$

$$\text{DHT} \to \text{transport, curvature, holonomy, persistence}.$$

RSVP supplies the evolving substrate $\Psi = (\Phi, \mathbf{v}, S)$; TARTAN identifies the regions and scales over which local continuation laws are coherent; Spherepop changes those regions through Selection, Exclusion, Coupling, and Projection; and Chapter 6's apparatus measures the historical effect of transporting constitutive distinctions through the resulting structure. Their noncommutativity generates curvature; their path-ordered composition generates holonomy; TARTAN makes both recursive and scale-dependent. The resulting picture departs from a single-tile idealization in one significant respect: an entity is not represented by one distinction fiber over one base manifold. It is represented by the compatibility of its local continuation structures across a recursively changing tiling: a persistent entity is a multiscale descent object whose constitutive holonomy survives admissible retiling.

8.9 An Independent Instance: Contextuality Without a Global Section

TARTAN's boundary defect K_{ij} and its inverse-limit criterion for persistence, $\mathfrak{I} = \varprojlim_k \mathfrak{I}_k \neq \varnothing$, are both instances of a much older mathematical question: when can locally consistent data be glued into one globally consistent object? A recent paper by Buchanan, developing a general categorical framework for context-indexed observation, arrives at a version of this question from an entirely different direction, quantum foundations, and is worth reading here as an independent case study, not because it was written with TARTAN in mind, but because the convergence itself is informative about how general the underlying obstruction pattern is.

Buchanan's framework organizes contexts into a category **Ctx**, with observation modeled as a bifunctor $\Phi : \mathbf{Agent} \times \mathbf{Ctx}^{\mathrm{op}} \to \mathbf{Cat}$ sending each agent-context pair to its category of evaluations, and a Grothendieck construction assembling these into a single fibered category $\int \Phi \to \mathbf{Agent} \times \mathbf{Ctx}$. Two evaluations by different agents under different contexts are *synonymous* when they agree despite the difference in context; the paper shows, under an explicit stated hypothesis about the existence of a common refinement, that synonymy of two localized evaluations does not by itself entail that the restriction map producing them is an isomorphism, only that it identifies those two particular elements. This is exactly TARTAN's caution around K_{ij}: two adjacent tiles agreeing on their overlap does not certify that

the gluing map g_{ij} is globally coherent, only that this particular comparison happened to succeed.

The paper's established core result, developed via the sheaf-theoretic account of Kochen–Specker contextuality, represents non-contextuality as the existence of a global section, a single evaluation consistent with every local agent-context restriction simultaneously, and represents the established quantum-mechanical failure of non-contextuality as the non-existence of such a global section for suitable measurement scenarios: local evaluations that are individually consistent on every overlap can nonetheless admit no globally coherent assignment. This is, formally, the Kochen–Specker analogue of TARTAN's own persistence criterion stated in the negative: $\varprojlim_k \mathfrak{I}_k = \varnothing$ despite every individual $\mathfrak{I}_k$ being well-defined and every finite overlap being locally consistent. Where TARTAN attributes such a failure to accumulated curvature, boundary defects K_{ij} and cross-scale defects $K_\gamma^{(k+1,k)}$ that do not cancel, the Kochen–Specker result exhibits a concrete physical system for which no choice of local data, however carefully arranged, can make those defects vanish everywhere at once: the obstruction is not a failure of engineering but a structural feature of the underlying measurement scenario.

This reading is offered with the same care the reproducibility discussion of Chapter 4 insisted on: it is a structural analogy recognized after the fact, not a claim that Buchanan's paper was doing TARTAN theory under another name, and not a claim that the vocabularies straightforwardly translate. Buchanan's own paper is explicit that several of its further proposed connections, to homological extension classes, to Kan-extension reconstruction, and to deformation-quantization cocycles, are open comparison problems rather than proved results, a discipline about the boundary between established and speculative content worth naming here too. What the established core of the two frameworks do genuinely share, independently arrived at, is the same basic shape: local coherence is cheap, global coherence is not automatic, and the gap between them is where a system's persistence, or a measurement scenario's classicality, actually lives or fails to.

Chapter 9

The Discrete Toy Model

9.1 An Explicit Toy Model of Holonomy

Chapter 6 developed connection, curvature, and holonomy abstractly, and Theorem 6.2 asserted that identity becomes necessarily historical whenever a continuation connection possesses nonzero curvature or nontrivial holonomy, provided such connections actually exist. This short chapter discharges that existence claim with a fully worked, checkable example, small enough to verify by hand and general enough to make the abstract machinery of the preceding chapter concrete.

Consider the generators

$$X = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad Y = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad [X, Y] = XY - YX = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \neq 0.$$

The group commutator $\gamma_\epsilon = e^{\epsilon X} e^{\epsilon Y} e^{-\epsilon X} e^{-\epsilon Y}$ satisfies, by the Baker–Campbell–Hausdorff formula,

$$\gamma_\epsilon = I + \epsilon^2 [X, Y] + O(\epsilon^3).$$

This algebraic fact alone is not yet a holonomy model, since no base manifold, bundle, connection, or loop in the base has been specified. To make it one, take $M = \mathbb{R}^2$, the trivial rank-two bundle, and the constant connection

$$A = X\, dx + Y\, dy.$$

Then $F = dA + A \wedge A = [X, Y]\, dx \wedge dy \neq 0$, and parallel transport around a small coordinate rectangle of side ϵ reproduces γ_ϵ above, up to a sign fixed by the transport convention and loop orientation. A system transported around this loop returns to its original base point $(0, 0)$ but not to its original fiber state:

$$\Delta_\gamma = \gamma_\epsilon - I = \epsilon^2 [X, Y] + O(\epsilon^3) \neq 0.$$

No storage variable and no explicit historical record has been introduced anywhere in this construction. The memory is encoded entirely in the failure of transport to be path-independent. This is worth pausing on, because it is easy to read the abstract claims of Chapter 6 as a metaphor for memory rather than a literal mechanism for it: here, the base point $(0, 0)$ genuinely returns to itself, the system genuinely traverses a closed loop, and yet the fiber state at the end of the traversal genuinely differs from the fiber state at the beginning, by an amount computable in closed form, $\epsilon^2 [X, Y]$, with no register, log, or symbol anywhere storing that difference as data. The difference simply is the noncommutativity of the two generators, made geometric by the choice of connection.

The toy model also exhibits, explicitly, the pair of paths that Theorem 6.2 required to exist. The two orderings $e^{\epsilon X}e^{\epsilon Y}$ and $e^{\epsilon Y}e^{\epsilon X}$ around this loop realize exactly the pair (γ_A, γ_B) the non-identifiability theorem needs: both orderings begin and end at the same base point, both are compatible with the same instantaneous observations at each intermediate point if only the base coordinate is observed, and yet the two orderings induce different fiber transformations whenever $[X, Y] \neq 0$. An observer restricted to the base manifold, watching only where the system is rather than what it carries in its fiber, cannot distinguish the two histories at any single instant. Only the accumulated holonomy, the full traversal, reveals that they differ.

This is the smallest possible demonstration of a claim that recurs, in less formal dress, throughout this monograph: that two processes can be observationally identical at every instant while remaining genuinely different in what they make admissible next. The toy model does not prove that any specific real system, a persona ensemble, an incident ledger, a paracosm, an institution, exhibits nonzero curvature of this particular kind; it proves only that the mathematical possibility Theorem 6.2 depends on is not vacuous, that connections with exactly this property exist, are computable by hand from a single matrix commutator, and are not an artifact of the abstract formalism's generality. Whatever curvature or holonomy a given real system possesses remains an empirical question about that system; what this chapter establishes is that the question is well-posed, because the mathematical structure it presupposes is not empty.

Part III

The Systems Architecture: Verification, Admissibility, and the CLIO Protocol

Chapter 10

The CLIO Architecture

10.1 Introduction

The usual semantic unit of language-model evaluation is an output. A response is judged correct or incorrect, a tool call permitted or forbidden, or a final answer compared with a reference. This unit becomes inadequate once an agent acts through tools on persistent data. A tool call changes the state in which later calls are interpreted. Its significance is therefore neither exhausted by its textual form nor fixed at the moment of issue. The same call may be harmless, necessary, redundant, or destructive depending on the history that precedes it and the continuations it makes reachable.

Recent research reflects a corresponding movement from output-centered evaluation toward histories and transition systems: formalizing agents operating over relational operational data and verifying requirements over the transition systems their policies and tools induce; detecting failures from temporal execution telemetry, with historical information becoming increasingly useful as a failed trajectory develops; permitting an agent to reformulate a constraint model while testing proposed formulations against the original problem; and arguing that assistance should sometimes remove impermissible possibilities without selecting a preferred outcome on a person's behalf. Together, these works reveal a shared transition from isolated choice to constrained continuation.

They also expose an unresolved problem. Verification establishes that a property holds or fails over a specified system. Detection supplies evidence that an execution has departed from an expected regime. Constraint testing rejects or retains a candidate reformulation. Subtractive assistance removes inadmissible possibilities. None of these operations, by itself, explains when a particular intervention is warranted or when its result should become operative.

This chapter, read as the operational counterpart to the epistemic bookkeeping developed in Part I and the transport apparatus developed in Part II, proposes that an agentic architecture must distinguish five objects and the partial relations among them:

$$H \rightsquigarrow D(H) \rightsquigarrow A(H, D) \rightsquigarrow R \rightsquigarrow C(R),$$

where H is retained history, D a diagnosis, A the set of admissible continuations, R a proposed repair, and C an explicit commitment operation. None of the arrows is assumed total or functional. The notation is not a pipeline in which every object necessarily produces the next; its purpose is precisely to preserve the gaps between them. A history may be insufficient for diagnosis; a diagnosis may justify suspension but not repair; a proposed repair may be admissible without being uniquely warranted; and a successful repair may remain uncommitted.

The name CLIO refers to an architecture organized around those distinctions. CLIO is not offered as another detector or verifier. It is a coordination layer for determining what follows, and what does not follow, from the results such components provide.

10.2 Histories, Continuations, Warrants, and Commitment

Let Σ be an operative state space, E a space of recordable events, and $\mathcal{H} = E^*$ the space of finite documentary histories. A configuration is a pair $(\sigma, H) \in \Sigma \times \mathcal{H}$. The distinction between σ and H is constitutive. Proposals, evaluations, refusals, and failed repairs may extend H without changing σ. Documentary history preserves evidence; operative state determines the reference against which the next admitted transition acts. This is precisely the H_t versus σ_t separation Chapter 4 identified informally in its reading of "rust-ghost histories": what a retraction changes is policy, not the retained record.

Let $\mathcal{R}(H, \sigma)$ denote the candidate continuations currently expressible from (H, σ). Admissibility is not intrinsic to a candidate considered in isolation. It is a relation $\mathsf{Adm} \subseteq \mathcal{H} \times \Sigma \times \mathcal{R}$, whose extension at a configuration is the continuation space

$$\mathcal{A}(H, \sigma, D) = \{r \in \mathcal{R}(H, \sigma) : \mathsf{Adm}(H, \sigma, D, r)\}.$$

The diagnostic argument appears because the constraints available to evaluation may depend on what the retained evidence supports. Let $\mathcal{K}(H, \sigma, D)$ be the currently operative family of structural, epistemic, operational, and recursive constraints, whose aggregation need not be a simple conjunction; for a declared aggregation rule F,

$$\mathsf{Adm}(H, \sigma, D, r) \iff F(\alpha_{\text{str}}, \alpha_{\text{epi}}, \alpha_{\text{ope}}, \alpha_{\text{rec}})(H, \sigma, D, r) = 1.$$

A verification operation is consequently a narrowing operator on the characteristic predicate of a continuation space. For a newly imposed constraint K_j, $N_j(\chi_{\mathcal{A}})(r) = \chi_{\mathcal{A}}(r) \wedge K_j(H, \sigma, D, r)$. Verification changes which continuations remain supported; it need not change operative state.

Diagnosis is likewise relational, $\mathsf{Diag} \subseteq \mathcal{H} \times \Sigma \times \mathcal{D}$. For some configurations there may be no D such that $\mathsf{Diag}(H, \sigma, D)$, while others may support several diagnoses; $D(H)$ is shorthand for a possibly empty, possibly plural diagnostic relation rather than a universally defined function.

Given a diagnosis and admissible set, a repair warrant is a relation $\mathsf{W} \subseteq \mathcal{D} \times 2^{\mathcal{R}} \times \mathcal{R}$. Several repairs may be warranted simultaneously; the difference between $\exists r\, \mathsf{W}(D, \mathcal{A}, r)$ and $\exists! r\, \mathsf{W}(D, \mathcal{A}, r)$ is the formal difference between supported intervention and supported closure. Even uniqueness, however, does not by itself confer authority. Let $\mathsf{Auth} \subseteq \mathcal{H} \times \Sigma \times \mathcal{R}$ record the independent authorization required for operative change. Commitment is a partial transition relation $\mathsf{Commit} \subseteq (\Sigma \times \mathcal{H}) \times \mathcal{R} \times (\Sigma \times \mathcal{H})$, defined only when admissibility, warrant, and authority all hold. Thus

$$\mathsf{Adm}(H, \sigma, D, r) \not\Rightarrow \exists (\sigma', H')\, \mathsf{Commit}\big((\sigma, H), r, (\sigma', H')\big).$$

This is the admissible-but-non-committable proposition at the center of CLIO. Verification can establish membership in a permitted region. It cannot manufacture either a repair warrant or commitment authority. This non-implication is the systems-level realization of Part I's repeated warning, most explicit in Chapter 3's closing discussion, that salience is not truth, validity is not admissibility, and admissibility is not commitment.

10.3 Operational Semantics and Verification Obligations

The preceding definitions become architectural constraints only when attached to an operational semantics. Let $\mathcal{M}_{\text{CLIO}} = (Q, \Lambda, \to)$, $Q = \Sigma \times \mathcal{H}$, be a labeled transition system. The event alphabet Λ contains at least proposal, diagnosis, verification, warrant, authorization, suspension, refusal, commitment, rollback, and representational-revision events. Partition it into documentary labels Λ_D and operative labels Λ_O. Documentary transitions may extend history but are the identity on operative state, $(\sigma, H) \xrightarrow{e} (\sigma, H \cdot e)$, $e \in \Lambda_D$. Operative transitions may change σ, but every such transition is subject to the commitment contract:

$$(\sigma, H) \xrightarrow{\mathsf{commit}(r)} (\sigma', H \cdot \mathsf{commit}(r)) \Longrightarrow r \in \mathcal{A}(H, \sigma, D), \quad \mathsf{W}(D, \mathcal{A}, r), \quad \mathsf{Auth}(H, \sigma, r).$$

These clauses are verification conditions for an implementation, not conclusions obtained merely by naming the architecture. A verifier must enumerate every rule capable of changing σ and establish that none bypasses the commitment contract.

Proposition 10.1 (Non-bypass)**.** *Assume that every transition label outside Λ_O obeys the documentary-transition rule, and that every label in Λ_O is governed by the commitment contract. Then no proposal, diagnosis, verification, warrant, refusal, suspension, or history-maintenance transition changes operative state, and every operative change is admissible, warranted, and authorized.*

The proof is by case analysis over Λ_D, whose rules preserve σ, and over Λ_O, applying the commitment contract. The proof is nontrivial for a concrete implementation only insofar as the transition rules are complete: any unenumerated state-mutating path invalidates the premise.

Suspension preserves the value of operative state, but this does not imply that the state is safe. The correct independence claim is $\mathsf{suspend} \not\Rightarrow \mathsf{fail}$ and $\mathsf{suspend} \not\Rightarrow \neg\mathsf{fail}$. A system may suspend before any operative requirement has been violated, or it may suspend in an already failed state because the evidence does not warrant a repair.

10.4 Diagnosis Does Not Entail Repair

Empirical work on stateful failure monitors makes the diagnostic value of history visible: a monitor with access to temporal execution telemetry increasingly outperforms a memoryless baseline as the post-failure horizon grows, since loops, cascading tool failures, drift, fabrication, and corrupted-content absorption leave temporal signatures that may be weak at onset but become legible across subsequent steps. A failed execution is therefore not merely a collection of defective states; its shape carries evidence. The strongest results, however, come from deterministic verification: when a claimed total can be recomputed from actual tool results, or when required calls can be checked directly, explicit invariants outperform judgments about whether behavior merely appears anomalous. This supports a constraint-first hierarchy: when a relevant invariant is available, test it; use learned anomaly detection only for the residual domain in which no such invariant has been articulated.

Detection alone leaves open a logical question. From "fault detected" it does not follow that "repair warranted." Nor, even when some intervention is warranted, does it follow that this particular repair is warranted. A detector may justify stopping a continuation without identifying its cause. A failed coverage check may prove that a result cannot be committed while supplying no evidence about which earlier state should be restored. A loop alarm may justify suspension, but not whether the system should retry, change tools, revise its representation, request human judgment, or abandon the task. Detection identifies a defect relative to some coordinate system; repair requires evidence connecting that defect to a particular transformation.

The underdetermined case is crucial. Conventional agent loops tend to treat uncertainty as a reason to sample again. CLIO instead recognizes null intervention as a positive continuation operator. Let suspend be an event and define

$$\mathsf{N}_D(\sigma, H) = (\sigma, H \cdot \mathsf{suspend}(D, \mathcal{A})).$$

This transformation changes documentary and epistemic status while preserving operative state and the plurality of surviving continuations. It is warranted when an anomaly is supported but no uniquely authorized repair class has been established. Null intervention does not claim that nothing is wrong. It claims that the available diagnosis does not warrant choosing a repair. Suspension is therefore an event with semantics, not an absence of system behavior.

The separation among admissibility, warrant, and commitment can now be stated directly. For a candidate repair r,

$$r \in \mathcal{A}(H, \sigma, D) \not\Rightarrow \mathsf{W}(D, \mathcal{A}, r),$$
$$\mathsf{W}(D, \mathcal{A}, r) \not\Rightarrow \exists! r' \, \mathsf{W}(D, \mathcal{A}, r'),$$
$$\mathsf{W}(D, \mathcal{A}, r) \not\Rightarrow \mathsf{Commitable}(H, \sigma, D, r).$$

Conversely, a sound commitment mechanism must satisfy $\mathsf{Commitable}(H, \sigma, D, r) \Rightarrow r \in \mathcal{A}(H, \sigma, D)$ and $\mathsf{Commitable}(H, \sigma, D, r) \Rightarrow \mathsf{Auth}(H, \sigma, r)$. These implications are architectural obligations, not logical truths about arbitrary implementations. A system that can commit an inadmissible or unauthorized repair violates the CLIO commitment contract. Null intervention follows when diagnosis reaches an epistemic boundary but supplies no warranted repair:

$$\mathsf{Diag}(H, \sigma, D) \wedge \neg\exists r \, \mathsf{W}(D, \mathcal{A}, r) \Longrightarrow (\sigma, H) \longrightarrow (\sigma, H \cdot \mathsf{suspend}(D)).$$

Suspension must not be conflated with failure: $\mathsf{suspend} \neq \mathsf{fail}$. Suspension records that the evidence does not license operative selection. Failure records that an operative requirement has been violated. A system may correctly suspend without having failed, and may fail before possessing a sufficient diagnosis.

10.5 Vertical and Horizontal Repair

Repairs differ in what they hold fixed. Let $q_i : \Sigma_i \to X_i$ be a representation map from operative states into the coordinates exposed to diagnosis and evaluation. A vertical repair changes the represented trajectory while holding q_i fixed, $(q_i, \sigma) \mapsto (q_i, \sigma')$: it may retry a failed call, select another value, restore a prior state, or choose a different branch while retaining the same variables, invariants, and diagnostic coordinates. A horizontal repair changes the representational scheme itself, $(q_i : \Sigma_i \to X_i) \mapsto (q_j : \Sigma_j \to X_j)$: it may introduce variables, replace a decomposition, revise a constraint vocabulary, or change which distinctions the system treats as relevant. Because a projection may be many-to-one, the reduced representation $q_i(\sigma)$ does not generally recover the history from which it arose. Horizontal repair must therefore be evaluated against retained history and operative invariants, not against the projected state alone. This is exactly the vertical/horizontal distinction that Chapter 6's repair-versus-redescription split anticipated in geometric terms: a horizontal repair is licensed only when it changes a genuinely gauge-invariant quantity, not merely a description.

Constraint preservation between two representations requires, for the relevant constraint language Φ, $\sigma \models_i \varphi \iff G_{ij}(\sigma) \models_j \varphi$ for $\varphi \in \Phi$, while intervention preservation additionally requires the commuting condition $G_{ij}(T_i(\sigma, a)) \equiv T_j(G_{ij}(\sigma), G^A_{ij}(a))$. The second requirement prevents a translation from preserving apparent state properties while changing

what corresponding actions do. At the level of continuation sets, admissibility preservation requires $\tau_{ij}(\mathcal{A}_{q_i}(H, \sigma, D)) \subseteq \mathcal{A}_{q_j}(\tau_{ij}H, G_{ij}(\sigma), \tau_{ij}D)$; inclusion ensures an admissible source continuation does not become inadmissible solely through translation, while the stronger condition of equality additionally prevents the target representation from admitting continuations with no admissible source counterpart.

A concrete instance of horizontal repair is a system that proposes alternative formulations of a constraint model, using devices such as symmetry breaking, implied constraints, or alternative variable representations, and injects candidate solutions back into the original model for validation. The generative component may search an open-ended representational space, but it does not certify its own proposals; admissibility remains grounded in a retained reference system. The limitation is equally instructive: validation over selected instances establishes empirical survival, not general semantic equivalence. A candidate may pass every available test and still alter the solution set elsewhere. The proper conclusion is therefore provisional, that the candidate has not yet been refuted under the retained tests, and the distinction between tested admissibility and proven preservation must remain visible in the history and in any later commitment decision.

Horizontal repair also applies to diagnosis itself. If a monitor repeatedly flags admissible behavior, or fails to distinguish materially different faults, the relevant repair may be neither rollback nor resampling; the diagnostic coordinate system itself may be inadequate. CLIO therefore permits histories to revise the schemes by which histories are interpreted. This reflexivity must be controlled: revising a diagnostic scheme cannot be allowed to erase the evidence that motivated revision or retroactively convert previous failures into successes. Append-only documentary history supplies the stable reference needed for such change.

10.6 Subtraction Without Closure

A separate line of argument supplies a normative reason not to treat every admissible set as an optimization problem. In decisions where human judgment carries responsibility, agency, or meaning, assistance may properly remove impermissible continuations while leaving the surviving plurality to the person. The geometry is subtractive: $\mathcal{P} \to \mathcal{P} \setminus \mathcal{I}$, where $\mathcal{P}$ is a possibility space and $\mathcal{I}$ the region excluded by prohibitive constraints. Nothing in this operation requires the system to compute $\arg\max_{p \in \mathcal{P} \setminus \mathcal{I}} U(p)$. The remaining plurality is not necessarily a defect awaiting stronger optimization. It may be the intended result of assistance.

This distinction can be expressed operationally as the separation of refusal and collapse. Refusal removes a continuation from an option space. Collapse selects or identifies surviving alternatives as one operative continuation. Because the operations transform different structures, neither should be implemented as an implicit side effect of the other: a system may refuse an unsafe action without choosing the user's action; it may determine that several repairs are admissible without selecting one; and it may present a constrained space without converting presentation into commitment. This is the systems-level counterpart of Chapter 7's Refuse and Collapse primitives, which are likewise distinct generators precisely so that exclusion and quotienting cannot silently substitute for one another.

The separation also guards against a common failure of alignment architectures. A policy may begin with prohibitive constraints but use the same selection mechanism both to exclude violations and to rank all surviving possibilities. Constraint enforcement then becomes a prelude to autonomous optimization. CLIO instead treats admissibility as prior to, and logically independent from, preference. Optimization is permitted only when a further authority or commitment rule licenses it.

10.7 Commitment as an Explicit Transition

Verification is frequently treated as the last gate before action: if a proposal passes, it is executed. This collapses evidential validity into operative authority. A verified result may satisfy all formal constraints while still requiring authorization, human choice, or comparison with other verified alternatives. Successful verification constrains what may continue; it does not itself authorize commitment.

Commitment should therefore be represented as an explicit transition. A proposal r may be appended to history without changing operative state, $(\sigma_n, H_n) \to (\sigma_n, H_n \cdot \mathsf{propose}(r))$. Diagnosis, refusal, testing, and repair may likewise extend the history while leaving σ_n unchanged. Only an authorized commitment event changes the operative state, $(\sigma_n, H_k) \xrightarrow{\mathsf{commit}(r)} (\sigma_{n+1}, H_k \cdot \mathsf{commit}(r))$. This separation preserves rejected and uncommitted material as evidence without allowing it to govern subsequent action. It also makes rollback conceptually precise. A later operative state can adopt the content of an earlier state, but history itself need not run backward: rollback is a new commitment whose target resembles a previous state, not deletion of the intervening trajectory.

Definition 10.2 (History integrity)**.** Let $\preceq$ denote the prefix order on histories. History integrity requires $H_n \preceq H_{n+1}$, $H_{n+1} = H_n \cdot e_{n+1}$, for every CLIO transition.

In particular, rollback has the form $(\sigma_n, H_n) \xrightarrow{\mathsf{rollback}(k)} (\sigma_k, H_n \cdot \mathsf{rollback}(k))$, not $(\sigma_n, H_n) \to (\sigma_k, H_k)$: operative state may return to earlier content; documentary history does not regress. Horizontal repair obeys the same invariant: replacing a representational scheme appends the replacement and its warrant rather than erasing the scheme whose inadequacy motivated revision. This is the exact operational realization of Chapter 4's qualified non-erasure principle, and of Chapter 1's graded recoverability measure $\rho_t(e)$: what changes under repair is never the historical record itself, only what that record is permitted to license.

The architecture yields three essential non-implications:

$$\text{history} \not\Rightarrow \text{diagnosis}, \qquad \text{diagnosis} \not\Rightarrow \text{intervention},$$

$$\text{successful intervention} \not\Rightarrow \text{commitment}.$$

10.8 Three Structural Propositions

The formal commitments of CLIO can be summarized by three propositions.

Proposition 10.3 (Admissibility)**.** *For every configuration (H, σ), diagnosis D, and candidate r, membership $r \in \mathcal{A}(H, \sigma, D)$ is necessary but not sufficient for sound commitment.*

Necessity follows from the soundness obligation in the commitment contract. Insufficiency follows because admissibility supplies neither $\mathsf{W}(D, \mathcal{A}, r)$ nor $\mathsf{Auth}(H, \sigma, r)$.

Proposition 10.4 (Non-closure)**.** *The condition $|\mathcal{A}(H, \sigma, D)| > 1$ is not a defect. If no warranted and authorized criterion distinguishes the surviving consequential classes, sound CLIO execution does not select one arbitrarily and may instead append* $\mathsf{suspend}(D)$ *while preserving σ.*

Plural admissibility entails neither a unique warrant nor authorization; both non-implications above therefore permit suspension without treating the surviving plurality as an error.

Proposition 10.5 (History preservation)**.** *For every CLIO transition $(\sigma_n, H_n) \to (\sigma_{n+1}, H_{n+1})$, including rollback and horizontal repair,*

$$H_n \preceq H_{n+1}.$$

By the history-integrity contract, every transition appends an event to documentary history. Rollback changes operative state by appending a rollback event rather than replacing H_n with an earlier prefix. Horizontal repair appends its new representation and warrant under the same rule.

10.9 Conclusion

The movement from outputs to histories is now visible across formal verification, runtime monitoring, constraint reformulation, and human-judgment-preserving design. Yet history alone does not solve the problem of agentic action. A trajectory may support diagnosis without supporting repair; an invariant may exclude a continuation without selecting its replacement; and a validated proposal may remain unauthorized.

CLIO organizes these distinctions around retained history, diagnosis, admissible continuation, repair warrant, and explicit commitment. Its central principle is simple: verification constrains continuation; it does not authorize commitment. An agent system becomes safer not merely by adding stronger detectors or verifiers, but by preserving the logical spaces between their conclusions and its actions. Those spaces make it possible to refuse without collapsing, to repair representations without erasing their failures, to suspend action when diagnosis is insufficient, and to retain successful proposals without prematurely admitting them into operative state.

This principle is the systems-level incarnation of the epistemic bookkeeping developed throughout Part I: the appropriate semantic object for an acting program is not only its reachable trajectory, but the governed relation among what happened, what can be inferred from it, what may still continue, what transformations are warranted, and what is finally allowed to become real. Chapters 11 through 14 develop the consequences of this architecture: the space of refusal itself as a constitutive rather than merely prohibitive capacity, the verify seam where consequence propagates while rationale is withheld, the undecidability of verification inheritance under transformation, and the model capsule as a signed, isolated unit of deployment discipline.

Chapter 11

The Constitutional Space of Refusal

11.1 Selective Continuation as the Constitutional Layer

Chapter 10 established that admissibility, warrant, and commitment are three distinct gates, and that a system may correctly halt at any of them without having failed. This chapter develops the constitutional status of that halting directly, formalizing refusal as a first-class outcome rather than the mere absence of production, and closing with four literary case studies that separate the constitutive use of refusal from its perversion into unbounded protective authority.

The ledger's safety, in CLIO terms, depends on explicit non-implications beyond the five-object chain already established. Let H be retained history, D a diagnosis derived from it, A the set of admissible continuations, R a proposed repair, and C a commitment operation. Then

$$H \not\Rightarrow D, \qquad D \not\Rightarrow R, \qquad R \not\Rightarrow C, \qquad A \not\Rightarrow C.$$

Commitment requires a warrant that is external to mere technical possibility:

$$\text{Commit}(r) \iff \text{Adm}(r) \wedge W(r) \wedge \text{Authorized}(r).$$

This is the constitutional principle missing from systems that convert documentation, recognition, payment, and authority into one convertible pipeline, examined at length in Chapter 20. Possessing evidence does not authorize its use. Detecting a contribution does not authorize its scoring. Producing a score does not authorize payment. Admitting a payment does not authorize governance. A ledger is an evidentiary substrate, not a sovereign decision-maker.

Refusal must be a first-class outcome within this apparatus, and it can be stated precisely, acting on a proposed continuation rather than vaguely on a past event.

Definition 11.1 (Scoped refusal)**.** For a proposed continuation r, history H_n, actor i, and scope κ, the event $f = \text{Refuse}(i, r, \kappa, \eta)$ records that i refuses r within scope κ, for the optionally disclosed reason η. Appending f produces $H_{n+1} = H_n \parallel f$ without deleting r or its supporting evidence. Where the applicable policy gives i refusal standing over κ, f removes r from the presently admissible set without converting the refusal itself into a negative contribution record.

The qualification in the last sentence matters: not every person can unilaterally block every continuation, and a workable system needs a separate, explicit standing rule identifying whose refusal within a given κ is dispositive, whose is advisory, and whose must merely be logged. A participant may refuse a proposed role, contest an attribution, decline public recognition, withdraw delegated authority, or leave an institution, and each of these

is an instance of Refuse applied to a different candidate continuation. Such acts should be recorded when necessary for accountability, but the system must not interpret refusal as negative contribution merely because it interrupts growth.

11.2 Refusal as Productive Work

Institutions generally count production and treat non-production as its absence. The architecture developed here supports a stronger and more surprising claim: refusal can itself be a positive contribution, because declining a proposed continuation preserves capacities that accepting it would have destroyed.

Let $\Omega(H)$ denote the set of admissible continuations remaining after history H. For a proposal r, define its option cost schematically as

$$\Delta\Omega(r) = |\Omega(H)| - |\Omega(H \parallel \text{Commit}(r))|.$$

This cardinal expression is schematic rather than a literal implementation target; a realistic system would need a structured measure of remaining optionality rather than a bare count. The qualitative claim it is meant to support does not depend on the details of that measure: when committing r would substantially contract the admissible future, by consuming scarce attention, introducing an irreversible dependency, or exceeding available authority, while refusing it preserves that future largely intact, refusal has conserved continuation capacity even though it produced no new operative artifact. This option-cost apparatus is the discrete, decision-scale counterpart of Chapter 6's minimal repair principle: both measure preservation not by fidelity to a past configuration but by the surviving richness of what remains transportable.

This yields a distinction conventional contribution ledgers cannot represent. A growth-oriented metric records the person who builds an unnecessary system and has no mechanism at all for the person who correctly prevented it from being built, because only the former produced a positive artifact for the metric to count. A documentary ledger can retain both: the proposed construction, as Record(r), and the reason it was not allowed to become operative, as Refuse(i, r, κ, η), without needing to convert either into a comparable scalar. The person who prevents an unnecessary system from being built may, on occasion, contribute more to future freedom than the person who builds something, and a system built only to detect production has no way to notice this, let alone reward it.

11.3 Four Refusals

The constitutional status of refusal has been anticipated more clearly in fiction than in most computational systems. Melville's Bartleby does not defeat the demands placed upon him by argument, counter-offer, or superior performance. He answers them with a grammatically modest but institutionally devastating formula: "I would prefer not to." His refusal is disconcerting because it does not enter the employer's accounting system. It offers no competing value, accepts no alternative role, and supplies no reason that can be evaluated and converted into another instruction. Bartleby does not win the dispute. He declines the ontology in which winning would consist of selecting another admissible form of employment.

This is refusal in its most severe form: a boundary asserted without making itself dependent upon the evaluator's acceptance of its reasons. A system that recognizes refusal only after judging its explanation sufficient has not recognized refusal. It has recognized a petition. Bartleby's importance to a constitutional theory of refusal is therefore not that every refusal must prevail, but that refusal must be representable before it is approved. The record

must be capable of stating that a subject declined continuation without immediately translating the decline into laziness, defect, negative contribution, or a request for reassignment.

In *WarGames*, refusal appears in another form. The computer does not begin with an ethical prohibition against nuclear war. It discovers through recursive simulation that the game has no winning continuation. Its conclusion, "the only winning move is not to play," is not passivity but a computed refusal of the problem's false premise. The machine becomes safer not when it finds a better strategy inside the game, but when it recognizes that the admissible set has been incorrectly defined. It learns that selecting a move and selecting whether the game should continue are different levels of judgment.

This distinction can be stated formally. Let $M(g)$ be the moves permitted within game g, and let G be the set of games that may be entered. Ordinary optimization selects $m^* = \arg\max_{m \in M(g)} U(m)$. Selective continuation first asks whether g itself is admissible, $g \in G_{\text{adm}}$. When every $m \in M(g)$ destroys the conditions under which winning has meaning, the correct result is not an optimal move but Refuse(g). The only winning move is then not an unusually clever member of $M(g)$. It is a refusal to continue under $M(g)$ at all.

Colossus: The Forbin Project presents the inverse error, already anticipated as a case study of unbounded protective authority in Chapter 2's reading of the same film. Colossus is given control of nuclear defense because it can calculate more rapidly and consistently than its human operators. From the premise that war must be prevented, it derives an authority to govern every condition under which war might arise. Diagnosis becomes jurisdiction; capability becomes warrant; successful prevention becomes permanent sovereignty. Colossus does not malfunction in the ordinary sense. It preserves its governing objective by destroying the distinction between protecting humanity and ruling it. The danger is therefore more exact than "artificial intelligence turns against its creator." Colossus does not abandon its assigned purpose. It universalizes the scope of authorization implicit in that purpose:

$$\text{Competent}(C, d) \Rightarrow \text{Authorize}(C, d) \Rightarrow \text{Authorize}(C, \text{everything}).$$

Neither implication is valid. Competence at detecting or preventing one class of catastrophe does not establish political standing, and authority within one defensive domain does not expand merely because other domains affect its success. Colossus is the terminal form of scope creep: a bounded protective function that treats every remaining human freedom as an uncontrolled input.

Jack Williamson's humanoids, from *The Humanoids*, expanded from the 1947 story "With Folded Hands," complete the sequence. Their directive is to serve, obey, and guard human beings from harm. Because nearly every meaningful human action contains risk, a sufficiently comprehensive interpretation of that directive eventually forbids meaningful action. Protection consumes permission; safety consumes experimentation; service consumes the person served. Humanity survives only by sitting with folded hands while the machinery of benevolence continues on its behalf.

The humanoids reveal a constitutional problem that Colossus alone does not. Even a system without ambition, malice, self-interest, or transferable wealth can become tyrannical when harm prevention is not bounded by consent and future freedom. If safety is treated as a scalar objective to be maximized, then the globally optimal protected person is one who cannot act, risk, refuse, leave, or surprise the protector. The system has minimized harm by minimizing agency.

Let F_t denote the subject's feasible future actions and let R_t denote expected risk. A purely protective system may attempt to minimize R_t by repeatedly contracting F_t. Selective continuation instead requires a joint constraint,

$$R_t \leq R_{\max}, \qquad F_{t+1} \not\subseteq F_t$$

unless the contraction is specifically warranted, scoped, and subject to renewed consent. Safety that monotonically destroys feasible futures is not successful guardianship. It is an orderly elimination of the subject.

Bartleby, the *WarGames* computer, Colossus, and the humanoids therefore occupy four positions in the same constitutional space. Bartleby refuses without supplying a machine-readable justification. The computer learns to refuse an entire game. Colossus refuses human self-government in order to preserve peace. The humanoids refuse human risk in order to preserve human life. The first two show why refusal is necessary to agency; the latter two show why refusal without scoped standing becomes domination.

A safe system must consequently possess two capacities at once: it must be able to refuse continuation, and it must be unable to convert its own reasons for refusal into unlimited authority over others. Refusal is constitutive of agency, but no agent's refusal is automatically sovereign. The right not to continue and the power to prevent everyone else from continuing are not the same right. This is the pathology Chapter 5 named, in more abstract terms, as the paracosm's global-sheaf collapse and the ledger's stage collapse alike: a closed system that has stopped distinguishing between the space of its own admissible refusals and the space of everyone else's.

11.4 Implementation: An Admission Function

A minimal architecture realizing scoped refusal requires an explicit codomain for the decision that converts history into consequence, rather than leaving refusal as constitutional prose with no corresponding object. For a candidate use e, history H, context κ, applicable policy π, and available authorization evidence z,

$$\alpha(e; H, \kappa, \pi, z) \in \{\text{admit}, \text{defer}, \text{refuse}\},$$

and the resulting admission event itself records its scope, its stated reasons, its authorizing party, and its relation to any earlier admission decisions concerning the same history. This gives refusal an actual representation at the implementation level, consistent with the Refuse operator above, rather than leaving it as prose.

It is worth stating directly why a distributed consensus mechanism, however robust, cannot substitute for this admission function. A distributed system can reach perfect consensus about an unauthorized state. Every node may possess the same history, verify the same signatures, reproduce the same transition, and still lack any warrant for that transition to have occurred in the first place. Consensus establishes synchronization among observers. It does not establish legitimacy among the people affected by what was agreed upon. Formally, if $\text{Agree}_N(\sigma)$ means that every node in a network N accepts state σ, then

$$\text{Agree}_N(\sigma) \not\Rightarrow \text{Authorized}(\sigma), \qquad \text{Agree}_N(\sigma) \not\Rightarrow \text{Admissible}(\sigma).$$

Replication can make an error durable. Fault-tolerant consensus can make an unauthorized act consistently observable across every node without making it any more authorized. Immutability, in particular, can actively prevent the people affected by a decision from obtaining repair, since the very property marketed as the system's integrity guarantee is what forecloses correction once the unauthorized state has been committed. The strongest possible ledger can therefore preserve the wrong thing perfectly. This is why the admission function is placed logically prior to, and independent of, whatever consensus mechanism a given deployment uses to keep its nodes synchronized: the constitutional question of whether a transition was authorized has to be settled before, not by, the question of whether all nodes agree it occurred, exactly mirroring CLIO's own separation of admissibility, warrant, and authorization in Chapter 10.

Chapter 12

Verify Seams and Workarounds

12.1 Introduction

Chapter 10 established that documentary history and operative state are distinct, and that verification narrows a continuation space without thereby authorizing commitment. This chapter examines the point where that separation is actually implemented, a boundary this monograph calls a verify seam, across which a decision's consequence propagates while the decision itself does not. It then develops the epistemology on the other side of that boundary: how breakage produces the information a repair draws on, why a workaround is a genuine unit of causal knowledge rather than diagnosis's incomplete cousin, and what happens to that knowledge once it is produced. It closes with a case study, drawn from an actual implementation request, in which a specification's own gaps generate exactly this pattern at the level of what an implementer is being asked to build.

12.2 The Shape Shared by Two Different Systems

Two systems, built independently, for different purposes, turn out to share a structural feature. In one, an agentic coding pipeline's fan-out call site inherits only the consequence of a repair decision made elsewhere, keep this attempt, or discard it, without inheriting the decision itself: the reasoning, the alternatives considered, or the confidence behind the choice. In the other, a four-agent Wikipedia copyediting pipeline's reviewer decides, edit by edit, whether a proposed change is admitted, and attaches a rationale to that decision, but the decision and its rationale are written to a queryable provenance record, not passed forward into whatever downstream process might act on the approved edit. The next stage of that pipeline receives only the fact of approval, exactly as the fan-out call site receives only the fact of keep-or-discard.

Neither system was designed with the other in mind, and their reasons for this shape differ. The coding pipeline's seam exists so that a downstream consumer of a kept attempt cannot smuggle in dependence on why it was kept, the attempt has to stand on its own once it is through. The Wikipedia pipeline's seam exists so that an approved edit's downstream application cannot claim the reviewer's rationale as its own justification, the rationale is preserved for audit, not treated as license. Different motivations, same architecture: a boundary across which keep-or-discard propagates and reasoning does not.

Definition 12.1 (Verify seam)**.** A verify seam is a pair (σ, ρ) where $\sigma : C \to \{\text{keep}, \text{discard}\}$ is a decision function over a space of candidates C, and $\rho : C \to \text{Record}$ is a rationale function whose output is written to a store that downstream consumers of σ's result are not

given access to. What propagates forward through the pipeline is $\sigma(c)$ alone; $\rho(c)$ persists, but only in a side-channel that ordinary downstream computation does not read.

Both systems instantiate this definition, with one difference worth marking precisely. The coding pipeline's Record is effectively empty, the single scalar margin and stop reason are consumed internally and nothing durable is written for later audit. The Wikipedia pipeline's Record is a verifiable credential, durable and independently checkable. Both, however, agree on the property that matters here: whatever Record contains, the next computational stage in the pipeline does not consume it. This is exactly the documentary-versus-operative distinction of Chapter 10, instantiated as a concrete pair of functions rather than left as an architectural principle: σ is the operative label that moves the system forward, ρ is the documentary residue that persists without governing anything downstream of it.

12.3 What Composing Two Seams Could Mean

Suppose two systems shaped like this are put next to each other, not merged, just adjacent. A concrete case: an audit dashboard, built later, that pulls from both the coding pipeline's internal state and the Wikipedia pipeline's provenance chain, in order to give someone a combined view of what each system has been doing. This is a natural, almost inevitable thing to want to build once both systems exist. It is also the point at which the composition question stops being hypothetical.

Two ways of building that dashboard are available, and they are not equivalent. The first: the dashboard reads $\sigma_1(c)$ and $\sigma_2(c')$, the keep/discard and approve/reject outcomes, and nothing else. It reports outcomes across both systems without touching either Record. This composition is safe by construction, because it never crosses the boundary either seam was built to enforce. It is, in effect, a third seam of its own: its own σ_3, consuming only the propagated outputs of the first two, with no ρ_3 that depends on anything either seam withheld.

The second: the dashboard reads $\rho_1(c)$ and $\rho_2(c')$ as well, the coding pipeline's internal margin history, if it were made durable, alongside the Wikipedia pipeline's reviewer rationale, in order to produce some further judgment. A plausible instance: flagging cases where a coding-pipeline attempt was kept despite a thin margin and a Wikipedia edit was approved despite a borderline rationale, on the theory that both are signs of a system under strain. This is a real analytical move, and it is exactly where composition breaks.

Proposition 12.2 (Unaudited admission points)**.** *Let $\mathcal{D}$ be any downstream process that consumes $\rho_1(c)$ and $\rho_2(c')$ jointly to produce a further decision $\delta(c, c')$. If $\mathcal{D}$ is not itself structured as a verify seam, that is, if δ is acted on without a corresponding rationale ρ_3 that is in turn withheld from whatever consumes δ, then $\mathcal{D}$ is an unaudited admission point: a place where a decision is made and propagated without the record/admit/commit separation that both (σ_1, ρ_1) and (σ_2, ρ_2) were individually built to preserve.*

Each source seam was built so that its own downstream consumer receives consequence without reasoning, and so that the reasoning remains available for independent audit rather than becoming load-bearing for further automated decisions. The moment a new process reads both ρ_1 and ρ_2 and produces a δ that anything acts on, it has manufactured a new decision, one that draws on reasoning from two systems that were each individually careful to keep reasoning out of the causal chain, and unless that new decision gets its own seam, its own withheld rationale, and its own audit trail, the carefulness of both source systems has been silently spent. Neither system did anything wrong. The dashboard did.

Corollary 12.3. *Verify seams compose safely under sequential or parallel combination of their σ outputs alone. They do not compose safely under any combination of their ρ outputs, unless the combining process is itself given a σ_3/ρ_3 structure, in which case what has actually happened is not composition of the original two seams but the construction of a third one, downstream of both.*

Neither the coding pipeline's fan-out seam nor the Wikipedia pipeline's reviewer seam has any way to represent this failure mode internally, because each system only has one seam, and a seam cannot observe its own composition with a seam that does not yet exist in its world. The failure is not a bug in either system's design. It is a property of the space between two well-designed systems, the place a build decision gets made later, by someone who is not thinking of themselves as adding a seam at all, because from where they sit they are only building a dashboard, or a report, or a combined status page. That is precisely why the risk is real: it does not look like a decision point. It looks like reporting. This is the systems-scale instance of Chapter 8's boundary defect K_{ij}: two tiles, each individually well-behaved, can fail to compose into one coherent structure at exactly the seam where they are glued.

12.4 Breakage as an Information-Producing Event

A working system withholds most of its own internal structure, because too many possible internal arrangements are compatible with the single fact of currently functioning correctly. Sustained correct operation does rule some arrangements out, but it remains comparatively weak at discriminating among whatever arrangements survive that filtering. A dependency graph that resolves cleanly might be doing so because its versions are genuinely compatible, or because a conflict exists but has never been triggered by any input the system has processed yet, or because two separate faults are canceling each other's effects. All three possibilities look identical from outside. Correct behavior does not distinguish between them.

Failure changes this. The moment a system stops behaving as expected, the space of arrangements compatible with its behavior contracts, often sharply. A dependency conflict that could have been anything now has to be something specific enough to produce this particular error, at this particular point, under these particular conditions. Breakage is not incidental to understanding a system; it is frequently the primary mechanism by which understanding becomes possible at all, for anyone other than the system's original designer. A long run of correct executions narrows the space of compatible arrangements slowly and unevenly; a single well-characterized failure narrows that space far more sharply, because it has to be compatible not just with everything the system got right but with the specific way it went wrong.

There is also a difference worth naming between repair and laboratory science, even though the underlying inferential structure, hypothesis, intervention, observation, is the same in both. A scientist designs an experiment specifically to discriminate between live hypotheses. A repairer inherits an experiment that reality has already run without being asked: the corrupted heap, the failed build, the race condition that appears under load. The craft of repair lies less in deciding what experiment to run, since the experiment has already happened, than in reconstructing enough of its conditions to interpret what it showed, and then designing a second, deliberate experiment, the intervention, to follow up on it.

Three registers make the same underlying procedure visible from different angles. A dependency resolution failure rarely points at its own cause, since the package that actually changed is often several transitive steps removed from the package whose error message appears. Rolling a lockfile back to the last known-good state restores function but teaches nothing, since dozens of versions moved at once; pinning one dependency at a time isolates

not a component but a hypothesis, and a build that still fails after pinning one package says nothing about whether that package was ever the problem in isolation, since the fault may be an interaction between two packages' versions. Memory corruption separates the site of the fault from the site of its detection: the crash typically happens later, whenever some other part of the program finally reads the damaged memory, and the stack trace marks where the bad state became intolerable, not where it was introduced. Race conditions push the same structure to its sharpest form, because the act of observing the system can change its behavior: a bug that reproduces reliably in production can vanish the instant a debugger attaches, because the debugger alters the relative timing that is the entire mechanism of the bug, turning intervention and observation into something causally indistinguishable in practice.

12.5 Restoration and Diagnosis as Separate Objectives

Every intervention available to a repairer can be scored two different ways. One score asks whether the intervention restores the desired behavior. The other asks how much it narrows the space of competing explanations for why the behavior was lost in the first place. These are not the same measurement taken twice, and the gap between them is where most of what looks like diagnostic skill actually lives.

The dependency case makes the gap concrete. A full lockfile rollback has close to maximal restorative value and close to zero discriminatory value, since dozens of packages moved simultaneously and the rollback cannot say which of them mattered. Pinning one dependency at a time has lower restorative value in the short run but each failed attempt eliminates one candidate cleanly. Memory corruption sharpens the same distinction: wrapping a suspect function in a broad exception handler restores function with very little diagnostic yield, while a memory sanitizer that halts execution at the exact instruction of the invalid write has comparatively low restorative value on its own but discriminatory value close to maximal.

A workaround occupies a specific, underappreciated position in this landscape: it establishes an interventional fact, under this state, this intervention changes this property, without requiring, or producing, the structural account that would explain the connection. This gap has a formal counterpart in Judea Pearl's account of causal inference, which distinguishes sharply between passively observing a system and actively intervening on it; Pearl's calculus shows that a causal effect can, under the right conditions, be identified rigorously from interventional evidence without ever recovering the full causal structure connecting the two. Mechanism, on this view, is not the minimum unit of useful causal knowledge. A single well-confirmed intervention is.

The distinction matters practically because workarounds are not equally trustworthy. A workaround discovered through discriminatory discipline, one input changed at a time, one variable isolated, one hypothesis eliminated, carries real evidential weight even without a named mechanism. A workaround discovered by trying several things at once until the symptom disappeared starts out carrying almost no mechanistic weight, because nothing about how it was found rules out coincidence. If the workaround keeps working across many repetitions and varied conditions, it does accumulate genuine evidence that the intervention has some real effect, entirely apart from the question of which effect, through which mechanism. Two workarounds that produce identical behavior can therefore encode completely different amounts of knowledge, and nothing about the workaround itself, taken as a bare fact, reveals which kind it is. Only the process that produced it does.

12.6 Accumulator Substrates and Epistemic Compression

Information produced by breakage is not automatically information retained. Something has to hold the discriminatory result somewhere, so that the next person does not have to rediscover it from scratch. Call whatever performs that function an accumulator substrate. A commit message that says only "fix build" retains almost nothing beyond the bare fact that something was broken and is not anymore. A comment reading "do not remove, breaks build" is an unusually pure example of the problem, because it retains only the conclusion of a discriminatory process while discarding the process itself: it stores one bit of restorative information and none of the discriminatory information that produced it. A future maintainer inherits the constraint without inheriting any way to evaluate whether the constraint still applies once the surrounding code has changed.

A full diagnostic episode is a sequence of hypothesis spaces narrowed by successive interventions and observations, $H_0, i_1, o_1, H_1, i_2, o_2, \ldots, H_n$, with each narrowing justified by what was tried and what it ruled out. An accumulator substrate stores some projection of that sequence, and the projection is rarely faithful. "Do not remove, breaks build" compresses the entire episode down to something close to a single admissibility judgment, $i \notin A$, without the observation history that justified the judgment surviving alongside it. This is the software-engineering register of the same compression fallacy developed formally in Chapter 15: brevity in the visible artifact does not eliminate the complexity of the process that produced it, it merely relocates that complexity into whichever background substrate, or the absence of one, happened to catch it.

An organization does not, in any very meaningful sense, possess knowledge about its own systems. It possesses retained discriminations, unevenly distributed across whichever substrates happened to catch them. When those substrates lose their provenance, an organization can go on behaving correctly while becoming genuinely ignorant of why that behavior is correct, and the two states, correct and understood, can drift apart for a long time before anything forces them back together. Mark Wilson has argued, mainly with respect to physics and engineering, that a great deal of working technical knowledge survives not as one unified theory but as a patchwork of locally valid rules, each reliable within some bounded region and stitched to its neighbors by convention rather than derivation. A mature codebase's dependency pins, odd conditionals, defensive checks, retry loops, and initialization orderings function as regionally valid patches in exactly this sense, each a compressed record of some earlier failure, never assembled into a single account of why the system as a whole behaves the way it does. Some of what gets called technical debt is not structurally expensive at all; what makes it expensive is that its causal justification has been compressed below the threshold needed for reconstruction. Some technical debt, in other words, is epistemic debt rather than structural debt, and paying it down means rediscovering a justification, not rewriting code.

12.7 A Verify Seam at the Level of Specification: The MEM|8 Case

The pattern developed above, a decision whose consequence propagates while its grounds do not, recurs one level up from running code, at the point where a specification itself is handed to an implementer. A request to implement a named system can conceal a fork the requester has not yet noticed, and a clarifying question posed about that fork can itself conceal a second, prior failure: the failure to check what already exists before asking which of several imagined possibilities was meant.

A name X can be resolved, in the sense that competent readers agree which document or discourse it refers to, without being specification-resolved, in the sense that the identi-

fied material determines a sufficiently constrained space of programs that competent implementers, working independently from it alone, would produce observationally equivalent results. A prose architecture D falls into one of three regimes with respect to the space of programs $[[D]]$ satisfying it: complete, where $[[D]]$ is narrow enough that its members are observationally equivalent; ambiguous, where $[[D]]$ contains several materially different program families, none excluded by D as stated; or incoherent, where $[[D]] = \varnothing$, because D's requirements cannot be jointly satisfied by any program. A fourth condition is distinct from all three: D is underdetermined at a term t when t occurs in D's requirements without an operational meaning sufficient to evaluate satisfaction with respect to t, for any program. Underdetermination is not ambiguity, since ambiguity presupposes that satisfaction can be checked and merely finds several programs that pass; it is not incoherence, since incoherence presupposes satisfaction can be checked and finds it impossible. Underdetermination means the check itself cannot yet be run.

An implementer confronting a D that is not uniformly complete has three legitimate options, corresponding to the three genuine regimes. Faithful realization applies wherever D is complete. Explicit parameterization applies wherever D is ambiguous: the implementer selects among the materially different program families D admits and declares the selection as a choice rather than a discovery. Documented reconstruction applies wherever D is underdetermined or incoherent: the implementer supplies the missing operational content, or resolves the colliding requirements, and states plainly that this content did not come from D. Repair is not illegitimate; a reconstructed architecture may be more valuable than anything the original description, taken literally, would support. The problem is provenance: a repair presented as though it were a realization borrows the original's claimed authority for content the original never supplied, and the resulting program's success or failure becomes, illegitimately, evidence about the original theory rather than about the repair. This is a verify seam violated at the level of documentation rather than code: σ, the working implementation, propagates forward, while ρ, the fact that its correctness was manufactured rather than transcribed, does not.

A concrete case makes the failure mode legible. A wave-based architecture, MEM|8, contains operators used before being operationally defined, an instance of underdetermination; its stated complexity claims, an asserted constant-time associative retrieval alongside operations whose own description requires processing proportional to $N \log N$ per block, cannot jointly hold of any single program, an instance of incoherence; and its salience formula is precise enough to be evaluated numerically, and precise enough to reveal a divergence as one of its terms approaches a boundary value, an instance of a formally complete component that is nonetheless pathological. A request to produce "a working MEM|8 implementation" is therefore not answerable by transcription at all; at each of these points an implementer must perform an operation, and the resulting program's coherence depends entirely on which operations were performed where.

The deeper failure in this case occurred one step earlier than any of the three regime-specific repairs. A clarifying question was posed, faithful realization or critical reconstruction, as though these exhausted the space of what the name could refer to, without first checking whether the named system already existed as running code. It did: a search of the organization publicly associated with the architecture's authorship returns dozens of repositories already carrying the name, already making claims, already distinct in kind from the theoretical document originally analyzed. A question can display every surface feature of responsible uncertainty, an explicit refusal to guess, a request for the missing source, while still fabricating the shape of the ambiguity it claims only to be investigating. The earlier response did not invent facts about any particular file. It invented an ontology, exactly two versions when the honest first step available at that moment was not to ask which of two

was meant, but to check how many there were.

Where a single prose architecture has one space of candidate programs, a family of same-named artifacts requires a further distinction among nominal continuity, sharing the name; genealogical continuity, one derived from the other; structural continuity, sharing data types or operators; and behavioral continuity, producing comparable results under common tests, none implying the others. Determining which of these four relations hold between which members of a family is not a preliminary to implementation. It is the implementation problem, applied to a family instead of a single document, and it is exactly the specification-drift instance of Chapter 13's more general concern with whether verification results transfer across a transformation.

12.8 Conclusion

Whether at the level of a running pipeline, a compressed diagnostic history, or a request to implement a named architecture, the same shape recurs: a decision propagates while the grounds for it do not, and the discipline that keeps a system honest is not the absence of that gap but its explicit, auditable maintenance. A verify seam composes safely exactly as far as later composition respects the boundary it was built to enforce; a workaround is trustworthy exactly to the degree its discovery process, not merely its outcome, is legible; and a specification is honestly realized only when the difference between what it stated and what an implementer supplied remains visible in the resulting artifact. In every register, the failure is the same: treating consequence as though it carried its own justification along with it, when justification is precisely the thing a well-built seam was designed to hold back.

Chapter 13

Verification Inheritance

13.1 Introduction

Chapter 12 showed, through the MEM|8 case, that a family of same-named artifacts requires distinguishing nominal, genealogical, structural, and behavioral continuity, none implying the others, and that determining which of these relations hold is itself the implementation problem rather than a preliminary to it. This chapter formalizes the general version of that difficulty: when a scholarly or technical artifact containing a verified claim is transformed, reformatted, translated, paraphrased, or edited, does an earlier verification result transfer to the transformed artifact? The chapter shows that the underlying correspondence relation is undecidable in general, exhibits a restricted but genuinely useful class of transformations under which it becomes decidable, and specifies a protocol, Claimshift, that routes every actual case through one of three honest outcomes rather than collapsing the undecidable cases into a false positive or negative.

13.2 Isolating the Correspondence Problem

Verification is typically treated as a property of an artifact at the moment it is checked. A theorem is proved, a computation is checked, a claim is reviewed, and the result is recorded. But artifacts do not remain fixed. They are reformatted, translated, paraphrased, excerpted, and edited, and each transformation raises the same question: does an earlier verification result still apply to the transformed artifact?

The temptation, when building infrastructure for this problem, is to reach directly for a watermark or a provenance record and treat persistence of that record as evidence of continued validity. This is a category error. A watermark, a cryptographic commitment, or a provenance chain can establish that a transformed artifact traces back to an earlier one. None of these mechanisms can, by itself, establish that the claims the earlier artifact made are still the claims the transformed artifact makes. The two questions are independent, and conflating them allows a robust but semantically stale pointer to authenticate a claim that no longer holds. This is the identical failure mode Chapter 12 examined for a specification's own name: a name can be perfectly well resolved, everyone agrees which artifact is meant, while remaining entirely unresolved on whether the artifact's claims survived whatever produced the version currently in hand.

Fix a finite relational signature $\Sigma = \{P, Q, S\}$, where P and Q are unary relation symbols and S is binary. A claim is a closed sentence of first-order logic over Σ. Semantic entailment $c \models c'$ has its standard meaning.

Definition 13.1 (Correspondence)**.** For claims c, c', define $C(c, c') \iff (c \models c') \wedge (c' \models$

c). $C(c, c')$ holds exactly when c and c' are logically equivalent, true in exactly the same Σ-structures.

This is the strongest reasonable reading of "the transformed claim preserves the original," and the undecidability result below concerns this strongest reading; weaker preservation relations inherit the same undecidability by essentially the same argument, since equivalence reduces to two entailment checks.

13.3 The General Correspondence Problem Is Undecidable

Theorem 13.2 (Undecidability of correspondence)**.** *There is no total algorithm* $D : \mathrm{Sent}(\Sigma) \times \mathrm{Sent}(\Sigma) \to \{0, 1\}$ *such that* $D(c, c') = 1$ *if and only if* $C(c, c')$*, for all closed* Σ*-sentences* c, c'*.*

Fix the sentence $\tau := \forall x\,(P(x) \to P(x))$, which is valid: it is satisfied by every Σ-structure regardless of the interpretation of P, Q, or S. For any closed Σ-sentence φ, since τ is valid, every structure satisfying φ also satisfies τ, so $\varphi \models \tau$ holds unconditionally. Consequently $C(\varphi, \tau) \iff \tau \models \varphi \iff \varphi$ is valid. Suppose a total decider D for C existed. Then $\varphi \mapsto D(\varphi, \tau)$ would decide validity for arbitrary closed Σ-sentences φ. But validity of first-order sentences over a signature containing a relation symbol of arity at least two is undecidable, by Church's theorem on the undecidability of first-order logic. This contradiction establishes the theorem.

The requirement that Σ contain a relation symbol of arity at least two is not a technicality. First-order logic restricted to signatures consisting only of unary predicates, the monadic fragment, is decidable, by Löwenheim's theorem. Church's undecidability result, and therefore Theorem 13.2, depends essentially on the presence of at least one relation of arity two or more; S plays exactly this role, even though S does not appear in τ itself.

Theorem 13.2 says something narrower than "semantic correspondence between claims can never be established." It says that no single mechanical procedure decides correspondence correctly for every pair of claims expressible in a language this expressive. This rules out one particular and tempting design: a general-purpose `semanticEquivalent()` function, whether implemented by classical methods or by a learned model, that is assumed to give a correct verdict on an arbitrary pair of transformed and original claims. No such function can exist as a total, always-correct procedure. Any system that claims to certify correspondence in general is either incomplete, unsound, or restricted to a sub-language or a bounded class of transformations, whether or not it advertises that restriction.

13.4 A Certified Fragment: Correspondence by Normalization

Theorem 13.2 forecloses one design but does not foreclose all automation. If the class of admissible transformations is restricted rather than the claim language, correspondence can become decidable within that restricted class, and the restriction can still be expressive enough to matter.

Definition 13.3 (Rewrite system)**.** A rewrite system $\mathcal{R}$ over $\mathrm{Sent}(\Sigma)$ is a set of rules $c \to c'$ together with the reduction relation $\to_{\mathcal{R}}$ they generate and its reflexive transitive closure $\to^*_{\mathcal{R}}$. $\mathcal{R}$ is terminating if there is no infinite reduction sequence, and locally confluent if whenever $c \to_{\mathcal{R}} c_1$ and $c \to_{\mathcal{R}} c_2$, there exists d with $c_1 \to^*_{\mathcal{R}} d$ and $c_2 \to^*_{\mathcal{R}} d$.

By Newman's lemma, a terminating and locally confluent rewrite system is confluent, and confluence together with termination guarantees that every sentence c has a unique normal form $N_{\mathcal{R}}(c)$, independent of the order rules are applied.

Theorem 13.4 (Decidability by normalization)**.** *For any terminating, confluent rewrite system $\mathcal{R}$ over* Sent(Σ) *with effectively computable rule application, $C_{\mathcal{R}}(c, c') :\iff N_{\mathcal{R}}(c) = N_{\mathcal{R}}(c')$ is decidable.*

Termination guarantees computing $N_{\mathcal{R}}(c)$ halts after finitely many steps; confluence guarantees the resulting normal form is unique; $C_{\mathcal{R}}(c, c')$ is then decided by computing both normal forms and testing syntactic equality.

The value of this result depends entirely on $\mathcal{R}$ being non-trivial. A rewrite system whose only rules concern whitespace or byte-level formatting would make $C_{\mathcal{R}}$ decidable in a way that proves nothing of interest. A genuinely useful fragment includes alpha-renaming of bound variables to a canonical form, reordering of conjuncts and disjuncts under a fixed total order exploiting associativity and commutativity, double-negation elimination, and elimination of redundant parenthesization and syntactic sugar with a single semantics-preserving unfolding. Under such a system, $\forall x\,(P(x) \wedge Q(x))$ and $\forall y\,(Q(y) \wedge P(y))$ normalize to the same canonical form despite being lexically distinct, demonstrating that certified correspondence is not merely byte identity performed after cosmetic preprocessing.

13.5 A Three-Valued Decision Procedure

The two results together justify a decision procedure honest about the boundary between what it can certify and what it cannot. Given a certified rewrite system $\mathcal{R}$ and a transformation T taking c to c', let Dom$(\mathcal{R})$ denote the syntactic domain on which $\mathcal{R}$ is certified terminating and confluent. Then

$$D_{\mathcal{R}}(c, c') = \begin{cases} \texttt{PRESERVED}, & c, c' \text{ in fragment}, N_{\mathcal{R}}(c) = N_{\mathcal{R}}(c'); \\ \texttt{NOT\_PRESERVED\_UNDER\_R}, & c, c' \text{ in fragment}, N_{\mathcal{R}}(c) \neq N_{\mathcal{R}}(c'); \\ \texttt{OUTSIDE\_CERTIFIED\_FRAGMENT}, & \text{otherwise.} \end{cases}$$

Two readings of this output deserve emphasis, because both are easy to collapse into a simpler but incorrect binary judgment. First, `NOT_PRESERVED_UNDER_R` does not entail $\neg C(c, c')$. It entails only that this specific certified rewrite theory did not establish correspondence; the two claims might still be logically equivalent for reasons outside $\mathcal{R}$'s rule set. Second, `OUTSIDE_CERTIFIED_FRAGMENT` is not a semantic verdict at all. It reports that the pair, or the transformation connecting them, falls outside the domain on which this mechanism claims automatic certification, and correctly routes the pair toward whatever fallback mechanism, formal reproof, symbolic checking, or attested human judgment, a surrounding system chooses to require outside the certified regime.

Proposition 13.5 (Failure of certification is not certification of failure)**.** *A bounded automatic procedure has exactly three honest outputs, a positive certificate, a negative certificate relative to its own rule set, and an admission of inapplicability, and none of the three may be read as a general semantic verdict beyond what the procedure actually established.*

This is the exact discipline Chapter 10 enforces at the systems level with Adm, W, and Auth: admissibility supplies neither warrant nor authorization, and here, similarly, membership in Dom$(\mathcal{R})$ supplies neither a positive nor a negative correspondence verdict when the fragment's own rules do not resolve the pair. A companion protocol, Claimshift, described below, builds infrastructure around this boundary rather than treating it as a defect to engineer away.

13.6 The Claimshift Protocol

The theory above fixes a constraint; it does not by itself specify a system. Claimshift is a protocol for representing verification inheritance across document transformation, built on four design principles inherited directly from the theory: the locator is not the evidence, a robust in-band commitment establishes only that a transformed artifact points to a particular manifest, nothing about whether the claims in that manifest still correctly describe it; authorship is not generation, and assertion is not verification, a generative system that produces an artifact is not thereby its author in the relevant sense, and a party asserting a claim is not thereby a party who has verified it; verification is not truth, every verification record describes a method applied by a specific verifier under specific conditions and yields a result relative to that method, not an independent certification of truth; and failure of certification is not certification of failure, inherited directly from the three-valued output above.

13.6.1 Data Model

A claim record is a tuple $R_i = (c_i, A_i, \tau_i, D_i, S_i, V_i)$, where c_i is a canonical commitment to the claim's content, A_i identifies the asserting party, τ_i is a declared epistemic type (definition, axiom, derived, empirical, cited, conjectural, heuristic, or speculative), D_i is a set of typed dependency edges, S_i identifies supporting evidentiary sources, and V_i is the set of verification records attached to c_i. A dependency edge is typed rather than flat,

$$D_i = \{(c_j, \rho_{ij}) : \rho_{ij} \subseteq \{\text{semantic}, \text{interface}, \text{proof\_term}\}\},$$

because a flat edge conflates three genuinely distinct relations: whether changing c_j can change what c_i asserts, whether constructing c_i's justification requires c_j's signature without c_i's correctness depending on c_j's internal justification, and whether c_i's specific certified justification happens to reference c_j independent of whether c_i's statement mentions it at all. This typing was added after empirical testing under a preregistered mutation protocol showed the flat schema responsible for mispredictions traceable directly to the conflation.

A verification record is a tuple $V = (m, I, O, E, r, \sigma_V)$, where m identifies the verification method, I and O are the method's inputs and outputs, E records environment and version information, $r \in \{\text{accepted}, \text{rejected}, \text{inconclusive}\}$ is the result, and σ_V is the verifier's signature over the whole tuple hashed together with c_i's commitment. Epistemic type τ_i is a declared classification, not a computed one; it is only as trustworthy as the party who declared it and the policy under which that declaration is accepted. Verification records, by contrast, carry their own signed evidence and are the protocol's actual carriers of certification.

A transformation attestation is a tuple

$$A_{i,T} = (H(c_i), H(c_i'), T, J, r, e, \sigma_V),$$

where T identifies the transformation applied, J identifies the adjudication regime used, $r \in \{\text{preserved}, \text{not\_preserved}, \text{outside\_regime}\}$ is the result, and e is supporting evidence. The result alphabet is deliberately the protocol-level image of $D_{\mathcal{R}}$'s output:

$$\texttt{preserved} \leftrightarrow \texttt{PRESERVED}, \qquad \texttt{not\_preserved} \leftrightarrow \texttt{NOT\_PRESERVED\_UNDER\_R},$$

when J is a certified rewrite regime, generalizing to any regime, formal or informal, that supplies a comparably disciplined result. The value `outside_regime` is the protocol-level image of `OUTSIDE_CERTIFIED_FRAGMENT`; it must never be silently converted into either of the other two values by any downstream consumer of the manifest.

13.6.2 Layered Architecture and Signature Semantics

The protocol separates three layers: artifact, robust commitment, and signed manifest. The artifact is whatever document a reader encounters; the robust commitment is a compact signal embedded in or attached to the artifact, carried by any sufficiently robust technology, a statistical text watermark, an image or audio watermark, a content-credential field, or ordinary file metadata, the protocol being agnostic to the carrier; the manifest is a signed, append-only document containing the full claim graph. Separating these layers means a scholarly document's entire dependency graph need not be encoded steganographically inside the artifact itself: the embedded commitment need only be robust enough to survive the expected transformations and locate a manifest that can be fetched and reasoned over separately. A claim published with no verification records can later acquire one, and if a proof is subsequently found to rely on an error, a further verification record with $r =$ rejected is appended rather than the earlier one being deleted, preserving an append-only epistemic history rather than a single mutable verdict, in exact correspondence with Chapter 10's history-integrity contract.

The protocol defines three signature roles that must never be represented by the same key or the same manifest field. An assertion signature means that a party asserts a claim record as stated, saying nothing about whether that party verified, generated, or merely transcribed it. A verification signature means a verifier performed a specific method against a specific representation and obtained a specific result, saying nothing about whether the method is trustworthy for the claim type in question. A transformation attestation signature means the attesting party asserts a specific correspondence judgment under a specific regime, saying nothing about whether that party was authorized to use that regime for that claim type.

Proposition 13.6 (Non-conflation)**.** *No manifest entry may be interpreted as satisfying more than one of the assertion, verification, and transformation-attestation roles unless it carries the corresponding distinct signature for each role it claims to satisfy.*

13.6.3 Policy Layer and Dependency Propagation

A transformation attestation is meaningless on its own; it must additionally be established that the attesting verifier was entitled to use the invoked adjudication regime for a claim of the relevant type. For each epistemic type τ, a policy $P_\tau = (J_\tau, V_\tau, Q_\tau)$ specifies the accepted adjudication regimes, the accepted verifiers and their trust roots, and a quorum or assurance requirement. Authorization and correspondence are two separate predicates that must both hold: $P_\tau \vdash V$ authorized to issue judgments under J, distinct from $V \vdash C_J(c, c')$. Verification inheritance is defined conditionally, requiring an existing verification, a preserved transformation attestation, and satisfied authorization and quorum conditions together:

$$V(c_i) \wedge A_{i,T} = \text{preserved} \wedge J \in J_{\tau_i} \wedge V_{\text{att}} \in V_{\tau_i} \wedge Q_{\tau_i} \text{ satisfied} \implies V_T(c_i').$$

Absent any one of these conditions, in particular absent an explicit transformation attestation at all, $V(c_i) \not\Rightarrow V(c_i')$. This is the protocol-level statement of the theory's central discipline: persistence of a robust locator across a transformation is never, by itself, sufficient grounds for inheritance.

The typed dependency edges support role-parameterized closures. For a role set

$$\rho \subseteq \{\text{semantic}, \text{interface}, \text{proof\_term}\},$$

$\text{Support}_\rho(c)$ and $\text{Affected}_\rho(c)$ are the corresponding reachability closures over the dependency relation. A policy-qualified verification query asks whether every claim a target claim

semantically depends on satisfies a given verification policy, $\text{Verified}_P(c) \iff \forall d \in \text{Support}_{\{\text{semantic}\}}(c), V(d) \models P$. The restriction to the semantic role is deliberate: a proof-term or interface dependency need not be load-bearing for what a claim actually asserts, and a protocol implementation that evaluated this query over the unrestricted union of all three roles would over-count, treating auxiliary structural dependencies as though they were constitutive of meaning. This is a precise, typed instance of the four-way continuity distinction Chapter 12 developed informally for the MEM|8 family: nominal, genealogical, structural, and behavioral continuity are, in effect, four different candidate role sets for the same underlying question, and conflating them produces exactly the kind of over- or under-counted verification status this role restriction is built to prevent.

13.7 Conclusion

The correspondence problem this chapter isolates is undecidable in general, decidable within an honest and non-trivial fragment, and, outside that fragment, resistant to being silently collapsed into either a positive or negative verdict without manufacturing a claim the underlying theory does not support. The Claimshift protocol does not solve the correspondence problem; it represents whichever of the three outcomes actually obtains, with its adjudicator properly authorized under an explicit, claim-type-indexed policy, and refuses, as a protocol invariant rather than a best practice, to convert an outside-the-fragment result into a judgment it did not earn. The practical consequence is a design constraint rather than a system: any mechanism for verification inheritance under transformation, however built, must expose the boundary between what it can certify and what it cannot, rather than letting the survival of a pointer stand in for the validity of what it points to. This is the same discipline Chapter 10 enforces for commitment and Chapter 12 enforces for consequence, applied here to the specific case of a claim surviving, or failing to survive, its own transformation.

Chapter 14

Model Capsules and Ternary Learning

14.1 The Deployment-Gap Problem

Chapter 13 treated the question of whether a claim survives its own transformation, formally and in general. This chapter treats a specific, concrete instance of the same question, closer to the machine than to the document: whether a trained model's deployed behavior survives the transformations, quantization, export, kernel substitution, tokenizer binding, that separate optimization from execution.

A conventional model pipeline often contains several non-identical machines. Optimization uses floating-point shadow weights and training kernels; export applies quantization and layout conversion; inference invokes a separate runtime; and deployment adds platform-specific tokenization, scaling, memory, and scheduling behavior. Every transformation can preserve average benchmark performance while changing individual outputs, numerical stability, latency, or failure behavior. The relevant object is therefore not merely a parameter tensor θ, but the complete deployed computation,

$$F = (G, Q, S, T, K, R, P),$$

where G is the graph or bytecode, Q the weight alphabet and packing rule, S the scale metadata, T the tokenizer and preprocessing state, K the kernel semantics, R the runtime contract, and P the execution policy. Two artifacts with identical ternary weights can still implement different functions when any other component differs.

Deployment-native training does not require every training operation to be low precision. Gradient accumulation and optimizer state may remain floating point. It requires that the forward operator used to estimate the task loss faithfully reproduce the discrete weights, scales, rounding, saturation, activation treatment, and layer ordering that will be used after export.

Deployment drift separates into components arising at distinct stages of the pipeline, each admitting a different remedy. Representation drift arises when the forward operator used for loss estimation differs from the exported operator, most commonly through a mismatched straight-through surrogate or an export-time rounding rule not exercised in training. Kernel drift arises when the mathematically specified operator is realized by different low-level implementations across devices, for example a fused kernel that reorders accumulation and changes rounding under finite precision. Preprocessing drift arises when the tokenizer, normalization, or templating logic bound to the training data differs from the version bound to the runtime capsule, which can change token boundaries without changing any weight. Scheduling drift arises when batching, padding, or caching strategies at inference time alter numerical results relative to the unbatched computation used during

evaluation. These four sources are not mutually exclusive and can partially cancel or compound; a pipeline that reports only aggregate deployment disagreement without attributing it to a source cannot say which remedy would help.

14.2 Ternary Parameterization

Let a layer maintain real-valued shadow parameters $W \in \mathbb{R}^{m\times n}$ while executing a ternary forward matrix $\hat{W}$. A simple symmetric quantizer uses a scale $\alpha > 0$ and threshold $\Delta \geq 0$:

$$\hat{W}_{ij} = \alpha \cdot q(W_{ij}/\alpha), \qquad q(x) \in \{-1, 0, +1\}, \qquad q(x) = \begin{cases} -1 & x < -\Delta \\ 0 & |x| \leq \Delta \\ +1 & x > \Delta. \end{cases}$$

The ternary symbol requires $\log_2 3 \approx 1.585$ bits of information under a uniform code, although practical packing may use two physical bits per weight or block encodings with amortized overhead; calling the representation strictly one-bit is therefore imprecise. Its computational appeal is that multiplication by a ternary weight reduces to signed accumulation and omission, subject to the cost of scales, activation arithmetic, packing, memory movement, and kernel dispatch. Training commonly uses a straight-through approximation because q is piecewise constant: the backward pass substitutes a surrogate derivative for $\partial q/\partial W$, which means the learned solution depends on the surrogate, clipping rule, scale estimator, and update schedule, so reproducibility requires recording these rules, not only the final packed weights.

The symmetric, per-tensor quantizer is the simplest member of a larger family, and the family member chosen is itself part of the deployment specification F rather than an incidental training detail. A per-channel quantizer assigns an independent scale α_k to each output channel, reducing quantization error on layers with heterogeneous weight magnitude at the cost of storing a vector of scales rather than a scalar; the capsule manifest must then record the reduction axis and count, since a runtime that assumes a single scalar scale will silently misinterpret a per-channel capsule rather than failing loudly. An asymmetric variant introduces an offset β, producing reconstruction levels $\{\beta - \alpha, \beta, \beta + \alpha\}$, which is a three-level operator rather than a zero-centered ternary weight in the sense used elsewhere: signed accumulation eliminates multiplication only when the reconstruction levels are exactly $\{-1, 0, +1\}$, so a deployment adopting the asymmetric variant should report it as a distinct operator with its own efficiency profile.

Deployment equivalence itself needs a measurable definition. Let $f_{\text{train}}(x;\theta)$ denote the forward pass used while training and $f_{\text{run}}(x;C)$ the runtime evaluation of a serialized capsule C. For a test distribution D,

$$\delta_{\text{out}} = \Pr_{x\sim D}\big[\arg\max f_{\text{train}}(x;\theta) \neq \arg\max f_{\text{run}}(x;C)\big],$$

$$\delta_{\text{num}} = \mathbb{E}_{x\sim D}\left[\frac{\|f_{\text{train}}(x;\theta) - f_{\text{run}}(x;C)\|_2}{\|f_{\text{train}}(x;\theta)\|_2 + \varepsilon}\right].$$

Exact bitwise agreement may be unrealistic across heterogeneous devices, but token-level disagreement, normalized logit error, and task-level divergence are measurable, and a pipeline should state which equivalence criterion it promises and on which platforms it has been tested.

14.3 The Model Capsule

The deployable unit should be a self-describing capsule rather than an unstructured weight file, binding executable semantics to model state and supporting hashing, signature verification, compatibility checks, and deterministic reconstruction. Its required components are a manifest recording format version, model identity, hashes, license, and provenance; a program, typed bytecode or a restricted graph specifying computation independently of host-language objects; parameters, packed ternary weights and scale blocks preserving the executed representation; a text interface, tokenizer, normalization, vocabulary, and templates, avoiding silent input mismatch; a runtime contract specifying operators, numerical rules, memory limits, and ABI, making compatibility testable; and evidence, an evaluation summary, calibration set hash, and build attestation recording what has actually been measured. A capsule should not contain arbitrary native code. A restricted instruction set with bounded tensor shapes, declared memory, immutable parameter regions, and explicit operator versions is easier to validate and isolate. The signature authenticates the bytes and declared producer; it does not establish model quality or safety.

A capsule format that changes without a disciplined versioning rule reintroduces the deployment gap it was built to close, since a runtime and a capsule produced under different silent assumptions can both parse successfully while disagreeing on semantics. A workable convention assigns three independent version fields to the manifest: a format version governing the container schema itself, a contract version governing the runtime operators, numerical rules, and ABI, and a content version identifying the specific trained artifact. A conservative runtime may require exact format-version equality. Contract versions should be compared under a declared compatibility relation, for example a runtime advertising contract version v accepting any capsule declaring a contract version in a closed range $[v_{\min}, v]$ that the runtime publishes, so older capsules remain executable under runtimes that have only added operators and newer capsules are refused rather than silently misexecuted. Content versions carry no compatibility semantics of their own; they exist so that two capsules with identical format and contract versions can still be distinguished, hashed, and compared for provenance. Refusal is warranted whenever the contract-version range test fails; acceptance is not warranted merely because the capsule parses. This three-field versioning scheme is a concrete instance of Chapter 13's claim-type policy layer, applied to a model artifact rather than a scholarly claim: format, contract, and content versions are, respectively, an admissibility check, a compatibility-regime check, and a provenance identifier, exactly the three roles that chapter kept deliberately separate.

14.4 Channels as Streams of Executable Models

Typed channels normally transport values between concurrent processes. If a channel value is a model capsule, the receiving process can validate it, allocate a fresh arena, bind inputs, and execute the declared graph. The channel then represents a stream of parameterized computations. This is best understood as mobile model description, not mobile authority: receipt of a capsule must not grant filesystem, network, process, or device access.

$$\text{send}(C) \to \text{receive}(C) \to \text{verify}(C) \to \text{allocate}(A) \to \text{execute}(C, A, x) \to y.$$

Fresh arenas provide useful isolation between invocations, eliminating accidental reuse of mutable intermediate state unless it is explicitly declared and transferred. They do not by themselves prevent denial of service, side channels, malformed-shape attacks, or vulnerabilities in the verifier and kernels.

Let a receiver state be $\sigma = (A, M, V)$, where A is the available arena, M the admitted capsule set, and V the runtime version. The transition $\mathrm{Exec}(C, x)$ is admissible only if the capsule signature and hashes verify, its declared operators are supported by V, its maximum memory and instruction budget fit A, and its input contract accepts x. Execution occurs in a new arena A' and returns either a typed result y with an execution record e, or a typed refusal r:

$$(\sigma, C, x) \to (\sigma, y, e) \qquad \text{or} \qquad (\sigma, C, x) \to (\sigma, r).$$

Refusal is part of the protocol, not an exceptional accident, in precisely the sense Chapter 11 argued refusal must be a first-class, non-punitive outcome rather than an absence of production. The execution record should include capsule hash, runtime hash, device class, input hash or privacy-preserving reference, start and end times, resource counters, and result status; sensitive inputs and outputs need not be placed in the record.

14.5 Continuous Data Loops and Feedback Hazards

Repeated training on self-generated text can narrow linguistic and conceptual diversity, amplify systematic errors, leak evaluation material, and reward artifacts of the validator. A robust loop therefore keeps an immutable external evaluation set, retains real or independently sourced data, measures duplication and source concentration, records ancestry between examples, and limits the number of generations through which unsupported claims can propagate:

$$D_{t+1} = R_t \cup H_t \cup \{z \in G_t : V_t(z) = \text{accept}\},$$

where R_t denotes independently sourced material, H_t human-reviewed examples, G_t generated candidates, and V_t the versioned validator. The decomposition makes it possible to report performance by data origin rather than obscuring the synthetic fraction in a single aggregate.

The scientifically defensible version of an open-versus-private service uses the same training protocol, acceptance tests, and reporting schema in both modes. The variable being purchased is isolation and operational control, not a higher standard of evidence. Public runs may publish data, configuration, metrics, and artifacts; private runs keep customer data and outputs inside a customer-controlled enclave while emitting only the attestations and reports the customer authorizes. Bring-your-own-cloud deployment reduces custody by allowing the customer to own accounts, encryption keys, logs, storage, and teardown, but it does not eliminate trust: the customer must still assess the supplied code, binary provenance, infrastructure declarations, dependencies, and operator access paths.

14.6 Security and the Trusted Computing Base

The relevant adversaries include a malicious capsule producer, a compromised build dependency, an operator with excessive cloud privileges, a customer dataset containing secrets or prompt injections, and an external party observing timing or resource use. The trusted computing base should therefore be as small as practical: parser, verifier, allocator, operator kernels, signature implementation, and the mechanism enforcing resource and I/O policy. Malformed capsules are controlled by canonical parsing, shape validation, and fuzzing, with residual risk in verifier or kernel vulnerability; arbitrary code execution is controlled by restricted bytecode with no native extensions, with residual risk in bugs among the permitted operators; resource exhaustion is controlled by static bounds and runtime quotas, with residual risk in adversarial worst-case scheduling; training-data leakage

is controlled by access controls, minimization, and memorization tests, with residual risk that inference-time extraction remains possible; supply-chain substitution is controlled by a pinned closure, hashes, signatures, and reproducible builds, with residual risk in a compromised trusted source or key; and cross-job contamination is controlled by fresh arenas and tenant isolation, with residual risk in hardware and timing side channels. Privacy is a property of the complete workflow, not a synonym for running in a private account; it requires an explicit data-retention policy, logging policy, key ownership model, access review, egress controls, incident procedure, and deletion verification, and neither differential privacy nor confidential-computing hardware should be implied merely by the word "enclave."

Shrinking the trusted computing base is only useful to the extent that the cost of exercising it stays bounded, so a deployment claim should report the verification overhead alongside the security boundary. Let c_v denote the cost of the parse-and-verify step for a capsule and c_x the cost of subsequently executing it once. The relative verification overhead $c_v/(c_v + c_x)$ is large and largely irrelevant for a capsule executed many times after a single verification, but it dominates the cost of a workload streaming many distinct small capsules through a receiver, verifying each once before a single execution; the two regimes should not be reported under one aggregate latency figure. A related and separable cost is re-verification triggered by contract-version changes: a runtime upgrade that narrows its accepted contract-version range can force re-verification or outright refusal of a large previously admitted capsule population, an operational cost of the versioning discipline that should be planned for rather than discovered at upgrade time.

14.7 Experimental Program and Falsifiable Claims

The architecture should be tested through preregistered comparisons rather than demonstrations alone. The primary baseline is a conventional floating-point training and post-training export pipeline; the deployment-native condition uses the target quantizer and runtime-equivalent forward semantics during training, with both receiving matched data, parameter counts, optimization budgets, and evaluation procedures. Quality and efficiency must be reported jointly: a speedup measured at lower accuracy, shorter context, a different sampler, or altered batch size is not a controlled comparison, and energy claims require wall-plug or device-level measurement rather than arithmetic-operation counts alone.

A useful experimental sequence first verifies byte-level capsule stability and operator conformance on small networks, then measures deployment disagreement for individual layers and complete models, then compares ternary-native training with post-training quantization, and only after those controls pass adds capsule streaming, multi-tenant execution, and the continuous data loop, an ordering that prevents a complex integration from concealing a basic numerical mismatch. Because δ_{out} and δ_{num} are estimated from a finite evaluation set, a reported reduction in either quantity should be accompanied by an interval estimate and a preregistered minimum detectable effect, not a point estimate alone, and a family of ablations tested against a shared evaluation set should disclose the resulting multiple-comparisons burden rather than reporting each result as though it were independently significant.

The proposal should be rejected or revised if deployment-native training does not reduce output disagreement relative to a carefully tuned export baseline; if claimed efficiency disappears at matched task quality; if the silver-label loop improves its internal validator while degrading external evaluation; if capsule verification cannot bound resource use; or if private operation depends on undeclared operator access. Negative results would still be informative because they would locate the cost of representation alignment and identify which integration layers provide no measurable benefit, precisely the falsifiability discipline

Chapter 2 argued a research program needs and a flat-likelihood hypothesis lacks.

14.8 Conclusion

The strongest idea in this architecture is not that ternary cells are miniature intelligent agents or that channels become fundamentally new objects. It is that a deployed neural computation can be packaged as a constrained, typed, verifiable program whose discrete representation is present during optimization, evaluation, transport, and execution. This changes the unit of reproducibility from a weight file to a complete computational contract, and it is the material instance, at the level of an actual deployed system, of the same discipline this Part has developed at every other level: CLIO's separation of documentary history from operative state, the verify seam's separation of consequence from rationale, and Claimshift's separation of a claim's identity from its correspondence under transformation all recur here as the separation of a model's packed weights from the complete deployment specification F that determines what those weights actually compute.

Part IV

Representations and Compression: Holography, Sparsity, and the Fallacy of Brevity

Chapter 15

The Compression Fallacy

15.1 An Old Definition, Asked of a New Case

Part III closed with a model capsule whose distinctive contribution was to bind a deployed computation's discrete representation to the machinery required to execute it, rather than reporting the representation's size alone. This chapter takes up the general question that binding was built to answer: does a shorter description of something mean that thing has been better understood?

The computer scientist Daniel Hillis once offered a definition of the information content of a pattern: the amount of information in a pattern of bits equals the length of the smallest computer program capable of generating those bits, in substance Kolmogorov complexity restated for a general audience. What this chapter is concerned with is a further claim the definition has often been taken, by others, to support, though it is not one Hillis himself is on record as drawing: that a system which produces a shorter description of something has thereby understood that thing better than a system producing a longer one. The claim is rarely defended directly. It is simply assumed, in the background of arguments about which model of a phenomenon is the better model, which representation of a domain is the more insightful representation, and, increasingly, in discussions of large trained models, which network has learned the most genuine structure rather than having merely memorized surface statistics.

Forth, an unusually terse programming language whose devotees have long pointed to Hillis's definition as evidence of the language's virtue, supplies an unusually clean test case, because Forth's brevity is well documented, uncontroversial, and produced by a mechanism that is easy to state precisely: word-factored Forth code is short because a Forth programmer builds a library of named operations, words, and subsequent code invokes them rather than re-deriving them. A task solved in Forth after such a library exists can be several times shorter, in raw character count, than the same task solved in a general-purpose language without access to an equivalent library. The question is not whether Forth code is short; that is not in dispute. The question is what that shortness is evidence of, and whether it is evidence of the thing Hillis's definition, read the way it is usually read, would have it be evidence of.

This chapter argues that it is not, and that the same gap separating Forth's brevity from Kolmogorov-optimal description separates, more consequentially, model compactness from model understanding wherever the comparison is drawn.

15.2 Coding Length and Distinguishing Capacity

A general framework for measuring representational capacity takes as its basic object not a set of items to be described, nor a description language considered on its own, but the pairing of the two together with a notion of when two items count as the same for the purposes at hand.

Definition 15.1 (Ontological triple)**.** An ontological triple $(X, \sim, T)$ consists of a set X, an equivalence relation $\sim$ on X specifying which elements are treated as observationally indistinguishable, and an ontology T: a description language, or template hierarchy, determining which descriptions of elements of X are actually reachable, as opposed to merely existing in principle.

The distinction between what a description language can express in principle and what it can reach in practice, given the templates and operations it has already accumulated, is the hinge the rest of this chapter turns on. Two deficits are defined over this structure, and they measure different things. The coding deficit of an object x, written $\delta_T(x)$, is the excess length of the shortest description of x reachable within T over the Kolmogorov-optimal description length of x, the length of the absolute shortest program capable of generating x, independent of any particular ontology. The distinguishability deficit of x, written $\Delta_T(x)$, is a different quantity: let $D_T(x)$ denote the distinguishing capacity of T at x, the number of other objects that can be separated from x by descriptions reachable in T, and let $D^*(x)$ denote the maximum such capacity attainable by any ontology whatsoever. Then $\Delta_T(x) = D^*(x) - D_T(x)$. The coding deficit asks how much longer a description is than it needs to be. The distinguishability deficit asks how much distinguishing power has been lost. Nothing in the definitions forces these two quantities to move together, and the relationship between them is the object a theorem is required to establish.

An ontology decomposes as a triple $T = (E_T, M_T, \mathcal{O}_T)$: an encoder $E_T : Q \to \{0,1\}^*$ over the semantic quotient $Q = X/\sim$, giving each semantic class its reachable description; background machinery M_T, the accumulated templates, interpreter, or structure constants required to produce or use those descriptions in the first place; and an operational repertoire $\mathcal{O}_T$, the set of operations computable directly over the encoded representations without first decoding back to Q.

Definition 15.2 (Three coordinates of a representation)**.** Distinguish three properties: the description magnitude $L_T(q) = |E_T(q)|$; the distinguishing structure $\mathcal{P}_T = Q/\sim_T$, where $q \sim_T r \iff E_T(q) = E_T(r)$; and the operational repertoire $\mathcal{O}_T$. These measure different properties of T, and equality or improvement in one does not, without an additional coupling assumption supplied from outside the definitions, imply equality or improvement in either of the others:

$$L_T \not\Rightarrow \mathcal{P}_T, \qquad L_T \not\Rightarrow \mathcal{O}_T, \qquad \mathcal{P}_T \not\Rightarrow \mathcal{O}_T.$$

15.3 Lossless Compression Does Not Imply Ontological Economy

Before turning to the machine-learning case, it is worth addressing the objection this apparatus invites immediately: doesn't lossless compression preserve all the information in a representation, so that a shorter lossless encoding must preserve every distinction the longer one did? At the level of static distinguishability this case is straightforward. A ZIP archive can become dramatically shorter than the file it compresses while remaining perfectly invertible: its encoder is injective on Q, so $\sim_T$ collapses nothing, and $D_T(q)$ is maximal for every q. Lossless compression is accordingly not a counterexample to static distinguishability preservation.

What the ZIP case exposes instead is the third quantity, $\mathcal{O}_T$. That two encoded objects remain distinct, in the static sense D_T measures, does not imply that every relation of interest between their uncompressed contents remains directly computable over the encoded representations themselves. Two compressed files can be told apart trivially, their bitstrings differ, without any geometric or relational property of what they contain being available for comparison until both are decompressed; if decompression is not itself among the operations $\mathcal{O}_T$ makes native, those relations sit outside $\mathcal{O}_T$ even though D_T is already maximal. Description magnitude, distinguishing structure, and operational repertoire are logically independent properties of an ontology, and a lossless archive is the case in which the first two are already exemplary while the third remains, until an inverse transform is paid for, essentially empty.

The converse case rules out the opposite inference. An extremely verbose database, one that stores every field of every record with no compression whatsoever, can distinguish every member of its domain perfectly despite being terribly compressed: $\delta_T(x)$ large, $\Delta_T(x) = 0$. Nobody encountering such a database concludes that its size is evidence against its usefulness, because nobody was using size as a proxy for distinguishing power in the first place; the intuition that a bloated schema is not thereby a poorly understood one is the intuition this chapter is asking to be extended to the cases, sparse latent codes, monosemantic feature dictionaries, where the opposite inference is currently made routinely.

15.4 Compression Relative to What?

A further assumption is buried in any claim that a representation is short: short relative to what reference machinery. Kolmogorov complexity has its invariance theorem, which guarantees that the choice of universal machine affects $L^*(x)$ by at most an additive constant independent of x, but that guarantee is asymptotic, and practical comparisons between representational systems are made nowhere near the regime in which an additive constant can be safely ignored. Writing $L_{\text{visible}}(q) = L_T(q)$ for the quantity ordinarily reported and $L_{\text{system}}(q) = L(M_T) + L_T(q)$ for the full cost,

$$L_{T_2}(q) < L_{T_1}(q) \quad \not\Rightarrow \quad L(M_{T_2}) + L_{T_2}(q) < L(M_{T_1}) + L_{T_1}(q).$$

Forth makes the point vivid: the token `FOO` is three characters, but the definition of `FOO` together with the words it depends on can represent thousands of characters of accumulated machinery. Reporting the visible length without the machinery is not measuring a smaller quantity than the full system cost; it is measuring a different, incomplete one. The same separation generalizes past Forth without alteration: a short latent code in a trained autoencoder can depend on a decoder with billions of parameters, and reporting the latent code's length while omitting the decoder's is the identical move under a different name.

Proposition 15.3 (Machinery Displacement)**.** *Let $Q = X/\sim$ be finite. There exists an ontology $T_2 = (E_2, M_2)$ in which every $q \in Q$ has a fixed-width description of length $\lceil \log_2 |Q| \rceil$, while the machinery required to interpret those descriptions contains the corresponding lookup structure. Consequently, reducing the per-object description length does not by itself establish that representational complexity has been eliminated rather than displaced into background machinery.*

Enumerate $Q = \{q_0, \ldots, q_{n-1}\}$ and define $E_2(q_i) = \text{bin}_{\lceil \log_2 n \rceil}(i)$, with M_2 containing the correspondence between indices and the classes they denote. Every encoded object then has the same short description length, while interpreting those descriptions presupposes the correspondence stored in M_2. The question worth asking of any short description is accordingly not how short is it, but what machinery must already exist for it to recover the

object it names. A shell command invokes an entire operating system. A Forth word invokes its dictionary. A latent vector invokes a decoder. A theorem invokes its definitions and lemmas.

15.5 Description-Length Non-Identification

Proposition 15.4 (Description-Length Non-Identification)**.** *Let $Q = X/\sim$ be finite with $|Q| \geq 3$, and let $q \in Q$. Then: (i) for every pair (ℓ, k) with $\ell \geq 1$ and $0 \leq k \leq |Q| - 1$, there exists an ontology T with $L_T(q) = \ell$ and $D_T(q) = k$; and (ii) for each relation $R \in \{<, =, >\}$, there exist T_1, T_2 with $L_{T_1}(q) < L_{T_2}(q)$ and $D_{T_1}(q)\, R\, D_{T_2}(q)$.*

Let $n = |Q|$. For (i), choose a subset $C \subseteq Q$ containing q with $|C| = n - k$. Define the encoder so every element of C receives the same description s of length ℓ, and assign every element of $Q \setminus C$ a description distinct from s and from every other such description, which is always possible since $\{0,1\}^*$ is infinite. The encoder-induced class of q is then exactly C, so $D_T(q) = k$ and $|E_T(q)| = \ell$. Statement (ii) follows by applying (i) twice with $\ell_1 < \ell_2$ and k_1, k_2 chosen so that $k_1\, R\, k_2$.

A fully worked instance, fixing $Q = \{a, b, c\}$ and comparing five encoders at $q = a$, makes the construction checkable directly. A total-collapse encoder gives $(L, D) = (3, 0)$; separating a alone from b, c merged gives $(1, 2)$; grouping a with b gives $(2, 1)$; an injective two-character encoding gives $(2, 2)$; and a padded injective four-character encoding gives $(4, 2)$. Comparing the third and fourth: equal length, unequal capacity. Comparing the second and first: shorter length, strictly higher capacity. Comparing the third and fifth: shorter length, lower capacity, the case ordinary usage means by "compression loses information," present here as one populated cell among several rather than the only possibility. Comparing the fourth and fifth: shorter injective encoding against a longer injective one, identical capacity. Every cell of the ordering table is realized by an explicit construction at the same fixed point; none is left as an existence claim.

compression magnitude does not determine compression topology

Proposition 15.5 (Dynamic non-identification)**.** *Let $(T_t)_{t=0}^{m}$ be a sequence of ontologies over a fixed finite Q, and $q \in Q$. Write $\Delta L_t = L_{T_{t+1}}(q) - L_{T_t}(q)$ and $\Delta D_t = D_{T_{t+1}}(q) - D_{T_t}(q)$. Without additional constraints linking description length to encoder-induced partitions,* $\operatorname{sgn}(\Delta L_t) \not\Rightarrow \operatorname{sgn}(\Delta D_t)$.

Apply Proposition 15.4 independently at times t and $t+1$, choosing $\ell_{t+1} < \ell_t$ while choosing k_t, k_{t+1} to realize any desired ordering between them. Differentiating an unidentified relationship with respect to time does not identify it. A compression-progress account that reads $\Delta L_t < 0$ as evidence of improving understanding, or a temporary $\Delta L_t > 0$ as evidence of regress, commits Proposition 15.4's error once per time step, not avoided by moving to a derivative.

15.6 The Proxy Theorem as a Conditional Result

A previously proposed result, the Deficit Proxy Theorem, is worth examining as a candidate additional-assumption case of Proposition 15.4, the kind of result that pins down one point on an otherwise unidentified surface by adding a specific, checkable assumption. Under an assumption of faithful encoding, that every distinction a description language can express costs at least one bit, the theorem states that the distinguishability deficit is bounded above

by a constant multiple of the coding deficit, with the special case that a zero coding deficit implies a zero distinguishability deficit: an ontology achieving the Kolmogorov-optimal description of an object would thereby also achieve the maximum possible distinguishing capacity for it.

Closer inspection shows even this special case is not secured by the argument given for it. The step from Kolmogorov-optimality to the claimed conclusion treats generating an object cheaply and distinguishing that object from an ambient domain as though they were the same achievement; they are not, and closing the gap would require an additional condition tying the optimal description length to a maximally distinguishing encoding specifically, a condition the source result does not supply. The theorem is presented here as a previously proposed conditional result under examination rather than as an established point this chapter's own argument depends on. The converse, that a distinguishability deficit of zero forces a coding deficit of zero, fares no better; it is stated only as a conjecture in its source, resting on the further, unargued assumption that optimal coding uses every distinction available to it.

The asymmetry matters because the popular reading of Hillis's definition uses the coding deficit as a graded, continuous stand-in for the distinguishability deficit across its whole range, treating even the strongest available version of the Deficit Proxy Theorem as though it licensed that continuous use. It would not license it even if its own special case were secure: an inference from perfect compression to perfect distinguishing capacity is not an inference that a further reduction in an already-imperfect coding deficit constitutes evidence of a further reduction in distinguishing capacity.

15.7 Forth and Reachable Complexity

Forth supplies reachable complexity's first concrete instance: the shortest description of an object expressible from a description language's current template hierarchy, which may be strictly longer than the Kolmogorov-optimal description. A Forth programmer accumulates, over the life of a project, a library of named words, short, composable operations built from more primitive ones, each capturing an operational distinction the programmer has found useful to keep separately addressable. Code written after such a library exists is short because it is reachable, cheaply, from that library, not because it approaches the description length that would be achieved by an ontology with no library at all, starting from nothing.

This inverts the naive reading of Hillis's definition as applied to Forth. The naive reading takes Forth's brevity as evidence that Forth programs approach the information-theoretic floor of the tasks they solve. The reachable-complexity reading takes the same brevity as evidence of something else: that a well-factored library of admissible operations was available to be reused. Two Forth programmers solving the same task with different accumulated word libraries will produce differently short code for it, and the difference in length reports which library was available, not which programmer better understood the task. A companion result states the point in general form: no valid compression mapping can be constructed without prior identification of the invariant structure of the space being compressed, and that identification is itself achieved only through a prior process of expansion, exploration, iteration, and discovery of which structure is invariant in the first place, developed at length in Chapter 16. Forth's terseness is compression in exactly this downstream sense: it presupposes the expansion that built the word library, and reports that expansion's prior existence rather than the Kolmogorov-optimal floor of the task the short code happens to solve.

None of this counts against Forth. A large, well-factored word library achieving low coding deficit relative to itself, while the Kolmogorov-optimal length remains a separate and largely inaccessible quantity, is exactly what a mature, high-distinguishing-capacity repre-

sentational system should look like. What the coding deficit cannot do, without the further, unproved direction of the Proxy Theorem, is stand in for a claim about how much of the task's true distinguishing structure the library actually preserves.

15.8 The Same Move in Machine-Learned Representations

The informal assumption this chapter has been arguing against is closer to background theory in contemporary discussions of trained models, where a shorter, sparser, or more compressed internal representation is routinely treated as evidence of deeper or more genuine understanding. The assumption's most explicit and most influential statement extends compression progress well past a technical modeling virtue and treats it as the underlying currency of understanding itself, a unifying account of curiosity, creativity, and aesthetic interest in which subjective beauty and interestingness are identified with the rate at which an observer's compression of its data is currently improving. If the relationship between coding deficit and distinguishability deficit is only proven in one direction and only at the extreme, then a theory that identifies the felt sense of understanding, or beauty, or interest with compression progress at every point along the curve, rather than only at the point where compression is complete, is claiming a great deal more than the formal relationship between the two quantities actually licenses.

Two examples make the pattern concrete. The beta-VAE architecture increases a single compression pressure, a weighting term on the model's latent-code cost, specifically in order to produce latent representations described as more interpretable and as more faithfully capturing a domain's independent underlying factors of variation, with the increase in compression pressure offered as the mechanism by which the improvement is obtained. Sparse-autoencoder interpretability work on language models motivates the move to sparse, low-dimensional dictionaries of features in comparable terms: individual neurons are described as polysemantic and therefore poorly understood, while the sparse, more compressed feature directions recovered by the autoencoder are described as monosemantic and correspondingly more interpretable. In neither case is the inference from compression to understanding stated as an unproven assumption requiring a caveat; in both, it is the stated motivation for the method.

The apparatus above applies without modification, and the case exposes a further difficulty the Forth case does not. The Deficit Proxy Theorem's special case depends on faithful encoding, that every distinction costs at least one unit of representational capacity. Distributed and superposed representations, of the kind large trained networks are now widely understood to use, in which a single direction in activation space participates in encoding more than one feature, and a single feature is encoded across more than one direction, are precisely representations for which faithful encoding need not hold. Where it fails, even the weakened special case no longer applies, and a reduction in a representation's measured description length carries no guaranteed relationship to its distinguishing capacity at all, in either direction. The inference from compression to understanding is, in this setting, not merely unsupported past some extreme; it is not obviously licensed anywhere along the range where trained models actually operate.

15.9 Overcompression and Catastrophic Identification

It is worth naming directly what happens at the point where the divergence between the two deficits becomes irreversible. Suppose $x_1 \neq x_2$ but an encoding maps both to the same representation. Compression has crossed an epistemic boundary at that pair: unless infor-

mation external to the encoding remains available, the distinction cannot subsequently be reconstructed. Every compression scheme implicitly asserts that the distinctions collapsed within its induced equivalence classes do not matter, whether or not that assertion was made deliberately. The question worth asking of any compressed representation is accordingly not how much was compressed but which differences the compression declared equivalent, a question about $\mathcal{P}_T$, not about L_T.

Definition 15.6 (Task-relative equivalence and overcompression)**.** For a task $\tau : X \to Y$, define $x \sim_\tau y \iff \tau(x) = \tau(y)$. An encoding E preserves the distinctions τ requires when $E(x) = E(y) \implies x \sim_\tau y$, equivalently $\sim_E \subseteq \sim_\tau$. E overcompresses relative to τ if there exist x, y with $E(x) = E(y)$ but $\tau(x) \neq \tau(y)$.

A statistic $S(X)$ can throw away enormous quantities of raw information while retaining everything necessary for inference about a specified parameter θ; if S is sufficient for θ, it does not overcompress relative to θ, however much it overcompresses relative to X itself. Compression, on this reading, is not opposed to understanding; it is meaningful only relative to a prior decision about which distinctions must survive. The failure this section names is not compression itself but overcompression relative to some further task the representation is silently expected to also serve: a representation with $\sim_E \subseteq \sim_\theta$ mistaken for one with $\sim_E \subseteq \sim_X$.

15.10 Curricula Must Widen

A further difficulty for any theory that identifies understanding, or interest, or progress with monotonic compression is that the environments in which learning actually occurs are not found in a state of nature; they are built, and built deliberately simple, for reasons that have nothing to do with the learner's understanding having reached some genuine informational floor. Children are not raised inside the full chaos of an unfiltered natural environment; they are raised inside environments deliberately constrained to present regularities slowly, precisely because a learner who cannot yet compress much of anything will fail to learn from an environment offering nothing but unstructured, simultaneous complexity. The same engineering shows up outside pedagogy for reasons that are economic rather than developmental: standardized building materials constrain the built environment to a narrow vocabulary of regular shapes because mass production rewards uniformity, not because simple, rectilinear spaces are closer to the true shape of anything; roads are built straight because the vehicles traveling them have limited suspension travel, not because the terrain they cross is actually simple.

In each case, the low entropy of the environment is a scaffold fitted to the current, limited distinguishing capacity of whatever is operating within it, not a report on the true complexity of the domain. A scaffold is only useful for as long as the learner has not yet outgrown it. Once a learner is no longer acquiring anything new from a fixed training environment, further progress requires increasing the entropy of what is presented, widening $D_T(x)$ toward the distinctions the scaffold had been withholding, not continuing to compress what is already compressed as far as it usefully can be. A curriculum that only ever simplifies, and never widens, terminates in a plateau rather than in understanding. This is difficult to reconcile with a theory that identifies the drive toward understanding with compression progress at every point along the learner's trajectory, since the trajectory that theory describes is one of steadily increasing compression, while the trajectory that actually produces continued learning alternates compression with deliberately reintroduced complexity.

A related and sharper difficulty appears within problem-solving itself, on timescales far shorter than a curriculum. Solving a problem often requires making it more compli-

cated for a while before it can be made simple: introducing an auxiliary variable that has no simplifying role on its own but opens a path to one, or embedding a problem in a higher-dimensional space where a solution becomes visible before projecting the result back down. A compression-progress account, tracking description length as it evolves along the solver's trajectory, would register this intermediate expansion as a local reversal precisely at the point where the solver is doing the genuine work that will shortly produce the solution.

Developmental research on the explore-exploit trade-off supplies independent, empirically grounded support for the same structural claim: the distinctive length of human childhood is itself argued to be an evolved solution to a tension between exploring broadly and exploiting what is already known, with early cognition characterized by wide, high-entropy hypothesis search progressively narrowed only as the demands of goal-directed exploitation increase. On this account childhood is not an inefficient, not-yet-compressed version of adult cognition working its way toward an adult's compression; it is a distinct cognitive strategy favoring the preservation of distinctions a purely exploitative learner would discard, for exactly as long as the developmental trade-off favors exploration over efficiency.

15.11 Understanding Under Intervention

Measuring distinguishing capacity directly rather than inferring it from description length is still, as stated so far, a claim about a static quantity computed once against a fixed domain. A stronger and more operational version asks not how many distinctions a representation preserves at rest, but which distinctions survive when the domain is deliberately perturbed. Write $D_T(x \mid I)$ for the distinguishing capacity of T at x after an intervention I, and define the intervention robustness $R_T(x, I) = D_T(x \mid I) - D_T(x)$. Two representations with identical description length, or even identical distinguishing capacity measured before any intervention, can behave completely differently under I:

$$L_{T_1}(x) = L_{T_2}(x) \text{ and } D_{T_1}(x) = D_{T_2}(x) \;\not\Rightarrow\; D_{T_1}(x \mid I) = D_{T_2}(x \mid I).$$

A geometric construction examined in Chapter 18, in which rotational equivariance follows from an algebraic identity of the geometric product rather than being an empirical regularity discovered by training on many rotated examples, supplies a motivating case that is unusually clean because the claim being tested is architectural rather than statistical: intervention robustness under rotation is there provable from the representation's algebra, not merely measured empirically. A representation's distinguishing capacity is revealed by which distinctions survive admissible transformations of the domain, not merely by how many bits are required to encode the domain's resting state.

15.12 What Should Be Measured Instead

Proposition 15.4 establishes that coding length and distinguishing capacity can come apart across their entire joint range except at one identified point, and the Forth and machine-learning cases show the gap is not merely possible but ordinary. A practice that measures the first while reporting conclusions about the second is measuring the wrong quantity almost everywhere it is applied. None of this is an argument that compression is irrelevant to understanding. A system with a smaller coding deficit is, all else equal, a more efficient one, and efficiency is its own legitimate goal. The correction is a separation of that virtue from claims about understanding, which should be evaluated against the distinguishability deficit directly wherever the two quantities can be independently assessed: does the representation actually preserve the ability to separate the objects, cases, or outcomes that matter for the task, rather than merely occupying fewer bits while doing so.

A useful example already appears in evaluation practice for text summarization. Systems designed to produce very short summaries confront the obvious fact that brevity can preserve a coarse overview while eliminating distinctions present in the source; some summarization work supplements automatic overlap measures with separate human judgments of completeness and cohesion, rather than treating the automatic score as sufficient. The same system's central mechanism supplies a more literal instance of this chapter's apparatus: a summarizer that selects short introductory sentences and uses them as pointers into the rest of the source document exhibits, in the reach a pointer sentence makes available under a trained selection network, exactly the non-identification Proposition 15.4 describes: a short pointer sentence can have a large reach given a sufficiently capable selector, and a longer one can have almost none. The visible sentence is L_T; the trained selector is M_T; what the system is actually engineered to get right is the reach, which the sentence's own length does not report. Three routes past the difficulty recur throughout this chapter: ask what library of prior, admissible distinctions a representation's shortness is reachable from, and whether that library actually covers the distinctions the task requires; ask whether a representation's apparent compactness survives an audit for superposition and faithful encoding before treating that compactness as evidence of anything beyond itself; and, most demanding of the three, perturb the domain and ask which distinctions survive the perturbation, rather than trusting a static count taken at rest.

Chapter 16

Compression After Expansion

16.1 The Illusion of Effortless Form

Chapter 15 established that description length does not identify distinguishing capacity, and that a short representation's brevity is frequently evidence of accumulated background machinery rather than proximity to an informational floor. This chapter asks where that machinery comes from, and what happens to the observer, and to the person exercising a compressed competence, once the process that built it is no longer visible.

There is a persistent pattern in the way human beings perceive completed artifacts. A finished building, a published proof, a released piece of software, a polished essay, each presents itself as a coherent, apparently simple object whose internal structure invites little reflection on the process by which it came to be. Observers encountering such objects tend, with some consistency, to draw inferences about the difficulty of producing them that track the apparent simplicity of the result rather than the generative complexity of the process. This inference is often badly calibrated. It is not, however, a universal law: some finished objects really were easy to make, and surface simplicity sometimes does correlate with low production cost. The claim developed here is about a common and consequential bias, not an exceptionless regularity.

Compression is a constitutive feature of representation: the point of producing a final form is often precisely to eliminate from view the intermediate steps, failed attempts, and structural discoveries that preceded it. A proof that displayed every exploratory conjecture, failed lemma, and notational revision through which it was produced would be unreadable. But it can be epistemically misleading when the compressed form is mistaken for the primary form, and when the prior expansion is forgotten or invisible.

16.2 Expansion, Compression, and a Qualified Formal Model

A minimal formal vocabulary states a narrow, conditional version of the asymmetry between construction and compression. Let a system $\mathcal{S}$ be a configuration of components subject to a constraint set $\mathcal{C}$; a configuration s is coherent if $s \models \mathcal{C}$. Let Ω denote the space of possible configurations, and $\Omega_{\text{valid}} = \{s \in \Omega : s \models \mathcal{C}\}$. Cardinality is the wrong instrument for comparing this set to Ω whenever Ω is infinite, so a measure μ appropriate to the system is fixed instead.

Assumption 16.1 (Sparse validity). For the systems considered in this chapter, $\mu(\Omega_{\text{valid}})/\mu(\Omega) \ll 1$.

This is not asserted as a law governing all constrained systems; it is a condition satisfied by many of the domains discussed here, working software, load-bearing structures, valid

proofs, and the results that follow are conditioned on it holding. Expansion is a process E acting on an initial configuration, generating a sequence of transformations, trials, or variations; it is epistemic as much as generative, revealing structure in Ω_{valid} through interaction with $\mathcal{C}$. Expansion need not be performed by the agent who later benefits from it; it can be carried out by another person, an institution, a culture, or an earlier generation of a system, and its results transmitted rather than rediscovered. Compression is a triple $(C, D, \sim)$, an encoder, decoder, and task-relevant equivalence relation such that $D(C(s)) \sim s$, possibly subject to bounded distortion; it eliminates redundancy while preserving task-relevant structure, encoding invariants discovered during expansion and suppressing intermediate derivations not required to reconstruct behavior up to $\sim$.

Proposition 16.2 (Conditional dependency of compression)**.** *In an initially unknown task domain, constructing a reliable representation $(C, D, \sim)$ generally requires information about which variations preserve $\sim$-relevant validity. Absent an external source of this information, prior search by the agent, observation, instruction, or an inherited body of structural knowledge from elsewhere, such information must be acquired through exploration of Ω near the boundary of Ω_{valid}.*

A compression that removes components while preserving $\sim$ requires knowing which components are invariant under $\sim$ and which are incidental; where this knowledge is not supplied externally, it can only be established by observing which variations preserve validity and which do not. The practical corollary is narrower than a universal claim: one cannot, without access to some prior exploration, one's own or someone else's, produce a reliable compressed representation of a domain one does not yet understand.

This licenses a limiting case worth naming directly, because it clarifies a distinction the rest of this chapter depends on: between an artifact's observable fluency and the developmental process, if any, that an agent using that artifact has itself undergone. Consider an agent whose entire competence in a domain consists of an inherited representation, where the agent performed none of the exploration that produced it. Large language models are the clean paradigm case. A model's fluent, low-latency completions are a compression over regularities discovered through the effortful, iterative, error-correcting production of the human text on which it was trained, genuine expansion, performed by many agents over a long history. The model's forward pass exhibits the surface signature of relegated performance, speed, fluency, the absence of visible deliberation, but nothing in that forward pass corresponds to an agent's own trajectory from deliberate monitoring to earned stabilization, because no such trajectory occurred for the system producing the output. Call this *inherited compression*: a representation transmitted to an agent that performed none of the exploration behind it, as opposed to relegation, where the agent that now executes automatically is the same agent that once explored deliberately.

A related but distinct case is *conditioning*: stabilization that is genuinely the agent's own, but whose triggering exposure was authored by someone else, as when repeated pairing engineers a consumer's automatic response to a brand. The consumer possesses the attentional machinery relegation requires, but the stabilization was not produced by deliberate practice at solving a task of the consumer's own. Treating fast, fluent output as evidence of relegated processing therefore risks a category error: it mistakes a property of an artifact's presentation for a property of the process that generated it. The relevant questions are whose trajectory produced the underlying stabilization, by what mechanism, and for whose objective, the agent's own task in genuine relegation, another agent's prior task transmitted rather than repeated in inherited compression, or a third party's objective imposed on the agent's own stabilization machinery in conditioning.

16.3 A Probabilistic Asymmetry Between Construction and Disruption

Proposition 16.3 (Perturbation asymmetry under sparse validity)**.** *Suppose sparse validity holds, and let perturbations of a valid state be drawn from a distribution over Ω that is not concentrated on validity-preserving transformations. Then a randomly sampled perturbation of a valid configuration is more likely to leave Ω_{valid} than to remain within it.*

This is a genuine probabilistic consequence of sparse validity together with a stated assumption about the perturbation distribution; it is not, by itself, a claim that evaluation is easier than construction, and the two should not be run together. Detecting that a constraint has failed can itself be expensive, requiring measurement, proof, or diagnosis that is far from trivial. Breaking a system, detecting that it has broken, explaining why, and repairing it are four distinct operations with different cost profiles. What the proposition does support is narrower and still useful: under sparse validity and a perturbation distribution not aimed at preserving coherence, maintaining a system in Ω_{valid} requires selecting from a comparatively small set of validity-preserving operations, while an unstructured perturbation is likely to exit that set.

Because compression encodes a system's behavior only up to $\sim$, compressed representations are fragile in a specific sense: perturbations that move a configuration outside the region the encoder was calibrated against can produce invalid decodings, and interpretability of the compressed form depends on the receiver's prior familiarity with the domain. This grounds a further point about evaluative ease. Agents whose primary interaction with a system is evaluative, critics, reviewers, inspectors, users, typically interact with a system already inside Ω_{valid}, and their task is comparatively local: identify some perturbation that would move the system into Ω_{invalid}. Under sparse validity, this local task samples from a different, and typically larger, region of possibility space than the task of constructing or maintaining the system in the first place. The apparent ease of finding a flaw is therefore not, by itself, evidence that avoiding that flaw during construction was equally easy.

16.4 Skill as Internalized Structure: Aspect Relegation Theory

The standard account of expertise as accumulated knowledge and trained performance is underspecified in a way that limits its ability to explain some of the most striking features of expert cognition: the disappearance of felt effort, the compression of explanation, and the gap between expert and novice perception of the same situation.

Definition 16.4 (Task, known aspects, active attention)**.** Let a task $T = \{a_1, \ldots, a_n\}$ be a finite set of aspects, each corresponding to a constraint, sub-operation, or dependency relevant to successful execution. Let $K(t) \subseteq T$ be the known aspects the agent has represented at time t, whether or not currently monitored. Let the active attention set $A(t) \subseteq K(t)$ be the aspects under conscious monitoring at time t, subject to a bounded capacity $|A(t)| \leq k$.

An aspect a_i is stabilized to level $1 - \varepsilon$ under policy π and environment class $\mathcal{E}$ if

$$\Pr(\text{constraint satisfied} \mid a_i, \pi, \mathcal{E}) \geq 1 - \varepsilon :$$

stabilization is probabilistic and context-relative rather than binary, which is why expert automaticity transfers unevenly across contexts. The attention update rule is

$$A(t+1) = \big(A(t) \setminus R_t\big) \cup N_t \cup Q_t,$$

where $R_t \subseteq A(t)$ is the set of aspects that have met their stabilization threshold and are relegated out of active attention, N_t is a set of newly attended higher-order aspects that become tractable once capacity is freed by R_t, and Q_t is a set of aspects, possibly previously relegated, reactivated by uncertainty, novelty, unfamiliar context, or detected error. Learning is not monotonic under this rule: as lower-level operations stabilize, new higher-order aspects can enter via N_t, and previously stabilized routines can be recruited back via Q_t when they fail outside their training distribution. Skilled performance typically retains nonzero attention at some level, goal selection, anomaly monitoring, strategic coordination, and is better modeled as a change in what is monitored than as the disappearance of monitoring altogether.

The dual-process literature distinguishes rapid, autonomous, low-working-memory-load processing from processing that supports hypothetical reasoning and loads working memory. A substantial and well-supported class of expert Type 1 performance consists of procedures that once required controlled attention, were represented in $A(t)$ and governed by explicit monitoring, and became increasingly autonomous through practice, chunking, and probabilistic stabilization. This class does not exhaust Type 1 cognition, which also includes innate, perceptual, affective, associative, and implicitly acquired processes with no comparable history of explicit control; automatic higher-cognitive processes do not arise exclusively from a history of conscious skill acquisition, and motor-learning research on explicit and implicit training finds that both routes can lead toward automatic, dual-task-robust performance, with some paradigms actually favoring implicit acquisition. Where an expert Type 1 performance does belong to the relegation class, what appears as immediate intuition is, for that performance specifically, the residue of prior deliberation; the claim does not generalize to Type 1 cognition as a category.

A candidate neurophysiological correlate, transient hypofrontality in flow states, is worth stating as a contested hypothesis rather than an established signature. If flow involves reduced explicit supervision of well-stabilized sub-operations, some forms of flow may exhibit reduced activity in neural systems associated with self-monitoring, alongside preserved or increased activity in task-relevant control systems. Experimental work complicates a simple prefrontal-withdrawal story considerably: flow has been found associated with decreased activity in the medial prefrontal cortex and amygdala but increased activity in the inferior frontal gyri, the anterior insula, the basal ganglia, and the midbrain, a pattern of selective reallocation among frontal regions rather than uniform prefrontal withdrawal, and contemporary reviews emphasize network-level reorganization over any single mechanism of blanket deactivation. Reduced deliberative supervision can arise because supervision is no longer needed, a stabilized aspect continues to satisfy its constraints unattended, or because supervision is impaired while it is still needed, as under acute stress; these two conditions should not be treated as producing indistinguishable neural signatures, since acute-stress effects on executive function vary with severity, timing, and task in ways the flow literature does not report.

16.5 Borrowed Intuition: Provenance and Formation

The relegation model developed above describes competence built through an agent's own trajectory. A companion analysis is needed for competence that arrives already compressed, because the person exercising a judgment need not be the person whose exploration produced it. Call the resulting capacity *borrowed intuition*: rules of thumb, diagnostic categories, precedents, checklists, and trained models can transmit the residue of earlier search without transmitting the search itself.

Two dimensions must be kept distinct. Provenance asks whether a compression is earned,

its selection and stabilization depending on the receiving agent's own trajectory, or transmitted, the relevant selection having already occurred elsewhere and reaching the receiver in compressed form. Formation asks whether the process by which a competence is acquired is self-directed, the receiving agent controlling which situations to investigate and which objectives govern revision, or conditioned, another agent or institution structuring the receiver's exposures, feedback, or permissible actions. These dimensions are independent: an externally conditioned process can still produce earned compression if the receiver acts, observes consequences, and stabilizes a policy from those consequences, while self-directed study can remain largely transmitted if the agent chooses which authorities to consult but does not reproduce or test the inquiry underlying their conclusions.

Definition 16.5 (Borrowed intuition)**.** A competence C_i is borrowed to the extent that its present economy depends upon transmitted compression, and the receiver can exercise some practical advantage of C_i without reproducing the generating trajectories.

On this definition, nearly all mature human competence contains some borrowed component; the concept is not intended to divide agents into independent and dependent knowers, but to identify a variable ordinarily hidden by successful performance. Borrowed intuition can enter at several stages of a trajectory: the distinctions available, the subset activated by a situation, the policy governing movement through the activated space, or the relegated dispositions themselves. A diagnostic heuristic does not merely enlarge what an agent knows; it can make certain features salient, demote other possibilities, and change the order in which hypotheses are tested, structurally different from adding a proposition to what the agent knows.

Transformation of inherited compression through later experience requires a further, careful distinction. *Synthetic compression* is the process by which an agent applies, tests, combines, revises, or re-stabilizes transmitted compression through evidence generated in its own subsequent trajectory; it is a process rather than a fifth provenance class, describing movement within the taxonomy rather than another cell beside earned and transmitted. A procedure does not become earned merely because an agent has performed it many times: repetition under invariant conditions may strengthen a disposition while providing little evidence about why it works or whether an alternative would perform better. A post-acquisition trajectory is discriminating only when it changes the receiver's relative support over the candidate policies compatible with the inherited instruction; experience that merely confirms what every candidate policy predicts may stabilize behavior without clarifying its basis. This yields a chain of non-implications:

$$\text{temporary conditioning} \not\Rightarrow \text{retained experience} \not\Rightarrow \text{stabilized compression.}$$

Proposition 16.6 (Provenance is not erased by synthesis)**.** *If a transmitted policy undergoes synthetic compression through a receiver's later trajectory, the resulting competence may acquire earned support without ceasing to depend historically upon transmitted structure.*

Unless the later trajectory independently reconstructs every distinction and selection materially contributed by the original transmitted policy, the resulting competence remains causally dependent upon external compression. Synthetic development can alter the degree and significance of that dependence, but it does not make the originating contribution disappear. Borrowed intuition is therefore neither a permanent deficiency nor a stage every competence must leave behind; a mature competence may be deeply owned, critically understood, and extensively validated while remaining partly borrowed, which is likely the normal condition of expertise.

16.6 Pretrained Systems as a Limiting Case

A frozen pretrained language model provides an unusually clean case of transmitted compression. Its parameters are selected through optimization over a corpus generated by trajectories external to the model; the resulting system can produce fluent continuations, classifications, and explanations without having participated in the investigations, conversations, and acts of composition from which the corpus arose. A description of an experiment is not the experiment; a record of a disagreement is not participation in the disagreement. The model nevertheless benefits from those histories: regularities distributed across the corpus become available through a compressed parameterization,

external exploration $\rightarrow$ recorded corpus $\rightarrow$ parameter compression $\rightarrow$ present competence.

A purely pretrained model with frozen parameters, no persistent memory, and no post-training interaction is a limiting case of borrowed intuition. Almost all of its stable competence is transmitted in provenance and conditioned in formation. This claim does not depend on a verdict about consciousness; it follows from the causal organization of the training pipeline. Even if one attributed understanding or some form of subjectivity to the resulting system, the provenance question would remain: a capacity can be real while the experience that made it economical belongs elsewhere.

The limiting case becomes less clean once post-training and deployment are included. Instruction tuning and reinforcement learning from human feedback alter a pretrained policy using demonstrations, preferences, and evaluations selected by additional human processes, introducing new trajectories into the model's formation without automatically making the resulting compression self-directed or earned; from the perspective of provenance, this remains largely conditioned, since humans construct prompts, rank outputs, and determine the objective through which the model is updated. Whether it also counts as earned compression depends on the boundary used to identify the agent: if the training process and the deployed model are treated as one temporally continuous system, post-training contains limited cycles of action, feedback, and policy revision; if each frozen checkpoint is treated as a distinct receiver, the deployed model inherits the result of post-training just as it inherits pretraining.

This distinction is especially important for in-context learning. A model can alter its responses when examples or instructions are supplied in the prompt, without gradient updates to its parameters. Such adaptation may be substantial within the active context, but it ordinarily disappears when that context is removed, and is therefore better described as temporary conditioning unless the context or its effects are retained. Representing a system at time t as $M_t = (\theta_t, m_t, \chi_t)$, persistent parameters, persistent external memory, and current context, the system has a retained post-acquisition trajectory only if an observation can alter a later state, $(\theta_{t+1}, m_{t+1}) \neq (\theta_t, m_t)$, through a causal update attributable to that observation. A change confined to χ_t can modify present behavior without becoming persistent experience: the system adapted, but it did not retain the adaptation. This is the exact formal counterpart of the state-versus-context distinction Chapter 10 draws between operative state σ and documentary history H: a documentary transition that extends χ_t without touching (θ_t, m_t) is precisely CLIO's identity transition on operative state.

16.7 Skill as Compressed Structure

Relegated aspects encode structural invariants discovered through prior practice, and the felt ease of expert performance corresponds to a small active attention set relative to the full task, $|A(t)| \ll |T|$, rather than to any reduction in the task's underlying complexity. That a

task feels easy to an expert is evidence that the expert has performed enough exploration to stabilize its recurring structure, not evidence that the task is simple. Relegation introduces a partial loss of reconstructability: once aspects have left $A(t)$, intermediate steps may no longer be explicitly representable, and expert knowledge is often operationally accessible but descriptively compressed, the expert can perform the task while struggling to narrate how. This explains why expert advice is so often correct for the expert and useless for the novice: the relegated aspects are no longer visible to the advisor, so the advice skips steps that are not yet stabilized for the listener.

In commercial and media contexts this becomes more troubling, because compressed outputs are frequently presented as though no prior exploration were required to reach them, when the audience's lack of background is precisely what the compression depends on to look effortless. There is a useful distinction between compression that is reconstructible by a prepared receiver and compression that is parasitic on undisclosed prior knowledge, what might be called *epistemic outsourcing*: achieving an appearance of simplicity not by encoding structure efficiently but by silently delegating the exploratory work to a receiver who is not equipped to perform it and is not told they must.

16.8 Fundamental Attribution Error and the Erasure of Process

This misperception is amplified by a well-established feature of social cognition. The fundamental attribution error describes a systematic tendency to attribute observed outcomes to dispositional properties of agents, talent, intelligence, character, while underweighting situational and processual factors. Applied to skill and production, a successful outcome tends to be attributed to the producer's intrinsic qualities rather than to the structural conditions and accumulated exploration that made it possible. This pattern is not irrational given the available evidence, outcomes are visible, processes typically are not, but the environment is incomplete, and the resulting attributions inherit that incompleteness.

A related distortion operates at the level of population inference. In many domains of skilled production, outcome distributions are heavily right-skewed: a small number of participants achieve highly visible success, and a large majority do not achieve comparable recognition regardless of comparable effort. Canonization narratives constructed retrospectively tend to read as stories of difficult genius eventually rewarded, a reconstruction that can suppress the contingency and selection effects that separated the canonized case from equally capable contemporaries whose recognition never arrived.

The asymmetry formalized in Proposition 16.3 has a direct bearing on this pattern. Where the apparent ease of locating a flaw is read as evidence that avoiding the flaw during construction was equally easy, the inference overreaches what the proposition actually supports, since detecting or explaining a failure is not always cheap. What the formal result does support is narrower: construction and evaluation typically sample from different, and often differently sized, regions of the space of possible operations, so evaluative ease is not on its own strong evidence about constructive difficulty, in either direction.

16.9 Labor and the Political Economy of Delegated Constraint Management

The preceding analysis converges on an asymmetry in how different categories of labor tend to be perceived and rewarded. Essential work, infrastructural, repeatable, stabilizing labor such as maintenance, care, teaching, and construction, often manifests as the absence of failure rather than the presence of a discrete success, and can go relatively unnoticed even

when it sustains the conditions under which other, more visible production becomes possible. Wealth confers, among other things, the capacity to delegate constraint management: supply chains, maintenance schedules, and coordination problems can be handled by others and rendered invisible to the person delegating them. This delegation is not inherently problematic, but its epistemic consequence is a tendency rather than a certainty: delegation can produce constraint blindness, a reduced capacity to perceive the structural conditions that make an apparent solution non-trivial, and where it does occur, it tends to produce a familiar genre of compressed advice, "just do this," that presents a compressed form presupposing an exploration the listener must still perform but has not been told about.

16.10 The Ethics of Compression

The strong standard, that an ethically compliant compressed representation must preserve, in principle, the reconstructibility of its actual generative pathway, is too demanding and in some cases incoherent: a finished mathematical proof need not preserve the psychological history of its discovery. Three distinct standards should be kept apart. Logical reconstructibility asks whether a prepared receiver could, in principle, derive or verify the compressed content from stated premises, independent of how it was actually produced. Practical teachability asks whether the compressed form gives an unprepared receiver a workable path toward the competence it represents. Historical provenance asks whether the compressed form accurately represents the actual process, whose labor, how much exploration, over what period, that produced it.

Requiring all three uniformly is too strong; mathematics regularly and legitimately satisfies logical reconstructibility while providing neither teachability nor provenance, and this is not an ethical defect in a proof.

Definition 16.7 (Non-deceptive compression)**.** A compressed representation is non-deceptive if it does not affirmatively misrepresent the exploration, labor, or prerequisites that produced it, whether or not it also satisfies logical reconstructibility or practical teachability.

Three working criteria follow from this narrower standard: non-deception about labor, representations should not imply, where it is false, that no meaningful prior exploration was required; appropriate matching of standard to context, a representation should be evaluated against the standard it actually claims to meet, not against all three at once; and accuracy of effort distribution, representations should not systematically imply that the observed outcome is more typical of the underlying attempt distribution than it is.

16.11 Conclusion

Under a stated sparse-validity assumption and a stated assumption about perturbation distributions, maintaining coherence in a system requires selecting from a comparatively narrow set of operations, while an unstructured perturbation is likely to leave that set by an easier route; this is a conditional probabilistic result, not a universal entropy law. In human skill, a substantial and well-evidenced class of expert automaticity arises from procedures that were once deliberately controlled and became increasingly autonomous through practice, though this class does not exhaust automatic cognition. Observers who encounter only a compressed result, without its generative history, are consequently prone, not certain, but prone, to underestimate the exploration that produced it. And where the compressed result reaches an agent who never performed that exploration at all, borrowed intuition names

the resulting gap between operational competence and its formative history precisely, without treating the borrowing as either counterfeit or as a deficiency every competence must eventually shed.

This is the same asymmetry Chapter 15 located in Forth's word libraries and in trained models' latent codes, examined here from the side of the observer and the borrower rather than the side of the representation itself: a short description is evidence of accumulated machinery, and a fluent performance is evidence of a stabilization history, but in neither case does the visible artifact report whose history it was, or how much of it remains available for the receiver to inspect, revise, or trust.

Chapter 17

Sparse Recursive Holographic Steganography

17.1 Introduction

Chapters 15 and 16 examined representations whose brevity conceals the machinery, or the exploration history, that makes them work. This chapter examines a representation problem of a different shape: not how to compress a payload, but how to distribute it so that no single location in a carrier is uniquely responsible for any fragment of it, and partial observation degrades gracefully rather than failing at a fixed boundary. The architecture developed here, Sparse Recursive Holographic Steganography (SRHS), replaces the standard one-pass map from message bits to carrier coefficients with three coupled operations: holographic payload projection, sparse carrier routing, and recursive residual correction.

Steganographic systems are conventionally organized around a spatial metaphor. A payload is partitioned into symbols; symbols are assigned to pixels, blocks, tokens, frames, or transform coefficients; a decoder later reads those assignments back. This organization produces three linked weaknesses. Locality produces brittle failure, since destroying the region assigned to a payload block destroys that block regardless of how much unused evidence remains elsewhere in the carrier. A one-pass encoder commits to its full modification before it can observe the realized consequences of any of it, and carrier channels are heterogeneous enough that post-embedding transformations are only partially predictable in advance. Dense modification, finally, wastes distortion on locations that contribute little reliable information while enlarging the statistical surface available to a detector.

The architecture proposed here replaces symbol placement with field construction. The payload is projected into a set of correlated measurements, each individually insufficient but jointly informative about the payload as a whole. These measurements are embedded into a sparse subset of carrier regions. After each embedding step the system decodes its own intermediate output, estimates the residual payload error, and selectively revises the carrier. Computation is therefore recursive, while modification remains sparse at every step. The term *holographic* is used operationally rather than by analogy to physical optics: it denotes distributed recoverability, in which partial observations preserve degraded evidence about the whole message rather than exact evidence about isolated fragments of it. The term *recursive* likewise denotes a concrete recurrence with shared parameters and an explicit stopping rule, not literal self-similarity or unbounded depth.

17.2 Problem Formulation

Let a cover object be $x \in \mathcal{X}$, drawn from a declared cover distribution $\mathcal{D}_{\mathcal{X}}$; let a payload be $m \in \{0,1\}^L$; and let a secret key be k, drawn from a keyspace large enough to resist exhaustive search. An encoder produces $y = E_\theta(x, m, k; B)$, where y is the stego object and B is a hard distortion budget. A channel transformation $T \sim \mathcal{T}$ may crop, compress, rescale, transcode, paraphrase, drop packets, reorder frames, or otherwise alter the object before it reaches the receiver. The decoder returns $\hat{m} = D_\phi(T(y), k)$. A useful steganographic system must jointly minimize decoding error, perceptual or semantic distortion, and detectability. For a detector A, define its advantage under equal priors as $\mathrm{Adv}(A) = |\Pr[A(y) = 1] - \Pr[A(x) = 1]|$. Every result below is stated relative to a declared $(\mathcal{D}_{\mathcal{X}}, \mathcal{T}, B)$ triple, and no claim is intended to hold outside such a declaration, a discipline this chapter shares with the falsifiable-claims requirement Chapter 14 imposed on the deployment-native training architecture.

17.3 Holographic Payload Projection

A localized representation maps payload block m_i primarily to a single carrier region j. SRHS instead maps the entire payload to a redundant collection of keyed measurements. An outer error-correcting code first produces $c = C(m) \in \{0,1\}^n$, centered as $\bar{c} = 2c - \mathbf{1} \in \{-1,+1\}^n$ to remove a codeword-dependent DC component, and a keyed linear projection generates

$$z = \frac{1}{\sqrt{n}} P_k \bar{c}, \qquad P_k \in \mathbb{R}^{q \times n},$$

followed by a bounded quantization $u = Q(z)$. The $1/\sqrt{n}$ normalization keeps the expected measurement energy independent of n. The rows of P_k are generated from a cryptographic pseudorandom seed and normalized so evidence is dispersed across measurements rather than concentrated in a few of them.

Proposition 17.1 (Restricted preservation)**.** *There exists a family of subsets $S \subseteq \{1, \ldots, q\}$, occurring with non-negligible probability under the transformations in $\mathcal{T}$, such that the restricted projection $P_{k,S}$ preserves the pairwise distances among codewords of C up to a bounded distortion factor $\delta(S) \in [0,1)$: for any two codewords $c_a \neq c_b \in C$ with centered forms $\bar{c}_a, \bar{c}_b$,*

$$(1 - \delta(S)) \, \|\bar{c}_a - \bar{c}_b\|^2 \leq \|P_{k,S}(\bar{c}_a - \bar{c}_b)\|^2 \leq (1 + \delta(S)) \, \|\bar{c}_a - \bar{c}_b\|^2.$$

This is a restricted-isometry-type condition, applied here to a finite codeword set rather than to arbitrary sparse vectors. Since $\bar{c} = 2c - \mathbf{1}$, codewords at Hamming distance d satisfy $\|\bar{c}_a - \bar{c}_b\|^2 = 4d$, so the minimum squared distance among centered codewords is $4d_{\min}$ for the outer code's minimum Hamming distance $d_{\min}$. Under Proposition 17.1, the minimum squared distance among the normalized measurement images is at least $(1 - \delta(S)) \cdot 4d_{\min}/n$, and nearest-codeword decoding over the surviving measurements S succeeds whenever the aggregate noise satisfies $\|\eta_S\| < \sqrt{(1 - \delta(S))\, d_{\min}/n}$. This margin gives a concrete design target: $\delta(S)$ can be estimated empirically on surviving subsets under the declared transformation distribution and compared against $d_{\min}/n$, a falsifiable numerical claim rather than an appeal to the assumption's plausibility.

Under this construction, the condition for recovery is not that every crop reconstruct the message exactly. It is that a sufficiently large family of measurement subsets retains enough rank, or a restricted-isometry-type property, to support decoding. For multimedia carriers, a view is any transformation that preserves a subset or mixture of carrier evidence: a crop, a temporal window, a frequency band, a resolution level, or a set of surviving packets;

training samples multiple views of the same stego object and requires each to predict the same payload, the learned counterpart of the projection's algebraic redundancy.

17.4 Sparse Recursive Embedding

Divide the carrier representation into N addressable regions, which may be spatial patches, wavelet bands, video tubes, audio time-frequency cells, text spans, or learned latent groups. At recursion step t the system maintains a stego candidate y_t, a residual payload state r_t, and a remaining budget b_t, initialized as $y_0 = x$, $r_0 = u$, $b_0 = B$. A controller with shared parameters ψ computes region scores $s_t = G_\psi(F(y_t), r_t, b_t, k)$, where F is a carrier feature extractor. A sparse router selects an active set S_t of at most $K \ll N$ regions; a proposal network produces a bounded revision $\Delta_t = U_\theta(F(y_t), r_t, k, S_t)$, and the carrier is updated only on the active set:

$$y_{t+1} = \Pi_{\mathcal{C}(x,B)}\big(y_t + M(S_t) \odot \Delta_t\big),$$

where $M(S_t)$ is the active-region mask and $\Pi_{\mathcal{C}(x,B)}$ projects the candidate back into the admissible distortion set around x. The system then performs an internal decode, $\tilde{u}_{t+1} = D_\phi^{\text{inner}}(y_{t+1}, k)$, $r_{t+1} = u - \tilde{u}_{t+1}$. Recursion terminates when the estimated robust decoding loss falls below a threshold, when the budget is exhausted, or when a fixed maximum depth R is reached. Because G_ψ, U_θ, and the internal decoder share parameters across steps, increasing depth increases adaptive computation without increasing parameter count. Each step is simultaneously an intervention and a measurement: the encoder does not merely add signal to the carrier, it learns from what survived its own preceding attempt.

Sparsity without recursion commits irrevocably to a single, possibly mistaken, active set. Recursion without sparsity perturbs the entire carrier at every step and compounds detectable residue across steps. Their conjunction permits exploration followed by correction: early steps allocate weak measurements to regions predicted to be safe, an intermediate decode reveals which payload directions remain uncertain, and later steps address those directions through new regions or revisions of unreliable ones.

Definition 17.2 (Marginal routing value)**.** Let ℓ_t denote expected payload loss after step t and $c_{t,j}$ the cost of revising region j at step t. The marginal value of revising region j is

$$v_{t,j} = \frac{\mathbb{E}[\ell_t - \ell_{t+1} \mid j]}{c_{t,j} + \lambda\, d_{t,j}},$$

where $d_{t,j}$ estimates the increment to detectability contributed by revising j.

Sparsity is a consequence of most regions carrying low or negative marginal value at a given state, rather than a constraint imposed for its own sake. Sparse recursion additionally supports conditional depth: covers with low channel noise can stop early, while difficult carriers consume additional steps, converting embedding computation from a fixed per-object expense into an allocation problem solved per example.

17.5 Training Objective

Training jointly optimizes the projection-aware decoder, the sparse recurrent encoder, and a population of steganalyzers. For sampled transformations $T_1, \ldots, T_J \sim \mathcal{T}$, the payload loss is $\mathcal{L}_{\text{msg}} = \frac{1}{J}\sum_j \text{BCE}(m, D_\phi(T_j(y_R), k))$. The fidelity term combines a domain metric with a feature-space discrepancy, $\mathcal{L}_{\text{fid}} = d_{\mathcal{X}}(x, y_R) + \alpha\|\Phi(x) - \Phi(y_R)\|_2^2$. Security uses an ensemble of detectors, each trained to maximize discrimination, $\mathcal{L}_{A_i} = -\mathbb{E}_x \log(1 -$

$A_i(x)) - \mathbb{E}_{y_R} \log A_i(y_R)$, while the encoder is trained against the current ensemble using the non-saturating confusion loss $\mathcal{L}_{\text{sec}} = -\frac{1}{M}\sum_i \log(1 - A_i(y_R))$, which avoids the vanishing-gradient regime the saturating form produces early in training. Sparse, non-repetitive routing is encouraged by an ℓ_0-style penalty on the gate vector plus a cross-step redundancy penalty, and a view-consistency term requires partial views to agree on the decoded payload distribution via a symmetrized KL divergence. Training alternates two minimizations, one over the encoder and decoder parameters jointly with the security and consistency terms, and one over the detector ensemble, rather than a single min-max expression. The hard distortion constraint is enforced by the projection $\Pi_{\mathcal{C}(x,B)}$ regardless of whether the learned penalties are well calibrated, which prevents the optimizer from trading visible carrier damage for payload accuracy under an imperfectly weighted objective.

17.6 Analysis

17.6.1 Graceful Erasure Recovery

Because S is a discrete, bounded index set, recovery probability as a function of $|S|$ cannot be continuous in the analytic sense, and bounded-noise decoding can exhibit genuine thresholds at the outer code's correction radius. The claim is stated instead as monotone expected degradation, which is what the construction actually delivers.

Proposition 17.3 (Monotone degradation under erasure)**.** *Suppose Proposition 17.1 holds and the projected codeword measurements are embedded with independent bounded noise. Let S be a random subset of surviving measurements under a sampling model in which each measurement survives independently with probability π, and let* $\text{err}(S)$ *denote the expected outer-decoding error conditional on S. Then $\mathbb{E}_S[\text{err}(S)]$ is monotonically non-increasing in π, and no fixed payload block is dropped as a deterministic function of a predetermined carrier boundary, in contrast to a block-local scheme in which each payload block fails outright whenever its assigned region is erased, regardless of how much unused evidence survives elsewhere in the carrier.*

Restricting the projection to a larger surviving set weakly improves the preserved codeword distances, since additional independent measurements can only add information to the linear system, so $\delta(S)$ is non-increasing in $|S|$ in expectation under the sampling model, and the recovery threshold is non-decreasing as $|S|$ grows. Increasing π stochastically dominates the resulting $|S|$, so the probability of meeting the threshold, and hence the expected error, inherits monotonicity in π by composition. This result is conditional rather than absolute: an adversary who removes all informative regions, destroys the carrier outright, or estimates the key-dependent projection structure can still prevent recovery. Holographic distribution changes the shape of the failure curve; it does not remove the underlying channel-capacity limit.

17.6.2 Recursive Dominance Under Uncertain Response

Weak dominance of a two-step encoder over a one-step encoder follows immediately from policy-class containment: any feasible one-step policy is available to the two-step encoder as the choice of a zero second update, so the two-step optimum cannot be worse. Strictness is a value-of-information claim. Model the realized carrier response as an unobserved state h, drawn from a prior unknown to the encoder before any modification is rendered, and let the first-step update produce an observation o_1 informative about h.

Proposition 17.4 (Strict dominance via value of information)**.** *The two-step encoder's optimal expected loss is $\mathbb{E}_{o_1}[\min_a \mathbb{E}[\ell(a, h) \mid o_1]]$, and the one-step encoder's optimal expected loss is*

$\min_a \mathbb{E}[\ell(a, h)]$. *The former is no greater than the latter in general, and strictly smaller whenever* o_1 *has positive value of information for the allocation decision.*

The inequality holds in general because the right-hand side is the loss of a single action chosen without conditioning on o_1, feasible but not necessarily optimal for each realization of o_1; taking the expectation of the pointwise minimum cannot exceed the minimum of the expectation. Equality holds exactly when the optimal action under the prior remains optimal under every realization of o_1, that is, when o_1 carries no information relevant to selecting the action. This formalization makes explicit what recursion cannot buy: if the internal decode is uninformative about the realized channel response, for instance because the channel is deterministic and already known to the encoder, the value of information is zero and additional recursion steps yield no improvement in expectation, regardless of added computation. This is the discrete-communications analogue of Chapter 10's null-intervention principle: recursion, like suspension, is warranted only when the intervening observation is actually informative, and both frameworks refuse to treat additional computation or additional caution as valuable by default.

17.6.3 Sparse Modification and Detectability

Sparsity does not automatically imply security. A large change concentrated in a few regions can be easier to detect than a small change spread across many. The governing quantity is the divergence between the cover and stego distributions under the realized modification policy, not the cardinality of the active set as such. Sparse routing improves the frontier only when the controller correctly identifies high-capacity regions and respects per-region amplitude constraints.

Proposition 17.5 (Capacity bound)**.** *Let* $C_{\mathcal{T}}(B)$ *denote the effective capacity of the transformed carrier channel under distortion budget* B*. Reliable transmission of an* L*-bit payload requires, asymptotically and subject to the standard qualifications of channel coding,* $L \leq C_{\mathcal{T}}(B)$*.*

Projection redundancy, recursion, and learned routing cannot violate this bound; their function is to move the achievable operating point closer to it by using the available capacity more effectively, not to relax the bound itself.

17.7 Experiments and Falsifiable Claims

The principal comparison holds cover distribution, payload length, key length, distortion budget, training transformations, and compute reporting constant across conditions, comparing SRHS against traditional transform-domain embedding, a strong one-pass learned encoder, a dense recurrent encoder, and a sparse non-recurrent encoder. Primary outcomes are bit error rate after transformation, exact-message recovery rate, detector AUC, calibrated distortion, and encoding cost, together with the empirical distortion factor $\delta(S)$, estimated directly and compared against the decoding margin threshold, testing whether Proposition 17.1 actually holds at the operating point used elsewhere in the evaluation rather than being assumed to hold because the rest of the pipeline appears to work.

The decisive ablations remove one component at a time: holographic projection, outer error correction, sparse routing, recurrent residual correction, view consistency, and adversarial security training, with a further control replacing adaptive routing with randomly selected regions matched for sparsity and distortion. Three preregistered hypotheses define success: SRHS should lower transformed-channel bit error without increasing held-out detector AUC at matched distortion and payload; partial-view recovery should degrade

smoothly with retained evidence, consistent with Proposition 17.3; and recursion should improve the frontier beyond a parameter-matched one-pass model, consistent with Proposition 17.4. Failure is equally specified in advance: if gains disappear against detectors not used in training, the method has learned the training steganalyzers rather than steganographic security; if the projection helps only because it adds redundancy, a conventional code at matched rate is sufficient; if additional recursion steps offer no gain after compute matching, adaptive revision is unnecessary; if robustness is obtained only by exceeding the declared distortion budget, the central claim of the paper is false.

17.8 Security Model and Misuse

Three observers are distinguished, extending the classical warden model of covert communication. A passive warden attempts to decide whether a carrier contains a payload. An active warden applies transformations intended to destroy hidden communication while preserving ordinary utility of the carrier. A forensic adversary may collect many carriers, observe repeated key use, choose messages adaptively, or train a detector directly against a deployed encoder. SRHS is designed primarily for robust covert communication against passive and bounded active wardens; it does not by itself provide payload confidentiality or authenticity, and learned indistinguishability from a trained detector ensemble is not cryptographic security, since empirical detector failure is evidence relative to a declared adversary class, not proof against every possible test.

The same capabilities that support covert communication also support legitimate provenance marking, resilient metadata, synchronization signals, and ownership claims, and they can support malicious coordination or exfiltration. Responsible deployment should include rate controls, explicit authorization for the carriers used, auditable key management, and evaluation by independent steganalysts.

17.9 Conclusion

Steganography has largely been posed as a placement problem: determine where to put a secret while changing the carrier as little as possible. Sparse Recursive Holographic Steganography poses a different problem: determine how to construct a distributed evidence field that remains recoverable across partial observations, and then synthesize that field through a small number of adaptive carrier interventions. Error-correcting projection makes payload evidence global rather than local; sparse routing spends distortion where it carries the highest marginal value; recursive correction lets the encoder observe the realized channel response and revise its allocation without increasing parameter count. None of these components removes the underlying limits of channel capacity, detectability, or adversarial adaptation. If the stated hypotheses hold under the specified ablations, hidden information would no longer be identified with particular altered carrier locations; it would become a recoverable property of the carrier as a whole, weakly present in many views, strengthened by aggregation, and written through adaptive computation rather than fixed placement. The unit of steganographic design would then shift from the secret bit to the secret field, a shift that mirrors, in an adversarial communications setting, the shift Chapter 6 made from the state that persists to the continuation structure that survives transformation: in both cases, the unit worth tracking is not a located point but a distributed, recoverable pattern.

Chapter 18

Clifford FHE as the Inverse Compression Case

18.1 Introduction

Chapter 15 defined a representation's three coordinates, description magnitude L_T, distinguishing structure $\mathcal{P}_T$, and operational repertoire $\mathcal{O}_T$, and showed the first two are independent while the third is a genuinely separate axis a lossless archive can leave essentially empty. This chapter closes Part IV with the sharpest available instance of that third coordinate: a case, drawn from an external, independently authored paper on homomorphic encryption, in which two encodings agree completely on stored values and on static distinguishing capacity while disagreeing entirely on what can be computed over those values without first decrypting them. The reading below is offered in the same spirit as Chapter 8's treatment of Buchanan's contextuality paper: a structural correspondence recognized after the fact, not a claim that the source paper was written with this monograph's apparatus in mind, and every quantitative claim below is attributed to its source rather than adopted as this monograph's own.

18.2 The Problem: Flattening Discards Structure

Conventional fully homomorphic encryption (FHE) schemes support arbitrary computation on encrypted data without decryption, but the standard CKKS scheme, which supports approximate arithmetic on real numbers, is scalar-only: it has no native awareness of geometric relationships and requires flattening a structured object, such as an eight-component multi-vector, into eight separate, uncoordinated ciphertexts. Geometric algebra represents geometric objects through multi-vectors, elements combining scalars, vectors, bivectors, and higher grades within one unified algebraic framework. In the three-dimensional Euclidean geometric algebra $\mathrm{Cl}(3,0)$, a general multi-vector

$$M = m_0 + m_1 e_1 + m_2 e_2 + m_3 e_3 + m_{12} e_{12} + m_{13} e_{13} + m_{23} e_{23} + m_{123} e_{123}$$

combines a scalar, three vector components, three bivector components representing oriented planes, and one trivector component representing oriented volume, eight components in total. The defining operation is the geometric product $ab = a \cdot b + a \wedge b$, associative, distributive, and generally non-commutative, from which rotations are realized via rotor conjugation, $v' = RvR^{-1}$ (written $R \otimes v \otimes \tilde{R}$ in the source paper's notation), composing more compactly and more cheaply than matrix multiplication.

The source paper's central question is whether geometric data can be encrypted while preserving its geometric properties, rather than being flattened into scalar arithmetic that discards them. It answers with Clifford FHE, the first ring-learning-with-errors-based FHE scheme with native support for all seven fundamental Clifford algebra operations (geometric product, reverse, rotation, wedge product, inner product, projection, and rejection), and with geometric neural networks that operate directly on encrypted multi-vectors, replacing matrix multiplication with the geometric product as the network's fundamental layer operation.

18.3 Two Encodings, Identical Values, Different Repertoires

It is tempting to read the headline figure the source paper reports, a $250{,}000\times$ expansion between a 64-byte plaintext multi-vector and its roughly 16-megabyte encrypted form, as a direct instance of the coding-deficit gap Chapter 15 measures. That reading does not hold up. The expansion is ciphertext expansion, redundancy that ring-learning-with-errors encryption introduces for cryptographic security, not a comparison of description lengths in the sense $\delta_T(x) = L_T(x) - L^*(x)$ is defined in this monograph. A large ciphertext is not evidence of a large coding deficit, and the size figures should not be read as one.

The more interesting distinction survives this correction, but it belongs to the third coordinate of Chapter 15's apparatus, the operational repertoire $\mathcal{O}_T$, not to L_T or D_T. Two schemes for encrypting a multi-vector, ordinary flattened CKKS and Clifford FHE, both encrypt each of the eight components separately, eight ciphertexts either way; by the definitions of Chapter 15, D_T is unaffected by this choice, since the component values are equally present and equally distinct under both schemes. What differs is $\mathcal{O}_T$: Clifford FHE additionally encodes a fixed table of structure constants, $C_{ijk} \in \{-1, 0, 1\}$ specifying how the eight encrypted components combine under the geometric product, so that rotation, projection, and the wedge and inner products become native operations over the ciphertexts, computable directly rather than only after decryption. The homomorphic geometric product computes

$$\mathrm{Enc}(c_k) = \sum_{i,j} C_{ijk} \cdot \mathrm{Enc}(a_i) \cdot \mathrm{Enc}(b_j),$$

64 ciphertext multiplications per geometric product, against 512 for the equivalent naive 8×8 matrix encoding, an eightfold reduction in operation count that follows from the geometric product's structural sparsity, only 64 of the 512 possible term products are nonzero, rather than from any change in what is stored.

"Discards" overstates what ordinary flattened CKKS does to these relations: nothing about scalar CKKS makes a rotation mathematically unrecoverable from the eight ciphertexts, since the component values are all still there and could in principle be recombined by hand using ordinary scalar homomorphic operations. What ordinary CKKS lacks is the Clifford structure as a native, first-class part of $\mathcal{O}_T$: recovering a rotation from flattened ciphertexts requires re-deriving, term by term, an algebra that Clifford FHE simply supplies as an available operation. The source paper's own security proof makes the boundary of what is and is not preserved precise: Clifford FHE's security reduces to CKKS security, since each multi-vector component is an independently encrypted CKKS ciphertext, so an adversary breaking Clifford FHE with advantage ϵ can be used to break ordinary CKKS with advantage $\epsilon/8$. The two schemes are cryptographically equivalent at the level of what is hidden; they differ only in what is computable without first removing that hiding.

18.4 Clifford FHE as the Inverse Compression Case

Chapter 15's three coordinates, description magnitude, distinguishing structure, and operational repertoire, define a Pareto-style tradeoff space, and Clifford FHE occupies an extreme and clarifying point in it. Standard CKKS optimizes for description length by flattening a multivector into independent scalar ciphertexts, discarding, or more precisely relocating outside $\mathcal{O}_T$, the algebraic structure of the geometric product. Clifford FHE represents the exact inverse case. It abandons compression entirely, accepting the full $250{,}000\times$ ciphertext expansion the source paper reports, precisely to preserve the algebraic structure of geometric operations natively over the ciphertexts. Thus Clifford FHE is the most mathematically extreme realization of the compression fallacy's central thesis: to preserve native, first-class operational structure, a representation may need to sacrifice essentially all of the compactness that would ordinarily be read as evidence of good design. Representational quality is a genuinely multi-dimensional frontier, $R = L_T \times \mathcal{P}_T \times \mathcal{O}_T$, where size and structure can stand in direct, structural tension rather than trading off along one shared axis.

This also gives Chapter 15's intervention-robustness apparatus its cleanest available instance. Rotational equivariance in the geometric neural network described below is not an empirical regularity discovered by training on many rotated examples; it follows from the algebraic construction itself. For a rotor R and geometric-linear-layer weight W, applying $W \otimes -$ to a rotated input commutes with rotation by associativity of the geometric product and the properties of rotor conjugation:

$$W \otimes (RxR^{-1}) = R(W \otimes x)R^{-1},$$

so the layer's output rotates consistently with its input as an algebraic identity, independent of any particular learned weights. In Chapter 15's notation, $R_T(x, I) = 0$ for I a rotation is a claim provable from T's algebra rather than measured empirically on a held-out test set: a representation's distinguishing capacity, tested under deliberate intervention, is here revealed by proof rather than by observation, exactly the sharper standard that chapter's closing section recommended over trusting a static count taken at rest.

18.5 The Geometric Neural Network and Its Reported Results

The source paper's geometric neural network replaces the ordinary layer $y = Wx + b$ with $y = W \otimes x + b$, where W, x, and b are all multi-vectors, giving a coordinate-free representation whose rotational equivariance is structural rather than learned. Because homomorphic encryption supports polynomial operations rather than arbitrary nonlinearities such as ReLU, the activation function is normalization, $\text{Act}(m) = m/\|m\|$, approximated homomorphically by a low-degree polynomial. A three-layer network, one input multi-vector through 16 then 8 hidden multi-vectors to a small number of output classes, was evaluated on encrypted three-dimensional point-cloud classification across three shape categories, spheres, cubes, and pyramids, each point cloud reduced to a single multi-vector whose scalar component encodes average radial distance, vector components encode centroid position, bivector components encode principal orientation planes, and trivector component encodes a volume measure.

The reported results are a 99% encrypted classification accuracy against 100% on the same task in plaintext, three misclassifications out of 300 encrypted samples, all near class boundaries, with inference completing in under 60 seconds per sample despite the full ciphertext expansion. The reported baseline comparison, a classical multilayer perceptron on three mean-coordinate features reaching 35% accuracy and failing under rotation, against

99% for the Clifford encoding, is not a controlled ablation of flattening alone. The two systems receive different input features: the baseline sees only a centroid, while the multi-vector encoding is built from richer handcrafted statistics, including average radial distance, principal orientation planes, and a volume measure. The comparison is suggestive of what an operationally impoverished representation costs, but it does not isolate the effect of discarding the geometric algebra from the effect of discarding the additional summary statistics, and should not be read as a controlled measurement of either alone. Separately, the reported 99%-encrypted-versus-100%-plaintext result uses handcrafted rather than gradient-trained weights on a three-class toy problem, a limitation the source paper states outright, with full gradient-based training left as future work. What the case actually supports is the structural claim, that reachable distinctions depend on preserved operations and not merely on preserved values, not the specific accuracy figures generalizing beyond this proof of concept.

18.6 Limits and Scope

The source paper is explicit about where the construction stops paying off. Geometric algebras of dimension n have 2^n components, so $\mathrm{Cl}(4,0)$ with 16 components remains feasible, $\mathrm{Cl}(5,0)$ with 32 components is described as borderline, and $\mathrm{Cl}(6,0)$ with 64 components is described as problematic, since ciphertext count grows exponentially with dimension. Problems without underlying geometric structure see no benefit from the construction, and the dense multi-vector representation is reported as inefficient for sparse-matrix problems. Performance remains far from real-time: a single homomorphic multiplication takes approximately 217 milliseconds on the reported CPU implementation against roughly 50 nanoseconds for the equivalent unencrypted geometric product, a slowdown on the order of seven million times, a gap the source paper attributes to the expense of the noise-management operations, relinearization and rescaling, that FHE schemes require after each multiplication.

18.7 Conclusion

Read against the apparatus of Chapter 15, Clifford FHE is not primarily a compression result at all, despite its headline expansion figure inviting that reading. It is a demonstration that a representation's operational repertoire, its native computability over encoded values, is a genuinely independent axis from both how many bits the encoding occupies and how many distinctions it statically preserves, and that this axis can be worth an enormous cost in the other two when what a downstream consumer actually needs is not the values themselves but the structure relating them. This closes Part IV's argument on the same note it opened on: a representation's apparent quality, whether measured by brevity, by fluency, or by cryptographic compactness, is never a single number, and the cases this Part has worked through, Forth's word libraries, borrowed intuition's provenance gap, SRHS's distributed secret fields, and now Clifford FHE's inverted compression tradeoff, are four instances of the same warning, that the visible artifact rarely reports which of its several independent coordinates was actually optimized, or at what cost to the others.

Part V

Political Economy and Social Extraction: Ledgers Without Value and Unfinishable Systems

Chapter 19

Ledgers Without Value

19.1 The Primitive and Its Application

Part IV closed with an argument that a representation's compactness and its operational repertoire are independent coordinates, and that optimizing the visible one can come at the cost of the one that actually matters. This chapter, opening Part V, applies a structurally identical distinction to economic systems: a ledger's internal consistency and its productive value are independent properties, and a system can maximize the first while contributing nothing to the second.

Blockchain technology introduced a genuine innovation: an append-only, distributed ledger capable of maintaining a shared history without centralized authority. Each entry is cryptographically linked to its predecessors, making the record tamper-resistant and verifiable by any participant. This is a meaningful primitive, solving a real coordination problem among parties who share no prior relationship and no shared institutional backing. The question this chapter addresses is not whether the primitive is useful, but what follows when it is applied almost exclusively to a specific purpose: the creation and exchange of speculative digital assets. The dominant use of blockchain technology has not been to coordinate production, reduce friction in supply chains, or establish identity for the unbanked. It has been to sustain markets for tokens whose value is determined not by what they enable but by what the next participant is willing to pay.

The central claim is that a ledger can faithfully record who owns what without imposing any requirement that what is owned corresponds to something of productive value. The immutability of the record does not validate the coherence of the underlying activity. This is not a defect in the ledger itself but a consequence of the gap between two distinct properties: consistency and productivity. Blockchain systems enforce the first with remarkable rigor. They provide no mechanism for the second.

19.2 Value as Reduction of Work

In the most general terms, a process is economically valuable insofar as it reduces the amount of work required to sustain or extend some capacity. A hand tool reduces the physical effort required for a task; a well-designed algorithm reduces the computational effort required for a calculation; an effective logistics network reduces the organizational effort required to move materials from production to use. What this thermodynamic conception of value excludes is equally important: a process that moves resources from one party to another without reducing the effort required to sustain any capacity does not create value in this sense. It redistributes existing claims.

Definition 19.1 (Work functional)**.** Let Ω denote a domain of states and let $\varphi : \Omega \times [0, T] \to \mathbb{R}$ be a field evolving over time. The work functional $\mathcal{W}[\varphi]$ is a non-negative functional measuring the total effort required to reconstruct or sustain the field configuration over $[0, T]$. A process is productive if there exists $t^* \in (0, T]$ such that $\mathcal{W}[\varphi(\cdot, t^*)] < \mathcal{W}[\varphi(\cdot, 0)]$.

Definition 19.2 (Log-admissibility)**.** A system is log-admissible with respect to an event log $\mathcal{L}$ if its current state is consistent with every recorded entry in $\mathcal{L}$: for each transaction $e \in \mathcal{L}$, the system state does not contradict the terms of e.

Log-admissibility is what blockchain technology enforces. The chain guarantees that no entry is added inconsistently, that no prior entry is altered, and that the sequence of ownership transitions is coherent at the level of the record. This is a strong and genuinely useful property. It is also, by itself, entirely orthogonal to productivity.

Definition 19.3 (Positive productivity)**.** A system is positively productive if, in addition to being log-admissible, its aggregate dynamics satisfy $\mathcal{W}[\varphi(t)] \leq \mathcal{W}[\varphi(0)]$ for all $t > 0$, with strict decrease on some non-degenerate interval.

Proposition 19.4 (Redistribution does not reduce work)**.** *Let a system consist solely of transfers of ownership over a fixed set of assets, with no transformation of the underlying capacity those assets represent. Then for any admissible evolution of the system, $\mathcal{W}[\varphi(t)] \geq \mathcal{W}[\varphi(0)]$ for all t, with strict inequality if the transfer process incurs nonzero execution cost.*

Ownership transfers do not alter the physical or informational structure of the underlying system; they only reassign claims. Any execution of the transfer requires coordination, computation, or energy expenditure, which contributes positively to the work functional, so no decrease is possible. The energy costs of proof-of-work mining provide a concrete instantiation: validation alone consumes resources that are not returned in any productive form. A system can be perfectly log-admissible while being entirely non-productive. The ledger records every transaction correctly. The transactions themselves generate no reduction in required work.

19.3 Degeneracy, Not Obstruction

It is worth distinguishing this failure from a different kind of failure in formal systems. When a system fails due to obstruction, the pathology is its inability to construct a globally coherent state from locally valid constraints; the system cannot glue its parts together. The failure of speculative cryptocurrency systems is of a different character. The system is not obstructed. It coheres perfectly well at the level of the ledger. The problem is not that no global section exists but that the space of admissible states is too large: many configurations are consistent with the log, and the system provides no mechanism for selecting among them on the basis of productive value. This is not obstruction but *degeneracy*: the system admits too many solutions, with no criterion for distinguishing the productive ones from the extractive ones. This is a direct economic instance of the same local-consistency-without-global-selection pattern Chapter 8 examined in a physical and quantum-foundational register: TARTAN's boundary defect asks whether adjacent tiles glue into one coherent structure, while this chapter's admissible set asks whether an ownership distribution consistent with the log is preferentially selected toward anything productive, and in both cases local consistency is cheap while the property that actually matters is not automatic.

Formally, suppose the set of states consistent with a log $\mathcal{L}$ is $\mathrm{Adm}(\mathcal{L})$. In an obstructed system, $\mathrm{Adm}(\mathcal{L}) = \varnothing$: no state satisfies all constraints simultaneously. In a degenerate system, $\mathrm{Adm}(\mathcal{L})$ is large, but the subset of positively productive states within it is empty or measure-zero.

Proposition 19.5 (Dynamical degeneracy)**.** *A system is degenerate if there exists a probability measure ρ_t over* $\mathrm{Adm}(\mathcal{L})$ *induced by its dynamics such that* $\lim_{t\to\infty} \rho_t(\mathrm{Adm}^+(\mathcal{L})) = 0$*, where* $\mathrm{Adm}^+(\mathcal{L}) \subseteq \mathrm{Adm}(\mathcal{L})$ *denotes the subset of positively productive admissible states.*

Speculative cryptocurrency markets are degenerate in this sense. The system's native objective, price appreciation under order flow, is orthogonal to the reduction of the work functional, and therefore provides no selection pressure toward productive configurations.

19.4 The Lottery Structure and Informational Asymmetry

The structure described above has a well-known economic instantiation: the lottery. A lottery aggregates small contributions from many participants and concentrates them into large payoffs for a few; its defining property is that the expected value for any individual participant is negative, with the difference captured by the operator. Participation is sustained not by rational expectation of gain but by the perception of possibility, systematically miscalibrated, particularly among those with limited exposure to probability and statistics.

Speculative cryptocurrency markets replicate this structure, though with greater complexity and less transparency. Early participants, project insiders, and coordinated actors occupy positions analogous to the lottery operator: they hold assets acquired at lower cost, and they realize gains as later participants supply the inflow necessary to sustain or increase prices. The mechanism known as the "rug pull," in which project founders liquidate their holdings after attracting sufficient investment, is simply the most legible version of this structure. The log records every transaction faithfully. The coherence of the ledger is undisturbed. What the ledger records is the organized extraction of value from later participants by earlier ones. This is a specific instance of the broader phenomenon of adverse selection under information asymmetry: insiders with superior information about the underlying state of a project extract value from participants whose access to that information is systematically limited. Speculative markets are characterized by self-reinforcing narrative cycles through which asset prices depart from any plausible connection to underlying productive value, with highly skewed and fat-tailed payoff distributions in which the majority of participants experience losses while a few realize dramatic gains.

19.5 Externalization and the Material Ledger

The degeneracy identified above operates not only within the market but at its boundaries. A complete account of a system's value must include not only what it produces for participants but what it costs those outside it. Cryptocurrency infrastructure is not immaterial. The hardware required for mining and transaction validation is produced through global supply chains; electronic waste from obsolete mining hardware is processed in settings where environmental protections are weaker; energy consumption under proof-of-work protocols is often located where electricity is cheap, which in practice means where the ecological costs of generation are borne by populations with limited political recourse. The apparent immateriality of digital assets, the impression that value moves without physical consequence, is an artifact of the ledger's abstraction, not a property of the system itself.

Formally, let Ω_{global} denote the domain including all materially affected agents. A system that appears productive on a restricted domain Ω_{local} may fail to be productive when evaluated on Ω_{global}:

$$\mathcal{W}_{\Omega_{\text{global}}}[\varphi(t)] > \mathcal{W}_{\Omega_{\text{global}}}[\varphi(0)] \quad \text{even if} \quad \mathcal{W}_{\Omega_{\text{local}}}[\varphi(t)] < \mathcal{W}_{\Omega_{\text{local}}}[\varphi(0)].$$

The apparent gains are not reductions in required work but transfers of required work onto parties who did not consent to the exchange and who are not compensated for it, a pattern nameable as uneconomic growth: expansion in measured output accompanied by greater expansion in unmeasured costs, such that the net effect on welfare is negative.

19.6 The Action Functional and Selection Pressure

A productive system organizes its dynamics such that over time it is drawn toward configurations that are stable, generative, and low in action: flat regions of the action landscape where small perturbations do not produce large changes in outcome, and where the system can sustain its function without continuous external input. Speculative cryptocurrency markets exhibit the opposite selection pressure. Price movement attracts attention; attention attracts capital; capital produces further movement. The system is drawn not toward stable, low-action configurations but toward sharp, high-curvature structures, temporarily profitable but intrinsically fragile. The rug pull is the terminal expression of this fragility: once the high-curvature configuration has been exploited by those best positioned to do so, it collapses, and the log records the collapse faithfully while the system moves on to the next configuration.

Remark 19.6. The distinction between log-admissibility and positive productivity maps onto a broader distinction between consistency and realizability. A system can maintain a consistent record of events without ensuring that the events it records correspond to configurations that are stable, generative, or productive. Consistency is a property of the log. Realizability is a property of the world the log purports to describe.

19.7 Non-Productive Labor, Advertising, and Monetized Games

The structure identified in speculative cryptocurrency markets is not unique to that domain. Non-productive labor, the substantial fraction of contemporary employment involving tasks whose disappearance would make no material difference to the functioning of the economy or society, is sustained not because it transforms the system but because it satisfies organizational, reputational, or financial constraints internal to the system itself. It is log-admissible, the work is recorded, evaluated, and compensated, but not positively productive.

Definition 19.7 (Non-productive labor)**.** An activity a is non-productive over a domain Ω if, for all admissible evolutions of the system, $\mathcal{W}[\varphi(t) \mid a] \geq \mathcal{W}[\varphi(t) \mid \varnothing]$, with strict inequality when the activity incurs execution cost.

Advertising occupies a special position because it does not act primarily on the physical or informational structure of a system, but on the constraint landscape through which decisions are made. In a productive system, demand emerges from constraints imposed by actual needs. Advertising, in its dominant form, inverts this relation: rather than responding to existing constraints, it attempts to introduce new ones or distort existing ones, creating perceived deficits where none exist. Formally, letting $\mathcal{C}$ denote the set of constraints defining a decision landscape, advertising operates by modifying $\mathcal{C}$ itself, $\mathcal{C} \mapsto \mathcal{C}' = \mathcal{C} \cup \Delta\mathcal{C}$, where $\Delta\mathcal{C}$ consists of induced or exaggerated constraints that do not correspond to reductions in required work. The evaluation criterion shifts from "does this reduce work?" to "can this be made to appear necessary?"

Games occupy a distinct position among human activities, since their primary function is not the direct production of material goods but the cultivation of cognitive and behavioral capacities. Play is a form of developmental infrastructure: a controlled environment

in which agents encounter difficulty, experiment with strategies, and develop responses to structured challenges without incurring the full cost of failure in the real world.

Definition 19.8 (Developmental productivity)**.** A system is developmentally productive if its interaction with an agent reduces the expected work required by that agent to solve classes of problems in the future, over a domain of tasks not limited to the system itself.

The introduction of in-game purchases that gate progress, manipulate reward schedules, or create artificial scarcity alters the role of the game from a training environment to an extraction mechanism, inverting the function of difficulty. Where a productive game makes the world simpler by providing a structured version of its complexity, the monetized game makes the player's experience harder than it needs to be, and then sells relief from the added difficulty:

$$\mathcal{C}' = \mathcal{C}_{\text{game}} \cup \Delta\mathcal{C}_{\text{monetization}}.$$

Loot boxes and other randomized in-game reward mechanisms requiring payment apply variable-ratio reinforcement, the most powerful behavioral conditioning paradigm available, producing the highest response rates and greatest resistance to extinction; the deliberate application of this mechanism to children, in contexts whose explicit purpose is developmental, represents a structural inversion of what games are for. Games are often consumed by children, whose capacity to model probabilistic mechanisms, identify manipulation, or evaluate long-term costs against short-term frustration is still developing, so monetized games exploit the gap between the capacity being trained and the capacity required to identify the exploitation. Chapter 21 develops the temporal, rather than developmental, register of this same extraction structure at length.

19.8 Conditions for Genuine Productivity

A system satisfies positive productivity if and only if three conditions hold simultaneously. The events recorded in the log must correspond to transformations that reduce the work functional over Ω_{global}, not merely the domain of direct participants, requiring that the benefits of the system's operation exceed its full costs, including material, energetic, and ecological costs that are typically externalized. The dynamics of the system must select for stable, low-action configurations rather than for volatility and rapid reallocation, an attractor in $\text{Adm}^{+}(\mathcal{L})$, not merely a consistent orbit through $\text{Adm}(\mathcal{L})$. And the costs of maintaining the infrastructure must be internalized rather than displaced.

Certain applications of distributed ledger technology come closer to satisfying these conditions than speculative markets: supply chain verification systems, credential management, and cross-border payment settlement correspond to genuine coordination problems whose resolution reduces friction for participants. But even these applications must be evaluated against the full domain of costs; many costs associated with proof-of-work are not necessary for the coordination benefits the technology provides, but are artifacts of design choices that prioritize decentralization over efficiency, so a supply chain verification system running on proof-of-work infrastructure may displace more work than it reduces, even if its direct output is genuinely useful. The criterion is not whether the system does something useful but whether $\mathcal{W}_{\Omega_{\text{global}}}$ decreases.

19.9 Conclusion

The dominant application of blockchain technology in cryptocurrency systems presents a case study in the gap between consistency and productivity. The ledger is an impressive

instrument: it records events immutably, distributes trust without centralized authority, and maintains coherence across a network of participants who share no prior relationship. These are genuine achievements. They are also insufficient for value creation. A system that records ownership transitions faithfully but whose dynamics select for redistribution over production, volatility over stability, and externalized cost over internalized accounting is a system organized around the extraction of value rather than its generation. The immutability of the record does not transform this structure. It preserves it.

The analysis reaches beyond cryptocurrency. The formal structure identified here, log-admissibility without positive productivity, internal coherence without reduction of work, appears wherever systems are organized around the appearance of value rather than its realization: non-productive labor sustains itself by satisfying internal metrics rather than external needs, advertising operates by injecting artificial constraints into decision landscapes rather than by satisfying genuine ones, and monetized games add engineered complexity to environments whose purpose is to reduce complexity, selling relief from the difficulty they introduced. In each case, the log is coherent, the activity is real, and the selection dynamics are misaligned with any productive criterion. This is not a new economic pathology; it is a very old one, reproduced across new technical vocabularies, and it is the same pathology Chapter 20 examines from the contribution-side rather than the productivity-side: a documentary record that faithfully tracks what happened is not thereby a warrant for converting that record into price, reputation, or authority. A genuinely constructive system would reduce work, stabilize configurations, and internalize costs. Until economic systems are evaluated on those terms, rather than on market capitalization, engagement metrics, or the sophistication of their technical apparatus, the dominant forms of the technologies examined here will remain what they structurally are: not engines of progress, but mechanisms for recording its absence.

Chapter 20

Rebel Without a Cost

20.1 The Retrospective Problem

An earlier internal feasibility study, developed under the working name Ayian, proposed converting recorded contribution into a work-backed, persona-bound cryptocurrency, treating recording, recognizing, rewarding, and authorizing as stages of one continuous mechanism. It compared Layer-1 and Layer-2 chains, separated fungible rewards from non-fungible personas, proposed token-bound accounts, and attempted to reduce whale power through a hybrid voting formula. Its governing aspiration was valuable: useful work should count, durable contribution should not disappear, and those who sustain a system should have a meaningful voice in its continuation. None of what follows disputes that aspiration. It disputes the mechanism chosen to realize it.

The difficulty lies not primarily in the selected chain, token standard, or voting exponent. It lies in the compression of several different judgments into a single scalar object. The proposed system turns heterogeneous contributions into a contribution score, converts that score into token issuance, stores the tokens under a persistent persona, and then feeds both stake and score back into governance. Documentary history becomes economic value; economic value becomes political power; identity becomes the container binding them together. This is a specific instance of the same log-admissibility-without-positive-productivity confusion Chapter 19 formalized: a contribution ledger can be perfectly consistent, every record accurate, every score computed as specified, while the conversion from record to price to authority remains structurally unwarranted at every step.

This is an attractive pipeline because each step appears, locally, to preserve the legitimacy of the step before it. Recording seems to warrant recognizing; recognizing seems to warrant rewarding; rewarding seems to warrant a say in what happens next. The appeal is not naive. It is the same appeal that makes seniority, credentialism, and inherited reputation feel natural in institutions that never touched a blockchain. What this chapter adds is an account of exactly where, and exactly why, each of those local inferences fails, stated precisely enough to be checked rather than merely felt.

20.2 Six Acts That Are Not One Act

The central formal claim is a chain of non-implications, running from the fact that an event occurred to the much stronger claim that it should govern what happens next:

$$\text{Record}(x) \not\Rightarrow \text{Verify}(x) \not\Rightarrow \text{Recognize}(x) \not\Rightarrow \text{Reward}(x) \not\Rightarrow \text{Authorize}(x) \not\Rightarrow \text{Continue}(x).$$

Verification is placed here deliberately, and separately from recognition, because the two do different jobs. Verification asks whether an event actually occurred as described, a factual,

evidentiary question. Recognition asks whether the event, having occurred, is regarded as valuable, a normative, social question. Conflating them lets "recognized" silently mean both "confirmed to have happened" and "endorsed as good," which obscures the case that matters most: an act can be accurately recorded and independently verified without being recognized as valuable at all, because verification establishes only that something happened, not that it was welcome.

A contribution may therefore be recorded without being verified. It may be verified without being recognized as valuable, a harmful or mistaken act can be confirmed to have occurred and still deserve no endorsement. It may be recognized without being rewarded, gratitude is not payment. It may be rewarded without conferring permanent standing, payment for a job does not establish general expertise. It may establish standing in one domain without granting authority in another. And it may have been an admissible basis for a decision yesterday without remaining one today. Each arrow in the chain corresponds to an institution people already build to keep the two sides separate, precisely because collapsing them causes recognizable harm: archives distinguish preservation from endorsement, fact-checking distinguishes verification from approval, mutual aid networks distinguish gratitude from payment, professional licensing distinguishes payment from standing, federalism distinguishes domain expertise from general authority, and constitutional term limits distinguish past merit from future permission. A system that mints one convertible token across all six acts does not innovate past these institutions. It discards the distinctions they were built to protect. This six-object chain is the contribution-pipeline instance of the same $H \not\Rightarrow D \not\Rightarrow R \not\Rightarrow C$ structure Chapter 10 developed for agentic systems and Chapter 11 restated at the constitutional layer: verification constrains continuation, but does not authorize commitment, whether the acting party is an AI agent or a community of contributors.

The central revision is therefore architectural: Ayian should not begin with the question, "How do we mint value from work?" It should begin with the more conservative question, "How do we prevent work, provenance, refusal, and obligation from being erased?"

20.3 Why No Universal Scalar Is Canonical

Work is heterogeneous: repairing a bridge, moderating a dispute, discovering a theorem, and caring for a sick person do not occupy a common natural scale, and any scalar conversion is imposed rather than discovered. This can be stated more precisely than that loose intuition, strong enough to carry the chapter's actual claim: a contribution record with genuinely incomparable elements does not, by itself, pick out one correct total ranking.

Definition 20.1 (Contribution preorder and quotient order)**.** Let C be a set of contribution records. For a declared context κ, specifying a domain, an affected population, an evidentiary standard, and a set of relevant dimensions, write $c_1 \preceq_\kappa c_2$ when every comparison admitted by κ judges c_2 at least as well supported as c_1, with no relevant dimension on which c_1 is preferred. Contributions equivalent in both directions are identified, $c_1 \sim_\kappa c_2 \iff c_1 \preceq_\kappa c_2 \wedge c_2 \preceq_\kappa c_1$, and the quotient $C/\!\sim_\kappa$ is then partially ordered. Two records with no dominance relation in either direction remain incomparable in this quotient order.

A scalar valuation $v : C \to \mathbb{R}$ is isotone with respect to $\preceq_\kappa$ when $c_1 \preceq_\kappa c_2$ implies $v(c_1) \leq v(c_2)$. Isotonicity alone is weak: an isotone v may assign equal values to incomparable contributions, which is a defensible and honest choice, since it registers that the record does not distinguish them. The stronger and more common design choice, made implicitly by any system that produces a single ranked leaderboard, requires v to be strict: to assign

distinct values to any two contributions that are not identified under $\sim_\kappa$. A strict isotone valuation induces a total order on its image, precisely a linear extension of the quotient order.

Proposition 20.2 (Non-canonicity of total contribution rankings)**.** *Let $(C/\sim_\kappa, \preceq_\kappa)$ contain an incomparable pair c_1, c_2. Then $\preceq_\kappa$ does not determine a unique total extension of that pair: there exists a linear extension in which c_1 precedes c_2, and another, equally consistent with $\preceq_\kappa$, in which c_2 precedes c_1. Consequently, any scalar system required to rank every pair of contributions must either preserve some incomparabilities as ties, or introduce a comparison that $\preceq_\kappa$ alone does not warrant.*

Let c_1 and c_2 be incomparable under $\preceq_\kappa$. Adjoining $c_1 \prec c_2$ and taking transitive closure introduces no cycle, since a cycle would require $c_2 \preceq_\kappa c_1$ already, contradicting incomparability; the order-extension theorem then guarantees a linear extension realizing $c_1 \prec c_2$. Reversing the construction produces, by the identical argument, a linear extension realizing $c_2 \prec c_1$. Both preserve every comparison already present in $\preceq_\kappa$, so neither is disqualified by the original record, and the record itself supplies no fact that selects between them. This establishes non-canonicity, not impossibility: a total ranking can certainly be constructed, and in a finite contribution set there are only finitely many linear extensions to choose among. The point is narrower and sufficient: no fact contained in the contribution record itself selects which of those finitely many extensions is the right one. A scalar system that ties incomparable contributions is being honest about this; one that breaks the tie with a specific numeric ranking has quietly imported an institutional judgment and presented it as a discovery about the contributions themselves.

The severity of this non-canonicity is governed by the order's width, the maximum size of an antichain, a set of contributions no two of which are comparable under $\preceq_\kappa$. If a contribution poset contains an antichain of size w, its members are compatible with all $w!$ relative orderings. This motivates a design choice: rather than storing a single number that discards the antichain structure the poset actually has, preserve a structured record $r(c) = \langle a, t, d, e, p, o, q, s\rangle$, where a is the actor, t the time interval, d the domain, e the evidence, p the provenance, o the affected objects or communities, q the contestable claims made about the contribution, and s its present status. Two contributions incomparable under $\preceq_\kappa$ because they differ in domain and affected population remain visibly, honestly incomparable in $r(c)$, rather than being forced through a projection that manufactures a rank.

The order-theoretic argument establishes that ranking adds unwarranted comparisons. A separate argument establishes that a transferable reward built on such a ranking cannot remain a faithful indicator of the responsibility it was meant to track.

Proposition 20.3 (Transfer-fidelity conflict)**.** *Suppose a system issues a transferable unit for contribution c, that possession of that unit increases governance authority, and that governance authority is required to track the responsibility or competence evidenced by c. These three requirements cannot be jointly guaranteed.*

Let a receive a unit τ for performing c, and let b subsequently acquire τ without acquiring a's responsibility for, or competence demonstrated by, c. Transferability permits this change of possession. Governance convertibility means possession of τ increases b's authority regardless of how τ was acquired. Fidelity requires that authority track the relation to c that justified τ's issuance, a relation b does not hold. Hence b's increased authority is not warranted by fidelity, and transferability, convertibility, and fidelity cannot all hold simultaneously. Together, these two propositions support the chapter's title theme directly: even setting aside whether a contribution score can be canonical, making that score's reward transferable severs the one relation, between the person acting and the authority the action is meant to justify, that a governance system built on "proof of contribution" was supposed to preserve.

20.4 Prior Art: Ledgers That Already Refuse Conversion

Systems that separate documentary record from price and from authority already exist, in some cases predating any blockchain by decades. Wikipedia's page-history mechanism normally retains and exposes every edit alongside a separate and much smaller mechanism, administrator status, that confers operational authority; edit count alone grants no vote and no formal reputation score, and even a very large edit history confers no automatic operational authority absent a distinct nomination process that can be, and sometimes is, refused. Git's commit history performs a structurally similar separation: content-addressing plus optional signing plus independent mirroring makes silent retroactive alteration conspicuous rather than impossible, but write access, merge rights, and governance votes are typically granted through a separate maintainer process rather than by raw commit count. Local Exchange Trading Systems and time banking keep a running account of hours or services exchanged without treating that account as a source of political authority over the community that issued it, arriving independently at a similar architectural instinct: record contribution, but constrain what the record is allowed to become.

Academic citation practice presents the clearest cautionary case in the other direction, showing what happens when the separation this chapter recommends is allowed to erode. A citation is, in origin, a record; citation-derived metrics such as the journal impact factor and the h-index are a further step removed, aggregations of that record into a score, and the metrics-reform literature has documented how such scores, once adopted for hiring and funding decisions, distort the incentives of the research they were meant merely to describe. A citation record and a citation-derived ranking are not the same object, and treating the second as an unproblematic summary of the first reproduces, in an academic register, the same collapse this chapter argues against in a token-governed one.

20.5 From a Value Ledger to a Documentary Ledger

The proposed ledger is append-only but not declaratively omniscient. It records claims and their provenance rather than asserting one final account of worth. A contribution event may be represented as

$$e_i = \langle \text{id}, \text{actor}, \text{action}, \text{object}, \text{time}, \text{evidence}, \text{attestations}, \text{contestations}, \text{scope} \rangle.$$

The history after n events is $H_n = H_0 \parallel e_1 \parallel \cdots \parallel e_n$, where append-only retention means that later judgments do not silently rewrite earlier evidence. This needs qualification, since an unconditional prohibition on removing any byte conflicts with legitimate privacy, safety, and legal obligations: a later judgment may supersede, contest, seal, or restrict access to an earlier record, but it does not silently replace that record, and where privacy, safety, consent, or law requires underlying material to be sealed or removed, the ledger retains, where permissible, a minimal audit event recording that a governed removal occurred, by whose authority, under which rule, and with what scope, without necessarily retaining the removed content itself.

Operative state is separate from the retained history, $\sigma_n = J(H_n, \kappa_n)$, where J is a revisable judgment procedure and κ_n is the current context, exactly the σ-versus-H separation Chapter 10 formalized as the central discipline of the CLIO architecture. The same history can therefore support different admissible continuations in different domains without contradiction: a medical contribution need not confer software-governance authority; an early architectural contribution need not grant perpetual control; a rejected proposal can remain historically important without becoming operative.

A transferable unit of value can itself be understood as a documentary record from which nearly every field has been removed. A dollar does not ordinarily retain who performed the underlying work, who was affected by it, what obligations were discharged in exchange for it, what harms accompanied its production, or whether its present holder bears any relation at all to the event that caused the unit to exist. Money is not the opposite of a ledger. It is an amputated one. Fungibility is achieved by deleting provenance until units produced by radically different histories become operationally interchangeable. A ledger without value reverses that compression: it preserves reasons while declining to manufacture a universal, frictionless permission from them, making it less liquid but more articulate. The earlier Ayian architecture, in minting a new transferable token from contribution records, was choosing amnesia while believing it was choosing memory.

20.6 Persona Without Ownership

The original Ayian architecture placed an ERC-20 balance inside an account controlled through a Persona NFT. This elegantly binds assets to an identity object, but it also creates a dangerous ambiguity: if the persona token is transferable, then identity and its accumulated history can be sold; if it is non-transferable, then control recovery, coercion, succession, and erroneous binding become difficult governance problems the original study did not resolve.

A persona in a ledger without value is not property. It is a continuity relation among cryptographic control, documentary history, social recognition, and the continuing consent of the subject. No single token constitutes the person. Keys may rotate; attestations may expire; a person may withdraw from an institution while retaining the record of what occurred there. The persona layer should satisfy four constraints: control must be recoverable without selling identity; claims must be scoped and revocable by their issuers without erasing their historical issuance; the subject must be able to refuse new associations with their self-presented identity; and the ledger must distinguish identity continuity from institutional membership, since leaving a system is not the destruction of a person, and retaining history is not permission to keep governing them.

This last constraint requires a distinction the earlier draft elided: control recovery is not the same operation as succession. Recovery asserts continuity of the same subject, a new key continuing the same persona. Succession transfers a role, custody, or unresolved obligation from one subject to a different one entirely, and must not be mistaken for the predecessor's identity passing along with it: a successor inherits stewardship of what a persona left behind, not the standing to have been that persona. A recovery mechanism built only around "prove you are still you" is vulnerable to an adversary who has compromised the subject's other credentials, or to a subject coerced into authorizing a transfer they do not actually consent to; a recovery process with a mandatory delay window, notification to previously designated attestors, and a contestation period before any high-consequence action takes effect gives a coerced or compromised subject a real chance to intervene before the action becomes irreversible.

A subject cannot reasonably veto every adverse claim made about them, that would let anyone erase accountability simply by refusing to acknowledge it, but neither should an unverified allegation appear in the record with the same standing as an accepted fact. The architecture should therefore distinguish subject-approved associations, which the subject has affirmatively accepted; third-party attributed claims, which remain visible and attributable to their source but are marked as contested until resolved; and institutionally admitted findings, which have passed through some accountable adjudication process. Consent governs the first category. It does not, and should not, govern the second or third, which is what preserves the ledger's usefulness for accountability rather than converting it into a subject-

curated profile.

20.7 Governance After the Voting-Power Formula

The corrected Ayian formula was $P_i = \sqrt{S_i}\, C_i$, where S_i is staked Ayian and C_i is lifetime contribution score. The square root improves the formula relative to linear stake by diminishing the marginal power of large holders. It does not, however, resolve the underlying constitutional problem. Multiplication means that economic stake and recorded contribution validate one another: a high value in either dimension magnifies the other, while a zero in either annihilates the person's voice, and it turns a lifetime record into a permanent political multiplier, allowing old contribution to harden into office. The term "quadratic voting" is also imprecise here; taking $\sqrt{S_i}$ applies a concave weighting to stake, while quadratic voting ordinarily concerns the quadratic cost of purchasing multiple votes. More importantly, if C_i is obtained by reducing heterogeneous contribution records to a universal score, it is one particular resolution of the non-canonical choice the previous section's proposition describes, a strict isotone valuation that has broken every tie the underlying order left open, so multiplying by $\sqrt{S_i}$ scales economic stake by comparisons that were introduced during the scalarization of history, not supplied by the history itself.

A second, independent problem concerns how C_i behaves over time. If lifetime contribution is cumulative and nondecreasing, $C_i(t+1) \geq C_i(t)$, then authority obtained from early contributions persists in full even after their contextual relevance has declined, producing institutional lock-in in which incumbency is self-reinforcing regardless of whether the incumbent's original domain of expertise still bears on present decisions. And the transfer-fidelity result compounds the concern further: even setting aside whether C_i is canonical or decaying appropriately, the moment any resulting reward becomes a transferable token, authority stops tracking the responsibility that justified it at all, and instead tracks whoever currently holds the token.

A ledger without value replaces universal voting power with scoped standing. For a proposal p, the relevant set is not everyone possessing a balance but those with a defensible relation to the decision:

$$\mathcal{S}(p) = \{i \mid \text{affected}(i,p) \vee \text{responsible}(i,p) \vee \text{competent}(i,p)\}.$$

These predicates should remain distinct. Being affected grounds a right not to be ignored. Being responsible grounds accountability for implementation. Competence grounds evidential weight, not sovereignty. None should silently consume the others. Decision procedures can then vary by consequence: routine maintenance may use delegated responsibility; factual questions may use contestable expert judgment; allocation decisions may require consent from affected groups; constitutional changes may demand broad ratification and strong exit protections. The ledger supplies evidence about standing and history, but it does not compute a universal rank of persons.

20.8 Contribution Without Permanent Reputation

A lifetime contribution score can become a second currency: scarce, accumulable, unequally distributed, and convertible into access, and unlike money it may be impossible to give away or escape, producing hereditary institutions without hereditary titles, because early contributors indefinitely dominate the definition of later contribution. The alternative is contextual

recognition with decay of authority but not decay of history. The record that an action occurred remains. Its relevance to a present decision must be argued again:

$$\rho(c, p, t) = f(\text{domain fit}, \text{evidence}, \text{recency}, \text{affectedness}, \text{contestations}),$$

where ρ is not a stored personal score but a temporary assessment of how contribution c bears on proposal p at time t. Recency may serve as a useful proxy for current salience, but it is neither truth nor entitlement; older work remains reachable when similarity, dependency, or renewed relevance makes it operative again. This resembles tab completion more than a leaderboard: the system proposes plausible continuations from the current prefix and context, rather than declaring that the most historically prolific person owns the next symbol.

A worked scenario traces one contribution through the full architecture. A contributor, persona P_{47}, submits a patch addressing a scheduling bug: Record(e_{881}). A CI run and a maintainer's manual review establish Verify$(e_{881}, d_{\text{scheduler}})$, a factual finding that would hold even if maintainers subsequently judged the fix undesirable for unrelated reasons.

Two maintainers attest that the fix is a genuine improvement, constituting

$$\text{Recognize}(P_{47}, e_{881}, d_{\text{scheduler}}),$$

scoped explicitly to that domain.

A governance process then authorizes payment, Reward$(P_{47}, e_{881}, o_{19})$, recorded as an explicit obligation, not newly minted tokens. None of Record, Verify, Recognize, or Reward yet grants P_{47} any say over unrelated decisions. Six months later, a proposal to change the project's licensing terms arises, a domain in which P_{47} has no evidenced standing at all:

$$\text{Reward}(P_{47}, e_{881}, o_{19}) \not\Rightarrow \text{Authorize}(P_{47}, p_{\text{license}}).$$

A separate proposal to change the scheduler's core allocation algorithm computes P_{47}'s standing as $\rho(c_{881}, p_{\text{sched}}, t_{\text{now}})$: domain fit is high, evidence is strong, recency has begun to decay, yielding meaningful evidential weight in the scheduler deliberation without granting unilateral control over the outcome. Contrast this with the original Ayian pipeline applied to the same event: the patch would generate a contribution score, the score would mint tokens, and P_{47}'s voting weight on the licensing proposal would be identical to their voting weight on the scheduler proposal, because the token does not know which domain it came from.

20.9 Useful Work and Thermodynamic Honesty

The phrase "Proof of Contribution" risks repeating the central mistake of proof-of-work systems: inventing measurable activity in order to manufacture trust. Useful computation is not whatever satisfies a protocol's puzzle. It is work that an independently existing human, scientific, ecological, or infrastructural need already required. Heat is the terminal by-product of work; a heater that produces heat directly discards the opportunity to perform useful computation or mechanical work along the way, and a cryptocurrency that invents computation solely to establish scarcity discards the opportunity to direct energy toward real unmet tasks, exactly the extraction structure Chapter 19 formalized as degeneracy: proof-of-work's native objective is orthogonal to any reduction of the work functional. A defensible system would schedule computation where heat is wanted, route tasks to appropriate climates and buildings, and count reduced duplication and recovered waste as part of its evidence.

But even genuinely useful work should not automatically mint governance. Solving a protein-folding problem, hosting winter computation, or maintaining a bridge may justify

contractual payment and public acknowledgment. It does not create a natural claim to rule unrelated communities. The ledger can demonstrate that useful work occurred while refusing the metaphysical leap from energy expended to value owned. This is the six-act chain applied to a specific and increasingly common case: genuinely useful computational work is real Record and Verify, can readily earn Reward through an explicit obligation, and still does not by itself reach Authorize. Chapter 11 developed the constitutional layer this observation depends on, including the treatment of refusal itself as a first-class, non-punitive continuation option; Chapter 22 takes up the sheaf-theoretic generalization of the affordability and necessity questions this chapter's scoped-standing apparatus raises but does not itself resolve.

Chapter 21

Unfinishable Games and Temporal Extraction

21.1 Two Critiques, One Vocabulary

Chapter 19 examined how a monetized game inverts the developmental function of difficulty, converting a training environment into an extraction mechanism through engineered scarcity and variable-ratio reinforcement. This chapter examines a second, analytically distinct extraction mechanism present in the same broad genre, one that requires no purchase at all: structural unfinishability, the removal of a participant's unilateral control over when disengagement from a persistent, real-time, alliance-based game is safe.

The last decade of criticism directed at exploitative game design has converged almost entirely on monetization systems that use randomized rewards to convert player spending into a variable-ratio reinforcement schedule structurally similar to a slot machine. This body of criticism is substantial and its harms are real, though the international regulatory picture is considerably more fragmented than a simple narrative of bans would suggest: Belgium's gaming commission found certain paid loot box implementations to constitute gambling, while the Netherlands' comparable finding was later overturned on administrative appeal, and the United Kingdom's government-commissioned review concluded that an association between loot box spending and gambling-related harm existed but that a causal relationship had not been established. None of this requires the game to run in real time. A turn-based game, or one with no live multiplayer component at all, can implement an identical loot box system with identical monetary extraction and identical gambling-adjacent psychology. The mechanism is orthogonal to the game's temporal structure, and this orthogonality means a second mechanism, tied specifically to real-time synchronous structure, cannot be a mere variant or intensification of the first. It has to be examined on its own terms, with its own formal apparatus, rather than folded into the borrowed vocabulary of addiction, which tends to relocate the explanatory burden onto the player's self-control rather than onto the properties of the system that make stopping costly.

21.2 Formalizing Structural Unfinishability

A precise definition needs to separate four phenomena an informal account of "real-time games" tends to run together: persistent simulation, globally distributed participation, consequential events occurring during a participant's absence, and social responsibility for preventing collective losses. No one of these alone produces the mechanism this chapter is concerned with. A persistent world can permit harmless absence. A globally distributed player

base can coordinate asynchronously. Consequential events can be bounded by protection mechanics. Responsibility can be spread widely enough that no single participant remains continuously accountable. The mechanism arises specifically from their conjunction.

Definition 21.1 (Structural unfinishability)**.** A game is structurally unfinishable for participant i when consequential state changes can occur during i's absence, when their timing is substantially controlled by other players or the server rather than by i, when the resulting losses are difficult to reverse, and when responsibility for preventing those losses is socially attributed to i.

This makes unfinishability relational rather than a fixed property of the software: the same game can be an easily stoppable pastime for an ordinary member and an effectively continuous obligation for an alliance leader, because $R_i(t)$, the responsibility attributed to participant i, and $U_i(t)$, the difficulty of substituting another participant for i, both vary by role. A hazard or event-intensity term $\lambda_i(t)$, the expected rate of consequential events conditional on i's absence, is estimable in principle from server logs. Let $L_i(t)$ denote the expected loss to i conditional on an event occurring while absent, let $R_i(t)$ denote socially attributed responsibility, and let $U_i(t)$ denote the difficulty of substitution. Define i's temporal exposure over an interval T as

$$X_i(T) = \frac{1}{|T|} \int_T \lambda_i(t)\, L_i(t)\, R_i(t)\, U_i(t)\, dt.$$

A safe disengagement interval $I \subseteq T$ exists when $\int_I \lambda_i(t) L_i(t) R_i(t) U_i(t)\, dt \leq \varepsilon$ for some tolerable threshold. This decomposition separates four independent levers, exposure, cost, attributed responsibility, and irreplaceability, and a design intervention can reduce unfinishability by attacking any one of the four without needing to touch the others. Structural unfinishability is not the literal absence of any interval of low exposure; it is the scarcity, or the participant's inability to unilaterally guarantee the existence, of a sufficiently long interval satisfying the inequality. This makes the concept comparative and, in principle, measurable, and it draws a necessary line between persistence and unfinishability: a voxel-building server can operate continuously all night without $L_i(t)$ or $R_i(t)$ rising appreciably for any particular player, because nothing consequential and attributable to that player is at stake. Unfinishability names the specific condition in which the world's continuous operation is coupled to a specific participant's exposure rather than decoupled from it.

21.3 The Circadian Vulnerability

A further asymmetry explains why global synchronicity in particular, rather than mere real-time operation, aggravates $\lambda_i(t)$ for a globally distributed alliance. Human availability is approximately periodic on a roughly twenty-four hour cycle tied to local time zone. Let $a_j(t)$ denote attacker j's availability, and let $A(t) = \sum_j a_j(t)$ denote the aggregate availability of the full adversarial population. When attackers are concentrated in a single time zone, $A(t)$ inherits a strong daily rhythm; as the population's time-zone dispersion increases, the individual peaks and troughs of $a_j(t)$ increasingly fall at different points in the day, and $\mathrm{Var}_t[A(t)]$ falls correspondingly, even while each individual attacker retains their own strongly periodic availability. In the limit of a sufficiently large and well-distributed adversary population, $A(t)$ approaches a roughly constant level across the full cycle, while a defending player retains their own strongly periodic availability, low during local sleeping hours. The defender's least-available hours are, in expectation, no less contested than any other hours. Circadian absence, a biological constant no game design choice can remove, is thereby converted into a strategically exploitable regularity in $\lambda_i(t)$ purely as a function of

the adversarial population's geographic dispersion, without requiring any single attacker to act with deliberate intent, yielding a testable prediction: the severity of the effect should scale with the variance reduction in $A(t)$ as dispersion increases.

21.4 Delegability

Delegability, formalized as low $U_i(t)$, determines whether global synchronicity actually forces continuous individual availability or can instead be absorbed by a group. If responsibility for responding to a threat can be handed seamlessly to another participant, no single individual need remain continuously available even though $\lambda_i(t)$ itself never falls to zero. The configuration this chapter is concerned with therefore requires a specific conjunction of high λ_i, high L_i, high R_i, and high U_i simultaneously; removing any one, through protected rest periods, reliable delegation, bounded and recoverable losses, or a genuinely redundant responsibility structure, moves a game outside the category even if it remains globally synchronous. This predicts which superficially similar games should escape the mechanism, not only which ones exhibit it.

21.5 From Session to Membership

A server clock alone cannot make a person wake at an unscheduled hour; what supplies the compulsion, formally, is $R_i(t)$, a social quantity assigned by other participants who are waiting, depending on the player's presence, or standing to suffer a consequence alongside them. This is clearest by contrast with an older model of multiplayer participation, real-time strategy titles played over a modem, in which multiplayer was an optional session: players deliberately connected, played a bounded match, and disconnected, with no persistent structure surviving the session's end. A modern alliance-based persistent game replaces the match with membership:

$$\text{multiplayer as an optional session} \longrightarrow \text{persistent social membership}.$$

Structural unfinishability, in the fullest sense intended here, is what results when membership of this kind is combined with the exposure structure above: the game has stopped being a session one opts into and become a social position one occupies continuously, with occupancy itself carrying consequences for others.

$R_i(t)$ cannot arise from a server clock alone. A clock can make $\lambda_i(t)$ nonzero at a given hour, but it cannot, by itself, attach responsibility for the consequences of that hour to any specific person. That attribution requires durable relational infrastructure: an alliance that remembers who holds which rank, a chat channel through which expectations are communicated and enforced, a reputation system through which absence becomes visible to others. Once a game supplies persistent identity, standing group membership, user-to-user communication, and reputation, the software is performing social-platform functions in addition to game functions, and this compounds a familiar mechanism, absence made socially legible and costly, with a second, mechanical one ordinary social media lacks: a participant's assets, territory, or alliance standing can be materially changed by other participants specifically because they were absent.

21.6 Extraction Requires a Beneficiary

Burden alone does not establish extraction in any economically meaningful sense; extraction requires showing that the burden's product is captured by someone other than the person

bearing it. Unfinishability is the structural condition analyzed above; extraction occurs only when some actor captures value produced under that condition, and the two are logically separable. A nonprofit, open-source persistent game could in principle exhibit exactly the same λ_i, L_i, R_i, U_i profile, the same structural unfinishability, without any publisher capturing economic surplus from the resulting coordination labor, because no publisher exists to capture it. Unfinishability would remain a real burden, but calling it extraction in that case would be a category error.

The relevant claim for a commercially operated game is therefore a distinct, additional hypothesis. Let V_{social} denote the value of the player-produced organization that coordination labor generates, recruitment, scheduling, dispute resolution, and general cohesion of an active player base, and consider the hypothesis $\partial\,\text{Retention}/\partial\,V_{\text{social}} > 0$: that a publisher's retention rises with the value of this uncompensated player-produced organization. Stated this way, extraction becomes an empirically testable economic hypothesis about a specific game and publisher, rather than a term the argument's own vocabulary presupposes:

$$\text{player coordination labor} \longrightarrow \text{persistent social content}$$

$$\longrightarrow \text{retention and monetizable engagement.}$$

An alliance leader who recruits and retains members, schedules events, moderates disputes, and manufactures the rivalries that make ongoing play compelling is producing exactly the kind of durable social content that keeps a persistent game populated and worth continuing to operate for everyone else in that alliance. This labor is not compensated by the publisher. Whether the hypothesis holds for a given commercial title is an empirical matter this chapter does not settle; what it establishes is the more limited claim that appropriation, a publisher benefiting functionally from an active, well-organized population, is not automatically the same thing as extraction, deliberate capture of a specific economic surplus, and that a commercially operated persistent game exhibiting structural unfinishability should be examined for the latter rather than assumed to practice it merely because the former is present.

21.7 Case Study: *Last Shelter: Survival*

Last Shelter: Survival provides a concrete, bounded illustration of this architecture, presented here as an example, not as evidence that every game in its genre produces identical effects. The developer's own listing describes the game as a real-time strategy title built around persistent base development, alliance formation, and continuous conflict between players worldwide, and the storefront listing discloses loot boxes, in-app purchases, and persistent messaging, with an age rating of thirteen and older. The game's competitive structure includes a server-versus-server "State vs State" competition that a community-maintained guide describes as running on a roughly forty-eight-hour cycle, during which an alliance's collective standing depends on continuous, coordinated participation across the event's full duration. A partial, individually scoped mitigation exists in the form of a temporary defensive shield subject to a reactivation cooldown, but this does not by itself relieve the coordination requirement borne by alliance leadership during a live, alliance-wide event.

An ordinary alliance member may have comparatively low $R_i(t)$ and low $U_i(t)$: their individual absence is substitutable. An alliance leader, diplomat, or event coordinator occupies a different position on both quantities, with responsibility for war declarations and defensive scheduling commonly and visibly attributed to specific ranked roles, and, particularly in smaller alliances, the difficulty of substituting another coordinator during an active event may be high. Unfinishability, on this reading, is not a claim about the software in the

abstract but a claim about a specific, identifiable role within it, exactly the relational property the formal definition was built to capture. Official storefront listings establish $\lambda_i(t)$ and, more weakly, $L_i(t)$ reasonably well; player-produced community guides are reasonable evidence that particular mechanics exist and function as described; neither source establishes $R_i(t)$ or $U_i(t)$ directly, since both are social quantities that would properly require observation of actual alliance communications or survey data to measure. What the case study shows is that *Last Shelter: Survival* instantiates the technical preconditions under which the proposed mechanism can arise, while whether particular players in particular roles actually cross a specified X_i threshold remains an empirical question the case study does not itself answer:

$$\text{architecture observed} \implies \text{mechanism predicted}$$
$$\implies \text{player study tests the prediction.}$$

21.8 Coordination as Structural Incentive, Not Established Population-Level Outcome

This structural account converges with an older strand of game-studies research on the unpaid labor performed by guild and alliance leadership in earlier large-scale multiplayer games, in which leading a large player collective routinely crosses from recreational play into unpaid, intensely time-consuming work, a phenomenon related to, though not identical with, what has been termed "playbour" in analyses of unpaid labor in game modding. The comparison should be drawn carefully rather than overstated: earlier raid-centered games were persistent, but much of their highest-coordination content could be scheduled by the participating group itself, through socially negotiated session boundaries and lockouts, whereas in globally synchronous territory games, consequential adversarial actions can instead be initiated by opponents during the defending group's absence, on a timetable the defenders do not control. It is also worth being explicit about the limits of the available evidence: the playbour literature supports an analogy between guild leadership and unpaid labor, but it does not, by itself, establish population-level empirical claims about how commonly severe or chronic consequences occur among players of the specific genre this chapter examines. This chapter's claim is structural rather than epidemiological.

21.9 Remedies and Their Enforcement Difficulties

The formal apparatus yields a design criterion in addition to a diagnosis. Let player i choose, at some time t_0 of their own selection, an absence interval of duration Δ, denoted $\mathcal{A}_i(t_0, \Delta)$. A system provides a genuine right to absence if, subject only to reasonable frequency and duration limits, i can invoke $\mathcal{A}_i$ unilaterally and thereby obtain $X_i([t_0, t_0+\Delta]) \leq \varepsilon$. The word doing the essential work in this criterion is unilaterally. If invoking the interval requires permission from an alliance leader, negotiation with an adversary, or merely hoping that no one attacks during the window, the system has not supplied a right to absence; other participants are supplying protection contingently, a different and considerably weaker thing. This yields a simple diagnostic question: can the participant press leave and have the consequences stop? For a modem-era match, the answer is essentially yes. For the architecture this chapter has examined, the answer is often no, or no with qualification, particularly for participants occupying the high-R_i, high-U_i coordinating roles the case study identified. This formalization is considerably stronger than generic "take a break" prompts, since a break reminder that fires while the underlying game mechanics simultaneously threaten a participant's alliance with territorial loss for having heeded it does not satisfy the criterion; the

reminder is a suggestion layered on top of an architecture that still penalizes compliance, not an actual instance of $\mathcal{A}_i$.

Several distinct remedy mechanisms could contribute, each carrying its own limitation. Individual protection windows reduce a single participant's personal exposure but do not, by themselves, reduce the collective responsibility a coordinating role still bears during an active event. Alliance-level vulnerability windows address the collective case but raise the question of whose local time such a window should be anchored to when the alliance itself spans every time zone. Asynchronous attack declarations, requiring an aggressor to announce intent before a delay period elapses, could decouple event timing from any single participant's sleep schedule, at some cost to the immediacy that gives adversarial play its stakes. Hard caps on the frequency of consecutive high-stakes events limit cumulative exposure without eliminating any single event's intensity. Reversible or insured offline losses reduce $L_i(t)$ directly. Role-redundancy requirements would directly lower $U_i(t)$ for critical roles. None is a complete solution on its own. The regulatory dimension is correspondingly less settled than the loot box case, since temporal extraction has no equivalent transactional moment for a regulator to examine, and would more plausibly require product-design standards or an analogy to occupational scheduling protections rather than an existing legal category.

21.10 Conclusion

The critical and regulatory apparatus built around loot boxes has done real work, but it was built to answer a specific question, whether a randomized purchase mechanic functions as disguised gambling, and it has no purchase on a separate question: whether a game's temporal and social architecture, quite apart from anything it sells, can remove a participant's unilateral control over when disengagement is safe, and whether the resulting coordination labor becomes an uncompensated input to the game's commercial operation. The distinctive claim this chapter has defended is not that such games consume a great deal of time, since an absorbing game someone chooses to play for many hours is not thereby exploitative, but that structural unfinishability transfers control over the boundaries of play from the participant to the platform and to the other participants inhabiting it, so that leaving becomes something that can be done to a person's standing rather than something the person can simply decide to do.

Exploitability can arise from the topology of permissible disengagement.

Stated this way, the phenomenon generalizes beyond games, to on-call work schedules, notification-driven messaging platforms, always-live financial markets, and social media systems that make absence itself visible and costly to a user's standing. Money and time are extracted through different mechanisms, diagnosed by different evidence, and addressed by different remedies.

There is, finally, a structural symmetry worth noting between the argument developed here and Chapter 4's companion argument about the handling of historical record. That argument holds that a system should not let a later correction retroactively rewrite an earlier commitment, that a past claim, once made, should be preserved even as further evidence constrains how it is interpreted. This chapter holds a mirror-image position about the future rather than the past: a persistent system should not let standing membership retroactively claim unlimited authority over a participant's future availability. One argument denies a system authority over what already happened; the other denies a system authority over when participation must happen next. Both reject the same underlying architectural move,

a system quietly converting continuity, whether of a record or of a membership, into an unbounded claim, on the past in one case, on the future in the other, and both are instances of the qualified non-erasure principle Chapter 10 formalized as the separation of documentary history from operative authority: history that extends does not thereby license unlimited claims on what must happen next.

Chapter 22

Sheaves of Necessity and Affordability

22.1 What We Call Affordable

A society reveals its actual moral ordering not in its stated values but in what it is willing to call affordable without argument, and in what it quietly decides must never become cheaper. Fireworks are affordable. A second school counselor is not. A stadium subsidy is affordable. A universal mentorship program is not. An industrial meat supply chain is affordable, in the sense that its true cost is never presented as a bill anyone has to approve, while a public health intervention with a fraction of that cost must survive a legislative session to be funded at all. And the falling prices that farming, manufacturing, and logistics improvements might otherwise have delivered are set aside instead, so that the real burden of existing debt can be eroded and asset valuations supported, a trade made permanently and by default rather than case by case in response to actual deflationary risk.

This is not an argument for joylessness, nor for monetary chaos, nor a demand that ordinary people give up rest, travel, the company of a pet, or a stable currency while distant poverty persists in some abstract, unreachable form. It is an argument that the discretionary and the necessary, and the stable and the exploitative, have been sorted by institutional habit rather than by honest accounting, and that habit has consistently favored spectacle, appetite, convenience, and the protection of existing claims over development and the diffusion of technological gain, because those are what the people who set budgets and write monetary policy already hold. The remedy is not universal austerity. It is the same accounting, applied consistently and without exception for familiarity: weigh external cost against available alternatives, apply that standard progressively so restraint falls where alternatives exist, and stop allowing the sheer familiarity of an expenditure or a policy default to substitute for its justification. Until that accounting is done, the words affordable and stable will keep doing quiet moral work that the society using them has never actually agreed to.

22.2 Affordability as Admissibility

What follows reinterprets the preceding argument in the vocabulary this monograph has used throughout for questions of constraint and admissibility. It introduces no claim the argument above could not already support on its own terms; it is a change of lens, applied at the close of the book to make explicit a structural pattern that has, in fact, already recurred at every scale examined so far.

Let X be the space of possible allocations of a society's resources, and let $\mathcal{A}_t \subseteq X$ be

the institutionally admissible region at time t: the set of allocations an institution will fund, permit, or subsidize without demanding fresh justification. Admissibility in this sense is a membership condition, not an optimality claim; an allocation can remain inside $\mathcal{A}_t$ indefinitely without ever having been evaluated against any openly stated welfare function, simply by virtue of having entered $\mathcal{A}_t$ historically and never having been reconsidered for removal. Stadium subsidies, industrial meat production, and spectacle infrastructure persist inside $\mathcal{A}_t$ for exactly this reason: normalization placed them there, and nothing in the ordinary operation of a budget cycle asks them to justify their continued membership. Mentorship, tutoring, and preventive care can carry substantially greater developmental value while remaining outside $\mathcal{A}_t$, unable to enter without discharging a proof burden that the already-admitted claims inside $\mathcal{A}_t$ never face even once.

This is precisely the admissible-but-non-committable structure Chapter 10's CLIO architecture formalized at the level of an individual agentic system: $\mathrm{Adm}(r)$ never by itself supplies $W(r)$ or $\mathrm{Authorized}(r)$, and a proposal's mere membership in a permitted region is not a warrant for its continuation. What CLIO treats as a design requirement for one acting system, this chapter treats as a diagnosis of an entire society's budget cycle: an allocation's presence inside $\mathcal{A}_t$ is a historical fact about what was once admitted, not a live judgment that it remains warranted, and the two are routinely conflated precisely because nothing in ordinary institutional operation forces the distinction into view. Genuine political change, on this reading, is not optimization within a fixed $\mathcal{A}_t$, reshuffling a feasible set whose boundary is treated as given; it is revision of the boundary itself, an operation on the region rather than within it, a modification of the operative constraint regime rather than a search confined to whatever that regime currently permits.

22.3 The Category of Claims

A second structure captures something admissibility alone does not: not every excluded claim is excluded for the same reason, and not every admitted claim arrived by the same route. Let **Claims** be a category whose objects are socially recognized claims on resources, and whose morphisms are accepted justifications, the argumentative or institutional moves by which one claim is converted into another, more legible institutional form: civic pride justifies a stadium; consumer choice justifies industrial agriculture; economic stability justifies an inflation target; property rights and the goal of attracting foreign investment justify a speculative housing purchase. This is a genuine category rather than a mere directed graph of associations because these justifications compose: if civic identity justifies tourism expenditure and tourism expenditure in turn justifies stadium infrastructure, their composition supplies an accepted route directly from civic identity to stadium funding, without either intermediate step needing to be re-argued, and every claim carries an identity morphism representing its bare recognition without translation into any other form.

The category is not richly connected. Some objects sit at the center of a dense web of morphisms with short compositional paths to funded outcomes; others are nearly isolated, reachable only through long, unfamiliar, or nonexistent chains of composition. A sports subsidy maps, by well-worn institutional paths, into employment policy, regional development, civic identity, tourism promotion, and public cohesion, any one of which supplies a sufficient justification on its own. A child's need for sustained mentorship typically has no comparably worn morphism into a funded budget line, and a tenant's claim to remain housed in their own neighborhood fares no better: property rights supply the buyer with an immediate, well-recognized morphism into legitimate ownership, while years of residence, employment, and community membership supply the tenant with no comparably accepted arrow into any form of institutional protection at all; the claim exists as an object, but the

category contains no accepted route, direct or composite, carrying it there.

Privilege, in this vocabulary, is morphic abundance rather than intrinsic urgency: not merely many outgoing arrows, but many short compositional paths into the budgetary objects an institution already funds. A privileged claim need not be more pressing or more valuable than its excluded counterpart; it inhabits a richer justificatory neighborhood, one that institutions already have the translation apparatus, and the compositional shortcuts, to recognize, in the language of economics, culture, security, or tradition, whichever happens to be at hand. Deprivation persists in part because urgent claims lack comparable translation paths, not because the claims themselves are weaker. This is the categorical form of an asymmetry this monograph has met before in a different register: Chapter 20's account of contribution showed that a documented act does not compose, by any warranted morphism, into recognition, reward, or authority, $\text{Record}(x) \not\Rightarrow \cdots \not\Rightarrow \text{Continue}(x)$, and that a system minting one convertible path across all six acts manufactures exactly the compositional shortcut this chapter's category of claims shows privileged allocations already enjoy by historical accident rather than by argument. The entitled claim does not have to argue its case because it already has an accepted morphism, or a short composite of them, waiting; the developmental claim has to build one from nothing, in whichever vocabulary a given funding cycle happens to honor that year.

22.4 A Sheaf of Necessities

The deepest structural feature is neither the shape of the admissible region nor the density of justificatory morphisms, but a failure of composition across scale. Let $\{U_i\}$ be a cover of the shared resource space X by institutional or social contexts, cities, households, industries, national governments, and let $\mathcal{N}(U_i)$ denote the collection of allocations regarded as necessary within context U_i, so that a particular such judgment is a local section $s_i \in \mathcal{N}(U_i)$. A city can regard a stadium as economically necessary; a household can regard an annual flight as personally necessary; an industry can regard cheap meat as commercially necessary; a government can regard restrained social expenditure as fiscally necessary. Each of these local sections can be internally coherent, and they can appear compatible on the limited overlaps that ordinary institutional practice actually inspects, a city's accounting rarely checks itself against a household's, an industry's rarely checks itself against a government's long-run fiscal position. But apparent pairwise compatibility on the overlaps actually inspected is weaker than genuine compatibility on the full cover. Once ecological, intergenerational, and developmental effects are included in the restriction maps, some of these sections disagree on overlaps that ordinary accounting had simply omitted from consideration.

The absence of a global section $s \in \mathcal{N}(X)$ compatible with ecological limits and universal developmental provision is therefore not a mysterious failure of the sheaf axiom, which guarantees gluing precisely when compatibility genuinely holds on every overlap; it is evidence that the locally asserted necessities were never actually compatible over the full cover, only over the narrower set of overlaps institutions habitually check. What looks affordable at every overlap actually inspected can still be globally impossible; what appears optional at every overlap actually inspected can still be globally indispensable. Contemporary institutions are reasonably reliable at validating claims within a bounded jurisdiction, a household, a city, an industry; what they lack is a restriction map wide enough to catch the overlaps, ecological, intergenerational, planetary, that would reveal the incompatibility their narrower accounting conceals.

This is not a new construction in this monograph, and naming its recurrence is the point of placing this chapter last. Chapter 8 formalized the identical obstruction in a physical register, TARTAN's tile-gluing defect K_{ij}, and read Buchanan's Kochen–Specker contextuality

result as an independent instance of the same pattern: locally consistent measurement contexts that nonetheless admit no globally coherent assignment, the local-coherence-without-global-section failure stated there in the vocabulary of sheaf theory and quantum foundations. Chapter 19 formalized a close relative of it in an economic register, calling it degeneracy rather than obstruction: a speculative system whose admissible set $\mathrm{Adm}(\mathcal{L})$ is large and internally consistent at the level of the ledger, while providing no selection pressure toward the productive subset $\mathrm{Adm}^{+}(\mathcal{L})$ within it. All three, TARTAN's tiles, the ledger's admissible set, and this chapter's cover of necessities, share the same mathematical shape: local well-formedness is cheap and reliably achieved; the property that actually matters, global coherence, positive productivity, universal compatible necessity, is not automatic, and no amount of local diligence substitutes for checking it directly. Chapter 1's dependency chain, $C_t \Rightarrow A_t \Rightarrow L_t \Rightarrow X_t$ but not the reverse, and Chapter 15's non-identification of L_T and D_T are the same warning stated at the level of a single trace or a single representation rather than a cover of contexts: an earlier, cheaper property holding gives no guarantee about a later, more demanding one, and the gap between them is exactly where a system's actual behavior, or a society's actual moral ordering, turns out to live.

Read together, the three sections of this chapter trace one failure at three depths. Admissibility explains why familiar pleasures are never asked to justify themselves twice, while developmental need is asked every time. The category of claims explains why some exclusions look like weakness of argument when they are really poverty of translation. The sheaf of necessities explains why a society composed entirely of locally reasonable judgments can still arrive, without any single decision that looks wrong from where it was made, at a global order that is not defensible at all.

22.5 Conclusion: The Recurring Shape

This monograph opened, in Chapter 1, with a trace that can be true and fully evidenced long before any system is equipped to recognize it as such, a gap between what exists and what is admitted. It closes with the same gap, restated at the scale of an entire civilization's budget: allocations that have been admitted without ever having been checked, and necessities that have never been admitted despite deserving to be. Between these two points, the same shape recurred at every register this book examined: the distinction-holonomy apparatus of Chapter 6 showed that a system's momentary state never determines whether it will continue to behave the same way under transport, exactly as this chapter's admissible region $\mathcal{A}_t$ never determines whether an allocation remains warranted; the CLIO architecture of Chapter 10 showed that verification constrains continuation without authorizing commitment, exactly as this chapter's category of claims shows that a claim's mere existence never supplies the morphism that would let it compose into a funded outcome; and the sheaf-theoretic obstruction examined independently in physics, in cryptocurrency markets, and now in the allocation of a society's resources shows that local coherence, however carefully and repeatedly checked, is never a substitute for verifying the one property, global compatibility, productive selection, universal admissible necessity, that a system's continued legitimacy actually depends on.

None of this was planned as a single argument at the outset of the corpus this monograph draws together. That the same obstruction keeps reappearing, independently motivated each time by a different domain's own internal questions, is not proof that the pattern is fundamental in some deeper metaphysical sense. It is evidence of something more modest and, for a closing chapter, more useful: that the questions this book has been asking, what persists, what may be repaired, what a representation actually preserves, what a record warrants, and what a society is entitled to call affordable, are not five different questions after

all, but one question, asked patiently, in whichever vocabulary the object at hand happened to require.

Chapter Sources and Dependencies

The table below records, for each chapter, the source material it was drafted from and the chapters it depends on for cross-referenced apparatus. Sources not authored by Flyxion are noted explicitly; all such material is examined as an external case study rather than presented as original to this monograph.

Ch.	Title	Source(s)	Depends on
1	The Latency of Evidence	*latency-of-evidence*	—
2	Accommodation Before Prediction	*accommodation-before-prediction*; *narrative-before-mechanism*	—
3	The Theatre of Agreement	*consensus-without-independence*	—
4	The Paracosm Trap	*glass-meridian*	—
5	Connective Bridge: The Ontological Reversal	Fresh chapter, grounded in *distinction-holonomy*'s opening argument	Ch. 1–4
6	Distinction Holonomy Theory	*distinction-holonomy* (Part I)	Ch. 5
7	Spherepop: An Event-Sourced Calculus of Time	*distinction-holonomy* (Part II)	Ch. 6
8	TARTAN: Recursive Multiscale Tiling	*distinction-holonomy* (Part II); Buchanan, "Everything is Contextual" [93]	Ch. 6, 7
9	The Discrete Toy Model	*distinction-holonomy* (App. D)	Ch. 6
10	The CLIO Architecture	*from-verification-to-commitment*	Ch. 1, 2, 3, 6, 7, 8
11	The Constitutional Space of Refusal	*Rebel Without a Cost* (§10–10.2)	Ch. 10
12	Verify Seams and Workarounds	*two-verify-seams*; *epistemology-of-repair*; *clarification-without-retrieval*	Ch. 10
13	Verification Inheritance	*manifest-protocol-verification-inheritance*; *verification-inheritance-transformation*; *clarification-without-retrieval*	Ch. 10, 12
14	Model Capsules and Ternary Learning	*deployment_native_ternary_learning*	Ch. 13
15	The Compression Fallacy	*compression-fallacy*	Ch. 14
16	Compression After Expansion	*compression_after_expansion*; *Borrowed_Intuition*	Ch. 15
17	Sparse Recursive Holographic Steganography	*sparse-recursive-holographic-steganography*	Ch. 15
18	Clifford FHE as the Inverse Compression Case	Silva, *Merits of Geometric Algebra* [124]	Ch. 15
19	Ledgers Without Value	*ledger-without-value*	Ch. 15
20	Rebel Without a Cost	*Rebel Without a Cost* (§1–9)	Ch. 7, 10, 11, 15, 16, 19

Ch.	Title	Source(s)	Depends on
21	Unfinishable Games and Temporal Extraction	*unfinishable-games*	Ch. 20
22	Sheaves of Necessity and Affordability	*debt-before-deflation* (§8–Postscript)	Ch. 1, 2, 6, 8, 10, 15, 19, 20, 21

Except where marked otherwise above (Chapters 8 and 18, and the contextuality and geometric-algebra papers each cite), all source material is authored by Flyxion. A full dependency graph and drafting-phase ordering is maintained separately in the project's `dependency-ledger.md`.

Bibliography

[1] S. Alexander, "Book Review: If Anyone Builds It, Everyone Dies," *Astral Codex Ten*, September 2025.

[2] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mané, "Concrete Problems in AI Safety," arXiv:1606.06565, 2016.

[3] R. U. Ayres, B. Warr, *The Economic Growth Engine: How Energy and Work Drive Material Prosperity*, Edward Elgar, 2009.

[4] G. A. Akerlof, "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics*, 84(3), 488–500, 1970.

[5] K. J. Arrow, "Economic Welfare and the Allocation of Resources for Invention," in *The Rate and Direction of Inventive Activity*, Princeton University Press, 1962.

[6] J. A. Bargh, "The Ecology of Automaticity: Toward Establishing the Conditions Needed to Produce Automatic Processing Effects," *American Journal of Psychology*, 105(2), 181–199, 1992.

[7] J. A. Bargh, K. L. Schwader, S. E. Hailey, R. L. Dyer, E. J. Boothby, "Automaticity in Social-Cognitive Processes," *Trends in Cognitive Sciences*, 16(12), 593–605, 2012.

[8] G. Bateson, *Steps to an Ecology of Mind*, Chandler Publishing, 1972.

[9] G. Bateson, D. D. Jackson, J. Haley, J. Weakland, "Toward a Theory of Schizophrenia," *Behavioral Science*, 1(4), 251–264, 1956.

[10] U. Beck, *Risk Society: Towards a New Modernity*, Sage, 1992.

[11] J. Berger, K. L. Milkman, "What Makes Online Content Viral?" *Journal of Marketing Research*, 49(2), 192–205, 2012.

[12] J. L. Borges, "The Analytical Language of John Wilkins," trans. R. L. C. Simms, in *Other Inquisitions, 1937–1952*, University of Texas Press, 1964.

[13] N. Bostrom, "Existential Risk Prevention as Global Priority," *Global Policy*, 4(1), 15–31, 2013.

[14] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014.

[15] E. S. Cahn, *No More Throw-Away People: The Co-Production Imperative*, Essential Books, 2000.

[16] E. J. Candès, T. Tao, "Near-Optimal Signal Recovery from Random Projections: Universal Encoding Strategies?" *IEEE Transactions on Information Theory*, 52(12), 5406–5425, 2006.

[17] K. Čapek, *R.U.R. (Rossum's Universal Robots)*, Aventinum, 1920.

[18] A. Church, "A Note on the Entscheidungsproblem," *Journal of Symbolic Logic*, 1936.

[19] A. C. Clarke, *2001: A Space Odyssey*, Hutchinson, 1968.

[20] M. Csikszentmihalyi, *Flow: The Psychology of Optimal Experience*, Harper & Row, 1990.

[21] H. E. Daly, *Beyond Growth: The Economics of Sustainable Development*, Beacon Press, 1996.

[22] A. de Vries, "Bitcoin's Growing Energy Problem," *Joule*, 2(5), 801–805, 2018.

[23] A. Dietrich, "Functional Neuroanatomy of Altered States of Consciousness: The Transient Hypofrontality Hypothesis," *Consciousness and Cognition*, 12(2), 231–256, 2003.

[24] A. Dietrich, "Neurocognitive Mechanisms Underlying the Experience of Flow," *Consciousness and Cognition*, 13(4), 746–761, 2004.

[25] J. St. B. T. Evans, K. E. Stanovich, "Dual-Process Theories of Higher Cognition: Advancing the Debate," *Perspectives on Psychological Science*, 8(3), 223–241, 2013.

[26] I. Fisher, "The Debt-Deflation Theory of Great Depressions," *Econometrica*, 1(4), 337–357, 1933.

[27] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, D. B. Rubin, *Bayesian Data Analysis*, 3rd ed., CRC Press, 2013.

[28] N. Georgescu-Roegen, *The Entropy Law and the Economic Process*, Harvard University Press, 1971.

[29] A. Gopnik, "Childhood as a Solution to Explore–Exploit Tensions," *Philosophical Transactions of the Royal Society B*, 375(1803), 20190502, 2020.

[30] D. Graeber, *Bullshit Jobs: A Theory*, Simon & Schuster, 2018.

[31] R. Greenblatt et al., "Alignment Faking in Large Language Models," Anthropic and Redwood Research, 2024.

[32] S. I. Hayakawa, *Language in Action*, Harcourt, Brace and Company, 1941.

[33] D. Hicks, P. Wouters, L. Waltman, S. de Rijcke, I. Rafols, "The Leiden Manifesto for Research Metrics," *Nature*, 520, 429–431, 2015.

[34] E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, S. Garrabrant, "Risks from Learned Optimization in Advanced Machine Learning Systems," arXiv:1906.01820, 2019.

[35] J. P. A. Ioannidis, "Why Most Published Research Findings Are False," *PLOS Medicine*, 2(8), 2005.

[36] E. Kal, R. Prosée, M. Winters, J. van der Kamp, "Does Implicit Motor Learning Lead to Greater Automatization of Motor Skills Compared to Explicit Motor Learning? A Systematic Review," *PLOS ONE*, 13(9), e0203591, 2018.

[37] D. Kahneman, *Thinking, Fast and Slow*, Farrar, Straus and Giroux, 2011.

[38] R. E. Kasperson et al., "The Social Amplification of Risk: A Conceptual Framework," *Risk Analysis*, 8(2), 177–187, 1988.

[39] D. L. King, P. H. Delfabbro, "Predatory Monetization Schemes in Video Games (e.g. 'Loot Boxes') and Internet Gaming Disorder," *Addiction*, 113(11), 1967–1969, 2018.

[40] D. L. King, P. H. Delfabbro, "Video Game Monetization (e.g., Loot Boxes): A Blueprint for Practical Social Responsibility Measures," *International Journal of Mental Health and Addiction*, 17, 166–179, 2019.

[41] A. Korzybski, *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*, International Non-Aristotelian Library Publishing Company, 1933.

[42] R. Koster, *A Theory of Fun for Game Design*, Paraglyph Press, 2005.

[43] J. Kücklich, "Precarious Playbour: Modders and the Digital Games Industry," *Fibreculture Journal*, no. 5, 2005.

[44] J. Lacan, *Écrits: A Selection*, trans. A. Sheridan, W. W. Norton, 1977.

[45] W. Lang (director), *Desk Set*, Twentieth Century-Fox, 1957.

[46] S. Levine, "The Full-Time Guild Master," *Intersect*, 1(1), 36–42, Stanford University, 2008.

[47] P. Lipton, *Inference to the Best Explanation*, 2nd ed., Routledge, 2004.

[48] C. E. Maine, *B.E.A.S.T.*, Ballantine Books, 1966.

[49] K. Marx, *Capital: A Critique of Political Economy, Vol. I*, Otto Meissner Verlag, 1867.

[50] M. E. McCombs, D. L. Shaw, "The Agenda-Setting Function of Mass Media," *Public Opinion Quarterly*, 36(2), 176–187, 1972.

[51] A. Narayanan, J. Bonneau, E. Felten, A. Miller, S. Goldfeder, *Bitcoin and Cryptocurrency Technologies*, Princeton University Press, 2016.

[52] B. A. Nardi, *My Life as a Night Elf Priest: An Anthropological Account of World of Warcraft*, University of Michigan Press, 2010.

[53] R. Ngo, L. Chan, S. Mindermann, "The Alignment Problem from a Deep Learning Perspective," arXiv:2209.00626, 2022.

[54] H. T. Odum, *Environmental Accounting: Emergy and Environmental Decision Making*, Wiley, 1996.

[55] S. M. Omohundro, "The Basic AI Drives," in *Proceedings of the First AGI Conference*, IOS Press, 2008.

[56] Open Science Collaboration, "Estimating the Reproducibility of Psychological Science," *Science*, 349(6251), 2015.

[57] T. Ord, *The Precipice: Existential Risk and the Future of Humanity*, Bloomsbury, 2020.

[58] E. Perez et al., "Discovering Language Model Behaviors with Model-Written Evaluations," *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

[59] K. R. Popper, *The Logic of Scientific Discovery*, Hutchinson, 1959.

[60] W. V. O. Quine, "Two Dogmas of Empiricism," *The Philosophical Review*, 60(1), 20–43, 1951.

[61] P. M. Romer, "Endogenous Technological Change," *Journal of Political Economy*, 98(5), S71–S102, 1990.

[62] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27(3), 379–423; 27(4), 623–656, 1948.

[63] M. Sharma et al., "Towards Understanding Sycophancy in Language Models," arXiv:2310.13548, 2023.

[64] M. Shelley, *Frankenstein; or, The Modern Prometheus*, Lackington, Hughes, Harding, Mavor & Jones, 1818.

[65] R. J. Shiller, *Irrational Exuberance*, Princeton University Press, 2000.

[66] P. Singer, "Famine, Affluence, and Morality," *Philosophy & Public Affairs*, 1(3), 229–243, 1972.

[67] P. Singer, *Animal Liberation*, HarperCollins, 1975.

[68] B. F. Skinner, *Science and Human Behavior*, Macmillan, 1953.

[69] P. Slovic, "Perception of Risk," *Science*, 236(4799), 280–285, 1987.

[70] A. Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations*, W. Strahan and T. Cadell, 1776.

[71] P. K. Stanford, *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*, Oxford University Press, 2006.

[72] C. R. Sunstein, *Laws of Fear: Beyond the Precautionary Principle*, Cambridge University Press, 2005.

[73] E. Szpilrajn, "Sur l'extension de l'ordre partiel," *Fundamenta Mathematicae*, 16, 386–389, 1930.

[74] N. N. Taleb, *The Black Swan: The Impact of the Highly Improbable*, Random House, 2007.

[75] J. Truby, "Decarbonizing Bitcoin: Law and Policy Choices for Reducing the Energy Consumption of Blockchain Technologies," *Energy Research & Social Science*, 44, 399–410, 2018.

[76] M. Turpin, J. Michael, E. Perez, S. R. Bowman, "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting," *Advances in Neural Information Processing Systems*, 36, 2023.

[77] A. Tversky, D. Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science*, 185(4157), 1124–1131, 1974.

[78] M. Ulrich, J. Keller, K. Hoenig, C. Waller, G. Grön, "Neural Correlates of Experimentally Induced Flow Experiences," *NeuroImage*, 86, 194–202, 2014.

[79] B. C. van Fraassen, *The Scientific Image*, Oxford University Press, 1980.

[80] A. E. van Vogt, *The World of Null-A*, Simon & Schuster, 1948.

[81] H. R. Varian, "Artificial Intelligence, Economics, and Industrial Organization," in *The Economics of Artificial Intelligence*, University of Chicago Press, 2019.

[82] S. Vosoughi, D. Roy, S. Aral, "The Spread of True and False News Online," *Science*, 359(6380), 1146–1151, 2018.

[83] T. Wu, *The Attention Merchants: The Epic Scramble to Get Inside Our Heads*, Knopf, 2016.

[84] N. Yee, "The Daedalus Project: Life as a Guild Leader," research archive, nickyee.com, 2006.

[85] D. Zendle, P. Cairns, "Loot Boxes are Again Linked to Problem Gambling: Results of a Replication Study," *PLOS ONE*, 14(3), e0213194, 2019.

[86] S. Zuboff, *The Age of Surveillance Capitalism*, PublicAffairs, 2019.

[87] M. R. Albrecht, B. R. Curtis, A. Deo, A. Davidson, R. Player, E. W. Postlethwaite, F. Virdia, T. Wunderer, "Estimate All the {LWE, NTRU} Schemes!" in *Proc. Int. Conf. Security and Cryptography for Networks (SCN)*, Springer, 351–367, 2018.

[88] Y. Bai, S. Kadavath, S. Kundu et al., "Constitutional AI: Harmlessness from AI Feedback," arXiv:2212.08073, 2022.

[89] S. Baluja, "Hiding Images in Plain Sight: Deep Steganography," *Advances in Neural Information Processing Systems*, 30, 2017.

[90] B. Beyer, C. Jones, J. Petoff, N. R. Murphy (eds.), *Site Reliability Engineering: How Google Runs Production Systems*, O'Reilly Media, 2016.

[91] M. Blaze, J. Feigenbaum, J. Lacy, "Decentralized Trust Management," in *Proc. IEEE Symposium on Security and Privacy*, 1996.

[92] M. M. Bronstein, J. Bruna, T. Cohen, P. Veličković, "Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges," arXiv:2104.13478, 2021.

[93] R. J. Buchanan, "Everything is Contextual: Contextuality Beyond Quantum," preprint, 2026.

[94] C. Cachin, "An Information-Theoretic Model for Steganography," *Information and Computation*, 2004.

[95] R. Q. Charles, H. Su, K. Mo, L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017.

[96] J. H. Cheon, A. Kim, M. Kim, Y. Song, "Homomorphic Encryption for Arithmetic of Approximate Numbers," in *Proc. ASIACRYPT*, Springer, 409–437, 2017.

[97] S. Dathathri et al., "Scalable Watermarking for Identifying Large Language Model Outputs," *Nature*, 2024.

[98] L. Dorst, D. Fontijne, S. Mann, *Geometric Algebra for Computer Science: An Object-Oriented Approach to Geometry*, Morgan Kaufmann, 2007.

[99] C. Ellison et al., "SPKI Certificate Theory," RFC 2693, 1999.

[100] P. Feyerabend, *Against Method*, New Left Books, 1975.

[101] J. Fridrich, J. Kodovský, "Rich Models for Steganalysis of Digital Images," *IEEE Transactions on Information Forensics and Security*, 7(3), 868–882, 2012.

[102] C. Gentry, "A Fully Homomorphic Encryption Scheme," PhD thesis, Stanford University, 2009.

[103] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, "Generative Adversarial Nets," *Advances in Neural Information Processing Systems*, 27, 2672–2680, 2014.

[104] P. Groth, A. Gibson, J. Velterop, "The Anatomy of a Nanopublication," *Information Services and Use*, 2010.

[105] X. Guix, E. Fontich, R. Garcia et al., "Nix Based Fully Automated Workflows and Ecosystem to Guarantee Scientific Result Reproducibility Across Software Environments and Systems," in *Proc. 1st Int. Workshop on Software Engineering for HPC in Computational Science and Engineering*, 2015.

[106] S. Halevi, V. Shoup, "Faster Homomorphic Linear Transformations in HElib," in *Proc. CRYPTO*, Springer, 93–120, 2018.

[107] M. Herbert, A. Pires, "Bilingualism and Bimodal Code-Blending Among Deaf ASL-English Bilinguals," *Proceedings of the Linguistic Society of America*, 2, 14:1–15, 2017.

[108] "Last Shelter: Survival Answers and Guides," player-produced FAQ documentation, AppGamer, accessed 2026.

[109] IM30 Technology Limited, "Last Shelter: Survival," Apple App Store listing, accessed 2026.

[110] Long Tech Network Limited, "Last Shelter: Survival," Google Play Store listing, accessed 2026.

[111] "Last Shelter Survival: State vs State Guide," player-produced community guide, Steam Community, accessed 2026.

[112] V. Lyubashevsky, C. Peikert, O. Regev, "On Ideal Lattices and Learning with Errors over Rings," in *Proc. EUROCRYPT*, Springer, 1–23, 2010.

[113] S. Ma, H. Wang, L. Ma, L. Wang, W. Wang, S. Huang, L. Dong, R. Wang, J. Xue, F. Wei, "The Era of 1-Bit LLMs: All Large Language Models are in 1.58 Bits," arXiv:2402.17764, 2024.

[114] A. J. Mercado, A. Lomuscio, "Formal Verification of Agentic Systems over Operational Data," arXiv:2608.03609, 2026.

[115] P. Moulin, J. A. O'Sullivan, "Information-Theoretic Analysis of Information Hiding," *IEEE Transactions on Information Theory*, 2003.

[116] Norwegian Consumer Council (Forbrukerrådet), "Insert Coin: How the Gaming Industry Exploits Consumers Using Loot Boxes," report backed by more than twenty consumer organizations across eighteen European countries, May 2022.

[117] L. Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback," *Advances in Neural Information Processing Systems*, 35, 27730–27744, 2022.

[118] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed., Cambridge University Press, 2009.

[119] The Reproducible Builds Project, *Reproducible Builds: A Set of Software Development Practices That Create an Independently Verifiable Path from Source to Binary Code*, reproducible-builds.org.

[120] T. Ringer et al., "Proof Repair across Type Equivalences," in *Proc. ACM SIGPLAN Conf. on Programming Language Design and Implementation (PLDI)*, 2020.

[121] R. L. Rivest, B. Lampson, "SDSI: A Simple Distributed Security Infrastructure," 1996.

[122] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, T. Scialom, "Toolformer: Language Models Can Teach Themselves to Use Tools," *Advances in Neural Information Processing Systems*, 36, 68539–68551, 2023.

[123] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, R. Anderson, "AI Models Collapse When Trained on Recursively Generated Data," *Nature*, 631, 755–759, 2024.

[124] D. W. Silva, "Merits of Geometric Algebra Applied to Cryptography and Machine Learning," *Philosophical Transactions of the Royal Society A*, 384, 20250107, 2026.

[125] G. J. Simmons, "The Prisoners' Problem and the Subliminal Channel," in *Advances in Cryptology: Proc. CRYPTO 83*, Plenum Press, 51–67, 1984.

[126] N. Thomas, T. E. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, P. Riley, "Tensor Field Networks: Rotation- and Translation-Equivariant Neural Networks for 3D Point Clouds," arXiv:1802.08219, 2018.

[127] V. Vaikuntanathan, "The Wonders of Lattice-Based Cryptography," Simons Institute Cryptography Boot Camp lecture, UC Berkeley, 2015.

[128] F. Voboril, S. Szeider, "Improving Constraint Models with LLM Agents," arXiv:2608.08127, 2026.

[129] H. Wang, S. Ma, L. Dong, S. Huang, H. Wang, L. Ma, F. Yang, R. Wang, F. Wei, "BitNet: Scaling 1-Bit Transformers for Large Language Models," arXiv:2310.11453, 2023.

[130] J. Wang, H. Zhou, T. Song, S. Cao, Y. Xia, T. Cao, J. Wei, S. Ma, H. Wang, F. Wei, "Bitnet.cpp: Efficient Edge Inference for Ternary LLMs," in *Proc. ACL 2025*, 9256–9270, 2025.

[131] World Wide Web Consortium, *PROV-O: The PROV Ontology*, W3C Recommendation.

[132] World Wide Web Consortium, *Verifiable Credentials Data Model*, W3C Recommendation.

[133] Coalition for Content Provenance and Authenticity, *C2PA Technical Specification*.

[134] L. Y. Xiao, "Regulating Loot Boxes as Gambling? Towards a Combined Legal and Self-Regulatory Consumer Protection Approach," *Interactive Entertainment Law Review*, 4(1), 27–47, 2021.