Mechanistic Reference: When Does a Mechanism Term Pick Out the Same Thing?

Elliot Tower
Independent Researcher
elliot@elliottower.ai

Abstract

Neither mechanistic interpretability nor the philosophy of mechanisms provides an account of cross-system mechanistic identity: the MDC account, Craver's mutual manipulability, and Woodward's interventionism all individuate mechanisms within a system and say nothing about when two mechanisms, in different systems, are the same mechanism. This gap matters because mechanism terms do not stay home—"the IOI circuit," "the mirror neuron system," "the language gene" are all applied across system boundaries as a matter of routine scientific practice. Many apparent replication failures are better understood as reference failures: the original finding and the replication are both valid within their domains, but the mechanism term does not refer across the boundary. Standard statistical and methodological safeguards are not designed to detect this failure mode.

Building on the Mechanistic Views and Mechanistic Validity frameworks (**??**), we formalize cross-system mechanistic reference using a transport hierarchy that specifies, for each level of mechanistic commitment, what evidence is required for a term to refer across systems and what inferences that reference licenses. We prove that inferential reach is strictly ordered: no quantity of behavioral evidence can substitute for structural evidence, regardless of accumulation. We define five reference failure modes—evidence misfire, claim laundering, mimic mechanism, reference debt, and zombie mechanism—ordered by severity and remediation cost, each triggering a distinct contraction pattern. We provide a Levi-rational contraction algorithm for principled partial revision under anomalous evidence: discard the weakest inferential steps, contract at the periphery not the center.

We motivate and test the framework by applying it to published findings on "the IOI circuit." Object-level transport fails even across random seeds of the same architecture—the heads that matter most functionally are the least stable, with mid-layer stability dropping to $\sim$0.70 (**?**). Role-level transport partially succeeds across seeds and scales (**?**). Structural-level transport fails across model families, with cross-family congruence dropping to 0.13 (**?**). The term "the IOI circuit" does not refer to the same object in GPT-2 and Pythia—not because either analysis is wrong, but because no transport condition at the structural level is satisfied. We apply the framework through fifteen worked examples drawn from mechanistic interpretability, neuroscience, pharmacology, and genetics—including cross-species mechanism transport in drug development (mTOR/rapamycin) and population-level overgeneralization in neurodegeneration research (the amyloid cascade hypothesis—a causal mechanism established in rare familial Alzheimer's variants applied wholesale to sporadic disease, driving $40B+ in failed trials)—and offer forward predictions about mechanism terms currently accumulating reference debt, several of which are now partially confirmed by independent empirical work.

1 Introduction

Consider three cases from recent science, each involving a mechanism term that appears to denote a stable, reusable explanatory object.

First, **?** identify what they call "the indirect object identification circuit" in GPT-2 Small, a network of attention heads that collectively implement name-copying behavior. The circuit has become a benchmark in mechanistic interpretability: other groups test whether "the IOI circuit" exists in Qwen, Gemma, Llama, and other architectures. In each case, the behavioral signature—correctly copying indirect object names—appears robustly across models. But recent work on numerical comparison circuits has demonstrated that "task behavior similarities do not imply mechanistic universality" (**?**). The models solve the same task via different internal mechanisms. When a researcher in 2026 says "we found the IOI circuit in Llama-3," does that term refer to the same mechanism that Wang et al. discovered in GPT-2?

Second, **?** describe "induction heads"—attention heads that implement a particular copying pattern—in one- and two-layer transformers, and present evidence that this mechanism underlies in-context learning. The term has since been applied to models with hundreds of billions of parameters, trained on vastly more data, exhibiting qualitatively different behavior. Work on "Selective Induction Heads" (**?**) has shown that what these larger models do is substantially more complex than what two-layer models do. The mechanism term was coined in a simplified setting and exported to systems where the referent, if it exists at all, has a different internal structure.

Third, when pharmacologists establish that "the mTOR pathway mediates the longevity effect of caloric restriction" in *Mus musculus*, and propose rapamycin as a human therapeutic on that basis, they are engaging in cross-species mechanism transport. The molecular components of the mTOR pathway are conserved across mammals, but the pathway's downstream effects, regulatory context, and interaction with other physiological systems differ substantially. The 90% failure rate of clinical drug development (**?**) suggests that this kind of cross-system mechanism transport fails far more often than it succeeds.

None of these questions can be answered by collecting more data. Each requires a prior specification of what "the same mechanism" means—a theory of mechanistic reference that determines when a term validated in one system picks out the same thing in another. And when that reference fails—as it frequently does—the scientist needs a principled method for revising their claims: not abandoning everything, but giving up exactly the right parts. The field currently lacks both.

The philosophy of mechanisms has made enormous progress on mechanism individuation *within* a system. The MDC account (**?**) characterizes mechanisms as "entities and activities organized such that they are productive of regular changes." **?** provides mutual manipulability criteria for constitutive relevance. **?** develops an account grounded in invariant change-relating generalizations. **?** provides an interventionist framework for causal claims. These accounts tell us what a mechanism *is*, and what evidence warrants claiming that a mechanism exists in a particular system. What they do not provide is an account of when two mechanisms, individually well-validated in their respective systems, are the same mechanism—or what to do when a mechanism term that seemed to refer across systems turns out not to.

This gap matters because mechanism terms do not stay home. Scientists generalize: from one model to another, from mouse to human, from one brain region to another, from one cell type to another. Every such generalization presupposes that the mechanism term refers across the boundary. When the presupposition holds, generalization is the engine of scientific progress. When it fails, the result is not merely a mistake but a particular kind of mistake—a reference failure—that standard statistical and methodological safeguards are not designed to detect.

We argue that many apparent "replication failures" are better understood as reference failures. The original study and the replication measured different things because the mechanism term was applied in a context where its referent did not transport. The original finding may be perfectly correct within its domain; the replication may be perfectly well-conducted within its own. The failure lies not in the statistics or the methods but in the unexamined assumption that the mechanism term denotes the same object in both contexts.

The central argument of this paper is as follows. A mechanistic view σ, in the sense of **?**, specifies an identity criterion $\sim$ over a domain of objects O. This identity criterion induces a quotient space $O/\sim$ whose equivalence classes are the candidates for stable reference. Within a single system, reference succeeds when there exists an equivalence class $[m] \in O/\sim$ that is stable under admissible measurement perturbations.

Cross-system reference requires something more: an *admissible transport map* $T : O_A/\sim_A \to O_B/\sim_B$ that preserves identity-relevant predicates. The existence of such a map is not guaranteed, and its failure is detectable. We characterize what admissibility requires for each view type in the commitment hierarchy, define five modes of reference failure, prove that inferential reach is strictly ordered, and—crucially—provide a formal theory of what to do when reference fails: a Levi-rational contraction algorithm that preserves the most entrenched sub-claims while discarding the least supported ones.

This paper builds on two precursors. **?** introduced the MechViews framework, which formalized mechanistic views as 5-tuples and showed that view incoherence generates unresolvable conflicts. **?** introduced MechVal, which provided evidence standards calibrated to view type. The present paper addresses two questions that both presupposed but did not answer: when does a mechanism term, fully validated in one context, pick out the same thing in another? And when it doesn't, what should the scientist do?

Readers familiar with software engineering may find the following analogy clarifying. A mechanism term is like a function name in a shared library: it promises that the same interface will behave the same way in any program that imports it. Reference failure is what happens when the interface signature matches but the implementation differs—a subtle type error that passes compilation but fails at runtime. The transport hierarchy formalizes what "the same behavior" means at each level of abstraction, and the contraction algorithm provides a principled procedure for downgrading the interface contract when a mismatch is detected, rather than discarding the library entirely.

The paper proceeds as follows. Section **??** traces a single mechanism term—"the IOI circuit"—through the transport hierarchy, using recent empirical results to motivate the formal framework. Section **??** reviews the relevant background, including a detailed treatment of belief revision frameworks and why standard approaches fail for mechanism claims. Section **??** develops the framework: within-system reference, cross-system transport, reference failure modes, the contraction algorithm, the inferential reach lattice, and the scaffold-path decomposition. Section **??** applies the framework through worked examples. Section **??** addresses pluralism, realism, forward predictions, and limitations. Section **??** concludes.

2 From Term to Transport: The IOI Circuit

Before developing the formal framework, we trace what happens when a single mechanism term—"the IOI circuit"—is transported across systems. This walkthrough motivates the transport hierarchy and demonstrates why a theory of mechanistic reference is needed.

? identified the IOI circuit in GPT-2 Small as a network of 26 attention heads across 7 functional roles: duplicate token detectors, S-inhibition heads, name mover heads, and backup name movers. The term has since been applied to Pythia, Qwen, Gemma, Llama, mGPT, and even non-transformer architectures. At each step, a researcher claimed to have found "the IOI circuit." But what exactly transported?

Recent work allows us to answer this question at each level of the commitment hierarchy.

Object level: heads do not transport. **?** train 50 independent refits of GPT-2-scale models with identical architecture and data but different random seeds. The result is striking: middle-layer attention heads—precisely the ones that constitute the IOI circuit—are the *least* stable across refits, with stability dropping to approximately 0.70 in layer 5 of an 8-layer model. Worse, the heads that matter most functionally (those whose ablation causes the largest perplexity increase) are the least stable. Even more troubling, **?** show that within a single model on a single task, different prompt templates activate different circuit structures: GPT-2's IOI circuit for ABBA prompts is distinct from its circuit for BABA prompts. Object-level transport—"head 9.9 is the name mover head"—fails not only across random seeds but across prompt formats.

Role level: functional roles partially transport. Despite the object-level instability, functional roles show greater stability. **?** find that head 9.9 serves as the name mover in both ABBA and BABA templates—the same role, implemented via completely different input signals (negative cosine similarity between the signals in the two contexts). The role transports; the implementation does not. **?** confirm that the same

algorithmic roles persist across training checkpoints and model scale: the "name mover" role is reproducible even when the specific head occupying it changes. **?** extend this finding across tasks: 78% of the heads in the IOI circuit also appear in the Colored Objects circuit, suggesting that the same components are recruited for the same functional roles across distinct tasks.

Subspace level: representations transport within families. **?** show that despite dramatic head-level instability, the residual stream is consistently stable across refits (measured by CKA). The representations converge even when the attention heads producing them diverge. **?** quantify this precisely using their CONGRUITY algorithm, which measures interpretive equivalence without requiring explicit interpretation: within model families (Pythia 160M to 2.8B), congruence ranges from 0.89 to 0.99. A model 17.5× larger implements the same IOI interpretation as its smallest family member. Subspace-level transport holds robustly within families.

Structural level: structure does not transport across families. **?**'s most consequential finding is the cross-family gap: GPT-2 vs. Pythia congruence on IOI drops to 0.13, compared to 0.73–0.92 within families. Different model families implement IOI via structurally different circuits, even though both solve the task correctly. The term "the IOI circuit" does not refer to the same structural object in GPT-2 and Pythia. **?** add a further dimension: cross-language circuit distance correlates with linguistic typological distance ($r = 0.83$–0.88 for syntactic and genetic measures), meaning structural transport is graded rather than binary.

Process level: the learning problem constrains but does not determine. The typological-distance correlation suggests that the process by which circuits form is partially determined by the structure of the learning problem: related languages produce more similar circuits, and circuits cluster first by template structure and only secondarily by language. The computational problem constrains the solution space but does not determine a unique circuit. **?**'s finding that AdamW substantially improves head stability over Adam, with no performance difference, reveals that training process choices modulate the degree of transport achievable at every level.

The walkthrough reveals the central claim of this paper in microcosm: *whether "the IOI circuit" refers across systems depends entirely on which level of the transport hierarchy the claim is made at.* At the Object level, the term fails even across random seeds. At the Role level, it partially succeeds. At the Subspace level, it succeeds within model families. At the Structural level, it fails across families. When a researcher says "we found the IOI circuit in Llama," the question is not whether this is true or false but what the term *refers to* under the operative view—and whether the transport conditions for that view are satisfied.

3 Background

3.1 The Philosophy of Mechanisms

The new mechanistic philosophy emerged in the late 1990s and early 2000s as a response to the covering-law model of explanation. Where the covering-law model took explanation to consist in subsumption under universal regularities, the mechanistic alternative holds that explanation consists in describing the mechanism responsible for the phenomenon of interest—identifying the relevant entities and activities, their organization, and how that organization is productive of the phenomenon.

The foundational statement is due to **?**, who define mechanisms as "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions." This definition emphasizes that mechanisms are not mere lists of parts but organized systems: the entities must be arranged in particular ways, and the activities must be temporally and spatially coordinated, for the mechanism to produce its characteristic phenomenon.

? provides a complementary account grounded in invariant change-relating generalizations. The emphasis on invariance connects Glennan's account to **?**'s interventionist framework: the interactions between mechanism

components are those that remain stable under a range of interventions. This provides a criterion for distinguishing genuine mechanism components from correlates.

? develops the most detailed account of constitutive relevance. Craver's mutual manipulability criterion states that a component X is constitutively relevant to a mechanism M if and only if (i) manipulating X changes M's behavior and (ii) manipulating M changes X's behavior. This bidirectional criterion rules out mere correlates.

For our purposes, the crucial observation is that all these accounts are theories of mechanism individuation *within a system*. They do not provide identity criteria that cross system boundaries.

3.2 Natural Kinds and Reference

The question of cross-system mechanistic identity is, at bottom, a question about reference: when does a mechanism term, introduced in one context, successfully denote an object in another?

The causal theory of reference, developed by **?** and **?**, holds that names and natural kind terms refer via causal chains originating in a "baptism" event where the term is introduced in direct contact with its referent. Mechanism terms aspire to natural kind status. When an interpretability researcher says "induction head," they intend to pick out a natural kind of neural network component. But mechanism terms cannot achieve Kripkean rigidity through the standard causal-chain route. There is no baptism event where one stands in direct causal contact with "the induction head." The term is introduced through a combination of behavioral observations and structural analysis, and its reference is fixed by a specification—implicit or explicit—of what properties an entity must have to count as an instance of the kind. This means that mechanism reference must be earned through what we will call *structural stabilization*: the demonstration that the term's referent is stable under the identity criteria specified by the operative mechanistic view.

? argues that natural kinds in the special sciences are "homeostatic property clusters"—sets of properties that tend to co-occur because of an underlying causal mechanism, but that need not all be present in every instance. We suggest that mechanism kinds are best understood as homeostatic property clusters whose co-occurrence is contingent on features of the system in question. The consequence is that cross-system reference requires explicit specification of which properties are identity-relevant and explicit verification that those properties are preserved.

3.3 Belief Revision for Mechanism Claims

When evidence shows that a mechanism claim is wrong, the scientist faces a practical question: which parts of the claim should be given up, and which should be kept? The philosophy of belief revision provides formal tools for answering this question.

When evidence reveals that a mechanism term does not refer as expected, the scientist must revise their claims. This is a problem of belief revision, and the choice of revision framework matters. We argue that standard frameworks—Bayesian updating and AGM contraction—are structurally inadequate for mechanism claims, and that **??**'s epistemology of contraction provides the right formal basis.

3.3.1 Why Bayesian Updating Fails

Bayesian approaches assign degrees of belief to propositions and update them via conditionalization. Three features of mechanism claims make Bayesian updating inappropriate.

First, Bayesian updating requires a prior probability distribution over mechanism hypotheses. But when a mechanism term might not *refer* in a new system, the sample space itself is undefined. One cannot assign $P(\text{IOI circuit exists in Llama}) = 0.7$ in any principled way, because the term "IOI circuit" might not pick out a well-defined object in Llama's quotient space at all. Reference failure is a presupposition failure, not an event in a well-defined probability space.

Second, the inferential reach theorem (Section **??**) establishes that evidence types are qualitatively different. Ten behavioral experiments do not add up to one structural experiment. Bayesian updating treats all

evidence as commensurable—different likelihood ratios operating on the same probability space. But the whole point of the commitment hierarchy is that evidence from the wrong level does not update the right quantity at all. It is not weak evidence; it is *irrelevant* evidence. Bayesianism has no mechanism for representing this qualitative incommensurability.

Third, scientists do not in practice maintain probability distributions over mechanism claims. They *accept* or *reject* claims and then reason from accepted claims as premises (we develop this point fully in the next section, where it motivates the choice of Levi's corpus model).

3.3.2 Why AGM Contraction Fails

? define rationality postulates for belief contraction on logically closed belief sets. Two features of AGM make it unsuitable for mechanism claims.

First, AGM contraction does not rank the conjuncts of a belief. When a mechanism claim is a conjunction (the mechanism exists AND it is causal AND it transports AND it uses this algorithm), AGM identifies the maximal consistent subsets but does not tell the scientist which conjuncts to give up. It offers multiple equally rational contractions and says "pick one." For mechanism claims, this indeterminacy is unacceptable: the scientist needs to know that "the Fourier circuit is sufficient to explain grokking's timing" is less entrenched than "Fourier features develop during training," and should be given up first.

Second, AGM's recovery postulate states that if you contract by A and then re-expand by A, you recover the original belief set. For mechanism claims, this is wrong. If you discover that "knowledge neurons" do not exist, contract the claim, and later find new evidence for localized factual storage, you should not restore the *original* claim with all its baggage—the term has been dissolved, the field has moved on, the word means something different now. Zombie mechanisms exist precisely because recovery fails: the term comes back but the referent does not. **?** explicitly rejects the recovery postulate, and his framework is better for it.

3.3.3 Why Levi's Framework Works

The fundamental reason to prefer Levi over Bayesian approaches is not merely formal but reflects how scientists actually reason about mechanisms. Scientists do not maintain probability distributions over mechanism claims; they accept claims and reason from them as premises. A researcher who has accepted that "the IOI circuit exists in GPT-2" does not continuously re-weight this claim; she treats it as the starting point for subsequent work. Levi's corpus model is designed for exactly this kind of agent—one who accepts propositions rather than weighs them.

Levi's central concept is the *corpus of knowledge*: the set of propositions an agent is committed to treating as certainly true for the purposes of inquiry. Unlike Bayesian approaches, Levi draws a sharp distinction between propositions that are accepted and propositions that are merely entertained. A proposition is a *serious possibility* relative to a corpus K if and only if its negation is not in K.

We argue that a mechanistic view functions as a partition generator. The identity criterion $\sim$ determines what counts as "the same" versus "different" mechanisms, and thereby partitions the space of possible mechanisms into equivalence classes. Evidence is meaningful only relative to this partition.

Contraction is the operation of removing a proposition from the corpus in response to anomalous evidence. Levi provides a rationality constraint on contraction: among the various ways of restoring consistency, the agent should choose the one that minimizes loss of *informational value*, as measured by an *entrenchment ordering* over the propositions in the corpus. The entrenchment ordering captures what AGM lacks: a principled ranking of sub-claims by their degree of embeddedness in the agent's inferential network. This maps directly onto the structure of mechanism claims, which are conjunctions of sub-claims with different degrees of empirical support.

3.3.4 Spohn's Ranking Theory as Generalization

? develops ranking theory, which handles iterated revision where Levi handles single-step contraction. A ranking function κ assigns non-negative integers to possible worlds, representing degrees of disbelief; condi-

tionalization on a ranking function produces another ranking function, which can itself be conditionalized further. This matters because mechanism claims face multiple successive anomalies over time—knowledge neurons challenged first by ROME, then by MEMIT, then by distributed storage arguments—and Spohn's iterated revision handles such sequences naturally where Levi's single-step contraction must be re-applied from scratch.

We use Levi's framework for Algorithm **??** because its entrenchment ordering maps directly onto the structure of mechanism claims. Spohn's ranking theory provides the natural generalization for iterative challenges, and we note it as the appropriate extension for the common case where mechanism claims face successive anomalies over time.

3.4 Mechanistic Views as Partition Generators

? formalize mechanistic views as 5-tuples $\sigma = (O, \sim, E, F, T)$, where O is a domain of objects, $\sim$ is an identity criterion, E is an evidence type, F is a family of allowable findings, and T is a set of licensed inference types. We adopt this formalism here, though we foreground O, $\sim$, E, and T while treating F as implicit.

The MechViews framework identifies a hierarchy of views ordered by commitment level: Object (physical constituents), Role (functional role), Subspace (representational geometry), Structural (gauge-invariant structure), and Process (developmental trajectory). Two key results from MechViews are relevant:

Proposition 2.1 (View Coherence) (from **?**). A mechanistic view is coherent if and only if the identity criterion is compatible with the evidence type.

Proposition 2.2 (Conflict Resolution) (from **?**). Two mechanism claims can be in genuine conflict only if they are made under views with the same identity criterion.

These results establish that mechanism claims are view-relative. What they do not provide—and what this paper supplies—is an account of when a mechanism term, fully validated under a given view in one system, successfully refers to a mechanism in another system, and what to do when it fails.

4 The Framework

4.1 Within-System Reference

We begin with the simpler case: what does it take for a mechanism term M to refer within a single system S?

Definition 4.1 (Within-System Reference)**.** *Let $\sigma = (O, \sim, E, F, T)$ be a mechanistic view and let S be a system. A mechanism term M* refers within S under σ *if and only if:*

(i) There exists an equivalence class $[m] \in O_S/{\sim_S}$ such that M denotes $[m]$.

(ii) $[m]$ is stable under admissible measurement perturbations.

(iii) The evidence supporting the claim is within the range of E_S.

The importance of condition (ii) is illustrated by **?**, who demonstrate that unconstrained Distributed Alignment Search achieves 100% Interchange Intervention Accuracy on *random* models. The method always finds what it is looking for, which means it never refers. This example also connects to **?**'s constructive empiricism: a method that achieves perfect empirical adequacy on every system, including random systems, has zero referential content. Constructive empiricism holds that science aims at empirical adequacy—saving the phenomena—rather than truth about unobservables. The mimic mechanism problem reveals the cost of this position for mechanism science: empirical adequacy without reference is not mechanism discovery; it is curve fitting.

4.2 Cross-System Reference and Transport

Cross-system reference is the harder case, and the one that matters most for scientific generalization.

Definition 4.2 (Cross-System Reference)**.** *Let σ be a mechanistic view, and let S_A and S_B be two systems. A mechanism term M* refers across S_A and S_B under σ *if and only if:*

(i) M refers within S_A and S_B under σ.

(ii) There exists an admissible transport map $T : O_A/\sim_A \to O_B/\sim_B$ such that $T([m_A]) = [m_B]$.

(iii) T preserves identity-relevant predicates.

What counts as "admissible" depends on the view, generating a hierarchy of transport requirements. Before presenting this hierarchy, two technical concepts require brief explanation for readers outside machine learning. A *gauge symmetry* is a transformation of a system's internal representation that leaves its observable behavior unchanged. The *Grassmannian* $\mathrm{Gr}(k, d)$ is the manifold of all k-dimensional linear subspaces of $\mathbb{R}^d$, providing a natural geometry for comparing representational subspaces.

We summarize the transport hierarchy in Table **??** and Figure **??**. The hierarchy is cumulative: each level's transport requirement strictly includes all requirements below it.

Table 1: Transport hierarchy: what each view requires for cross-system reference, and what successful transport licenses.

View	**Transport requirement**	**Licensed inferences**
Object	None (indices are system-specific)	Within-system ablation claims only
Role	Functional equivalence	Cross-system functional generalization
Subspace	Geodesic proximity on $\mathrm{Gr}(k, d)$	Variable-level intervention transfer
Structural	Gauge-invariant parallel transport	Cross-model structural comparison
Process	Dynamical trajectory equivalence	Developmental/training-origin claims

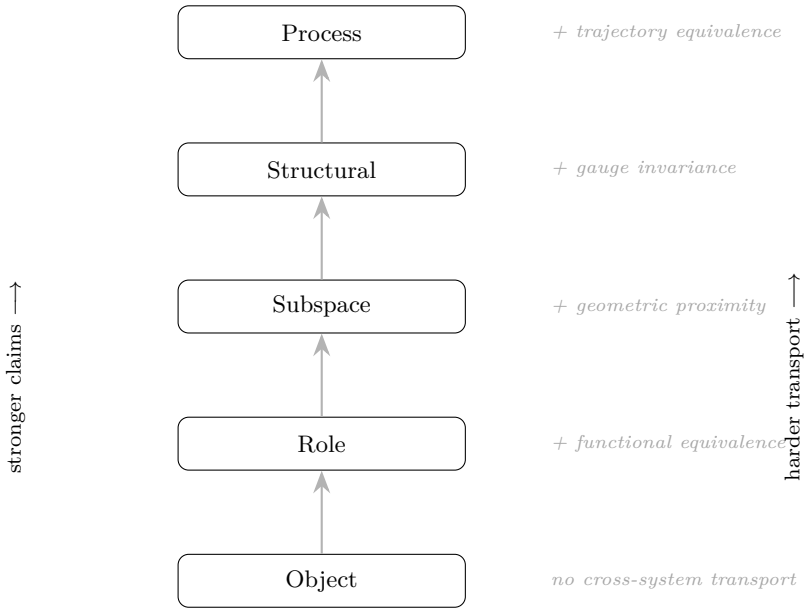

Figure 1: The transport hierarchy. Arrows indicate cumulative inclusion.

Proposition 4.3. *If admissible transport exists under view σ_j and $\sigma_i < \sigma_j$ in the commitment hierarchy, then admissible transport exists under σ_i. The converse fails (the IOI circuit demonstrates this: Role-level transport holds but Structural-level transport fails; see Section* **??***).*

By Proposition **??**, evidence at a higher level constrains but does not establish reference at lower levels. The IOI circuit illustrates both directions: Subspace-level transport within model families (Section **??**) entails Role-level transport, but the converse fails—Role-level transport across families does not entail Structural-level transport.

4.3 Reference Failure Modes

When mechanistic reference fails, it does not always fail in the same way. We identify five distinct modes, ordered by increasing severity, each corresponding to a specific violation of the conditions in Definitions **??** and **??**. Some are easy to fix (use better evidence); others require retiring terms that have become embedded in the field's vocabulary.

Evidence Misfire. The evidence is of the wrong type for the view under which the claim is made. Violates condition (iii) of Definition **??**.

Claim Laundering. Evidence admissible under one view is used to support claims characteristic of a higher-commitment view. Violates the inferential reach ordering (Theorem **??**).

Remark 4.4 (Distinguishing Misfire from Laundering)**.** *Ask: is the evidence admissible under* any *view? If no, the failure is misfire. If yes under σ_i but used for $\sigma_j > \sigma_i$, the failure is laundering.*

Mimic Mechanism. An apparent mechanism that satisfies the detection criteria without corresponding to a genuine feature of the system. Violates condition (ii) of Definition **??**.

Reference Debt. A mechanism term used in contexts where its cross-system reference has not been validated.

Zombie Mechanism. A mechanism term that continues to circulate after the referent it was introduced to denote has been shown not to exist. A zombie mechanism is distinctively dangerous because it is epistemically

contagious: unlike a merely false claim, a zombie continues to generate new claims, attract citations, and shape experimental design long after its referent has been dissolved. The term "knowledge neurons" still appears in papers citing ROME and MEMIT because the word survived the dissolution of its referent. The danger is not that the claim is wrong but that the term propagates.

These modes are ordered by increasing severity: misfires are corrected by replacing evidence; zombies require retiring terms that have become epistemically contagious.

Each failure mode triggers a different contraction pattern (Section **??**): misfires require evidence replacement, laundering requires claim downgrade, mimics require method constraint, debt requires transport validation, and zombies require term retirement.

4.4 Rational Contraction for Mechanism Claims

When evidence reveals a reference failure, the scientist must revise their mechanism claims. Simply abandoning the entire claim is often too drastic. What is needed is a principled method for partial revision—one that preserves the most well-supported sub-claims while discarding the least entrenched.

The key insight, drawing on **?**, is that mechanism claims have a structure analogous to scientific research programmes. Lakatos distinguishes the *hard core* of a research programme—fundamental assumptions "rendered irrefutable by the methodological decision of its protagonists" (p. 48)—from the *protective belt* of auxiliary hypotheses that "has to bear the brunt of tests and gets adjusted and re-adjusted, or even completely replaced, to defend the thus-hardened core" (p. 48). The *negative heuristic* forbids directing the *modus tollens* at the hard core; the *positive heuristic* guides modification of the protective belt.

Mechanism claims have exactly this structure. The hard core—the most entrenched conjuncts—consists of the basic empirical observations ("these neurons are activationally correlated with factual recall," "Fourier features develop during training"). The protective belt consists of the interpretive and cross-system claims ("these neurons *store* facts," "the Fourier circuit *explains* grokking's timing"). Rational contraction should follow the negative heuristic: preserve the hard core and revise the protective belt.

We formalize this insight as an algorithm:

Algorithm 1 Levi Contraction for Mechanism Claims

Require: Mechanism claim $C = c_1 \wedge c_2 \wedge \cdots \wedge c_n$ with entrenchment ordering $c_{\pi(1)} \preceq c_{\pi(2)} \preceq \cdots \preceq c_{\pi(n)}$; anomalous evidence A.

Ensure: Contracted claim C' consistent with A.

1: Identify the set $\mathcal{S}$ of all minimal subsets $S \subseteq \{c_1, \ldots, c_n\}$ such that A is logically consistent with $C \setminus S$.
2: **if** $\mathcal{S} = \emptyset$ **then**
3: View revision required. **Return** $\bot$.
4: **end if**
5: Select $S^* = \arg\min_{S \in \mathcal{S}} \sum_{c_i \in S} \mathrm{ent}(c_i)$.
6: **return** $C' = C \setminus S^*$.

The entrenchment ordering maps Lakatos's distinction: hard-core sub-claims have high entrenchment (they are supported by multiple independent lines of evidence and presupposed by many other accepted claims), while protective-belt sub-claims have low entrenchment (they rest on fewer lines of evidence and are presupposed by fewer other claims). The algorithm's minimization criterion formalizes the negative heuristic: give up the least entrenched (protective belt) conjuncts first.

Remark 4.5 (Idealization)**.** *Algorithm* **??** *models mechanism claims as propositional conjunctions and consistency as deductive entailment. In practice, mechanism claims are empirical and their consistency with evidence is a matter of degree. The algorithm captures the* structure *of rational contraction—give up the least entrenched conjuncts first—rather than providing a decision procedure that can be applied mechanically. The entrenchment ordering itself requires scientific judgment.*

Remark 4.6 (Progressive vs. Degenerating Programmes)**.** *Lakatos classifies research programmes as progressive or degenerating. A programme is progressive if its protective-belt modifications predict novel facts; it is degenerating if modifications are "fabricated only in order to accommodate known facts" (**?**, pp. 5–6). The contraction algorithm provides a diagnostic: if successive contractions of a mechanism claim preserve the hard core while yielding contracted claims that make novel predictions, the programme is progressive. If the protective belt keeps being patched with ad hoc modifications that accommodate anomalies without predicting anything new, the programme is degenerating, and the mechanism term is a candidate for zombie status.*

The algorithm has a crucial edge case: when $\mathcal{S} = \emptyset$, meaning the anomalous evidence is inconsistent with every partial conjunction of the original claim. In this case, no contraction suffices; what is needed is revision of the view under which the claim was made. This is the Lakatosian moment where the hard core itself is threatened—where Lakatos says the researcher should consider whether the programme has degenerated beyond recovery.

4.5 The Inferential Reach Lattice

The intuition behind inferential reach is simple: stronger claims require stronger evidence, and no amount of weaker evidence can substitute. Ten behavioral experiments do not add up to one structural experiment.

We now state the main theoretical result, which formalizes this intuition. The full proof is in Appendix **??**.

Definition 4.7 (Inferential Reach)**.** *Let σ_i be a mechanistic view with evidence type E_i and licensed inference type T_i. The* inferential reach *of evidence of type E_i is the set of inferences that such evidence can license:* $\text{Reach}(E_i) = T_i$.

Theorem 4.8 (Strict Ordering of Inferential Reach)**.** *Let σ_i and σ_j be mechanistic views with $\sigma_i < \sigma_j$ in the commitment hierarchy. Let C be a mechanism claim supported only by evidence admissible under σ_i. Then C does not license σ_j-characteristic inferences, regardless of the strength of the evidence.*

Proof sketch. Suppose C, supported only by E_i-type evidence, licenses a σ_j-characteristic inference $\tau \in T_j \setminus T_i$. By view coherence, licensing τ requires verifying the identity criterion $\sim_j$. But E_i can verify at most $\sim_i$-distinctions. Since $\sigma_j > \sigma_i$, the higher view imposes strictly more conditions on cross-system identity: $\sim_j$-equivalence requires all of $\sim_i$-equivalence plus additional conditions. Therefore E_i cannot verify the additional conditions, and C as supported by E_i is consistent with multiple $\sim_j$-equivalence classes. Hence C does not license τ. □

The theorem has a simple consequence: evidence cannot be "promoted" up the commitment hierarchy by accumulation. No quantity of lower-level evidence can substitute for higher-level evidence.

4.6 The Scaffold-Path Decomposition

The transport hierarchy raises a practical question: what kinds of evidence bear on each level? We argue that two fundamental modes of evidence—weight analysis and activation analysis—address formally distinct aspects of mechanism structure, and that their complementarity is not merely convenient but provably necessary.

The distinction maps onto what **?** calls the difference between a *capacity* and its *exercise*. Cartwright argues that "the generic causal claims of science are not reports of regularities but rather ascriptions of capacities, capacities to make things happen, case by case" (p. 3). A capacity is "a relatively enduring and stable" property "that [entities] carry with them from situation to situation" (pp. 2–3). The exercise of a capacity depends on context: enabling conditions, interfering factors, and the arrangement of other components.

Weight-space analysis reveals what **?** would call a mechanism's *capacity*: the set of computations that the model's weights can in principle support, independent of any particular input. Activation-space analysis reveals the *exercise* of that capacity: which weight-space capacities are actually recruited on the inputs observed.

These are formally complementary. Weight analysis is gauge-dependent: the same input-output function can be realized by many weight configurations related by gauge symmetry, so it reveals capacity only up to gauge equivalence. Activation analysis is input-dependent: the same weights produce different activations on different inputs, so it reveals exercise only for the inputs observed. Neither modality alone can determine both what a mechanism can do and what it does do.

In a neural network with gauge symmetries, a single measurement modality cannot in general determine both capacity and exercise simultaneously. Triangulation is therefore necessary for mechanism claims at the Structural View level or above.

This result refines **?**'s account of robustness reasoning: the value of convergent evidence lies not merely in reliability through redundancy, but in the fact that weight and activation evidence access genuinely independent aspects of the mechanism. It also responds to **?**'s challenge. Stegenga argues that the evidentiary force of convergent findings depends on assumptions about independence that are typically left unexamined. He is right. The scaffold-path decomposition makes the independence claim explicit and verifiable: weight evidence and activation evidence are independent not by stipulation but by construction.

The practical consequence is a concrete prohibition: a mechanistic claim at the Structural View level or above cannot be established by activation-patching evidence alone, regardless of how much of it one accumulates. This follows directly from the inferential reach theorem and the formal independence of weight-space and activation-space evidence. The prohibition is immediately actionable: any paper claiming structural-level mechanism identity on the basis of activation evidence alone is committing an evidence misfire, diagnosable from the methods section without examining the results.

5 Worked Examples

We apply the framework to concrete cases, demonstrating how each reference failure mode triggers a specific contraction pattern. Each example is structured as a contraction case study: we state the original claim as a conjunction, identify the anomalous evidence, and apply Algorithm **??** to produce the contracted claim.

5.1 Zombie Mechanisms

Knowledge Neurons. **?** introduced the term "knowledge neurons." **?** and **?** demonstrated through ROME and MEMIT that factual associations are distributed. Contraction: the original claim was $c_1 \wedge c_2 \wedge c_3$ (correlated, causal, localized storage). Rational contraction preserves c_1 and a weakened c_2 while discarding c_3 and the term itself. The contracted claim: "certain MLP neurons are part of the distributed circuitry involved in factual recall."

Fourier Features as the Grokking Mechanism. **?** identified Fourier features as progress measures for grokking. **?** demonstrated that the Fourier circuit reaches completeness ∼3,000 steps *before* grokking, with the timing controlled by weight decay. We work through this contraction in detail in Appendix **??**.

SAE Features as Canonical Atoms. **?** demonstrate that SAE features are incomplete and non-atomic. The contracted claim: SAE features provide a useful dictionary, but it is analyst-relative rather than model-intrinsic.

The Gender Bias Circuit. Multiple groups proposed "the gender bias circuit" in GPT-2-family models. Subsequent analysis revealed that implicit and explicit gender processing recruit substantially different circuits. The referent is distribution-specific and does not transport even across prompt types within the same model.

FOXP2 as "The Language Gene." FOXP2 was identified as "the language gene" (**?**), but it is a transcription factor regulating hundreds of downstream genes. The contracted claim: "FOXP2 is a transcription factor whose mutation disrupts neural development in ways that impair language."

5.2 Mimic Mechanisms

Unconstrained Distributed Alignment Search. **?** provide the cleanest example: DAS achieves perfect IIA on random models. The referential content of a method resides in its constraints, not its scores.

Probes Without Causal Validation. Linear probes (**?**) can achieve 95% accuracy on features the model does not use (**??**). This involves two compounding failures: the *mimicry* (condition (ii) violated—the feature is an artifact of the probe's capacity) and the *laundering* (Role View evidence used for Subspace or Structural claims).

5.3 Claim Laundering

Activation Patching and Necessity. **?** found that approximately half of 50 surveyed papers made claims exceeding their evidence level. In the MechVal hierarchy (**?**), activation patching provides Level 2 evidence (sufficiency); necessity requires Level 3 (ablation plus absence of compensation). Theorem **??** applies directly.

Induction Heads and In-Context Learning. **?** establish induction heads in one- and two-layer transformers via structural and causal evidence: these heads attend to the token following a previous occurrence of the current token, and ablating them degrades in-context learning. This is well-validated Role View evidence in small models. The laundering occurs in the export: **?** propose that this same mechanism underlies in-context learning in models with hundreds of billions of parameters. That inference—from Role View evidence in simplified models to a functional claim about large-model behavior—requires Process View or Structural View evidence that does not exist. The "selective induction heads" result (**?**) is the anomalous evidence that makes this laundering detectable: what large models do is substantially more complex than what two-layer models do. The original small-model claim is valid; the laundering is in the cross-scale export. Contraction: preserve c_1 (induction heads exist and are causal in small transformers) and c_2 (the attention pattern persists across scales); discard c_3 (induction heads explain in-context learning in large models).

Mirror Neurons and Human Empathy. **?** discovered mirror neurons in macaque premotor cortex. The laundering occurred when single-cell recording evidence was transported to humans via fMRI and used for empathy claims (**??**). The contracted claim—"macaque premotor cortex contains neurons responding to both execution and observation"—is valuable; the discarded claim—"mirror neurons are the neural basis of human empathy"—is a zombie produced by laundering.

Choice Subspaces Across Brain Regions. "Choice-selective subspaces" are applied across brain regions, tasks, and species. The term is applied on the basis of functional similarity (Role View) and used to support geometric claims (Subspace View). Optogenetic interventions add causal evidence but do not establish subspace identity.

5.4 Reference Debt

The Circuit Universality Hypothesis. **?** conjectured universal computational motifs. The hypothesis was ambiguous between Role View and Structural View readings. **?** partially called in this debt, and **?** quantified the gap: within-family interpretive congruence reaches 0.89–0.99 (Role View transport holds), but cross-family congruence drops to 0.13 (Structural View transport fails). The debt is partially serviced at the Role level and defaulted at the Structural level.

Induction Heads Across Scales. **?** defined induction heads in small transformers. **?** showed that large-model behavior is better described as "selective induction." **?** provide partial evidence for Role View transport—the same algorithmic roles persist across training and scale—but the mechanism's internal structure becomes substantially more complex. Role View transport may hold; Structural View transport fails.

Cross-Species Drug Mechanism Transport. **?** report that ∼90% of drugs entering clinical trials fail. Each failure is a potential reference debt default.

Pluripotency Network Across Cell Types. The Oct4/Sox2/Nanog network requires Structural View transport (regulatory network conservation), not merely Role View transport (same transcription factors present). **?** provides a route but it has not been established for most applications.

5.5 A Positive Case: Weight Circuit Transport

? demonstrate weight-space circuit convergence exceeding $r = 0.98$ across models trained from different random initializations, including zero-shot transfer from GPT-2 to Qwen. We state the original claim as a conjunction: c_1: weight-space circuit structure exists (structural observation); c_2: structure is stable across random initializations (within-architecture transport); c_3: structure transports to architecturally distinct models (cross-architecture transport); c_4: weight-space convergence predicts activation-level circuit behavior (scaffold-path bridge).

We walk through Definition **??** step by step. Condition (i) requires stable within-system reference: the weight-space circuit structure converges to $r > 0.98$ within each system, exceeding any reasonable stability threshold. The structure is not an artifact of a particular random seed; it is a reproducible feature of the trained model. Condition (ii) requires an admissible transport map: **?** construct a zero-shot transfer from GPT-2 to Qwen that preserves circuit structure without any model-specific tuning. The transport map exists and is explicit. Condition (iii) requires preservation of identity-relevant predicates: the same computational roles—name movers, duplicate token detectors, S-inhibition heads—are identified in structurally corresponding positions across architectures. The predicates that define the circuit's functional identity are preserved under transport.

This is the only worked example in this paper where all conditions of Definition **??** are met. The other fourteen examples demonstrate various failure modes; this one demonstrates that the framework's conditions are achievable, not merely diagnostic. The conditions are stringent—most mechanism terms in current use do not satisfy them—but they are not vacuously strong.

The positive case is important because it shows the framework has constructive content. Reference is not merely a standard that everything fails; it is a standard that well-designed methods can meet.

5.6 Summary of Worked Examples

Table 2: Worked examples mapped to reference failure modes.

Example	**Failure mode**	**Condition violated**
Knowledge neurons	Zombie	Def. **??**(i): referent dissolved
Fourier/grokking	Zombie	Def. **??**(i): partial referent
SAE canonical atoms	Zombie	Def. **??**(i): referent dissolved
Gender bias circuit	Zombie	Def. **??**(i): no unified referent
FOXP2 "language gene"	Zombie	Def. **??**(i): referent dissolved
Unconstrained DAS	Mimic	Def. **??**(ii): unstable
Probes w/o intervention	Mimic + Laundering	Def. **??**(ii) + Thm. **??**
Activation patching	Laundering	Thm. **??**: L2 $\not\to$ L3
Mirror neurons	Laundering	Thm. **??**: Role $\not\to$ Structural
Choice subspaces	Laundering	Thm. **??**: Role $\not\to$ Subspace
Circuit universality	Reference debt	Def. **??**(ii): unvalidated
Induction heads	Reference debt	Def. **??**(ii): partially defaulted
Drug mechanism transport	Reference debt	Def. **??**(ii): 90% default rate
Pluripotency network	Reference debt	Def. **??**(ii): unvalidated
Weight circuit transport	*Positive*	All conditions satisfied

6 Discussion

6.1 Pluralism Without Relativism

The framework is pluralist in that it accommodates multiple mechanistic views, consistent with **?**'s account of pluralism. But it is not relativist: given a declared view σ, reference succeeds or fails as constrained by the world. This navigates between naive realism (mechanism terms either refer or they do not, independent of the view) and pure instrumentalism (mechanism terms are merely useful fictions). The difference between constrained DAS (which finds genuine mechanisms) and unconstrained DAS (which does not) is precisely the difference between a method with referential content and one without.

6.2 Consequences for AI Safety

The reference problem has direct consequences for AI safety. If mechanism terms are used to construct safety assurance arguments—claiming that a model will not exhibit harmful behavior because its internal mechanisms have been characterized—then unexamined reference conditions create blind spots. A safety argument that relies on "the deception circuit" transporting from GPT-2 to a frontier model is only as strong as the transport conditions for the operative view. If the term transports at the Role level but not the Structural level, the safety argument covers the functional role but not the implementation—and it is the implementation that determines whether the model can circumvent the characterized mechanism via an alternative pathway. The framework makes these blind spots diagnosable before they become failures.

6.3 The Replication Crisis as a Reference Crisis

Some replication failures are better understood as reference failures: the mechanism term does not refer across the system boundary, so both the original finding and the replication are correct—each describes the mechanism in its own system—but the assumption that they are about the same mechanism is false. The most productive response is sometimes not "do the study again with more subjects" but "check whether the mechanism term refers in the new context."

The empirical scale of the replication crisis provides the baseline. **?** attempted to replicate 100 published psychology studies and found that only 36% of replications produced significant results, compared to 97% of the originals. This dramatic gap has been attributed to statistical power, publication bias, p-hacking, and questionable research practices. **?** locates the problem primarily in statistics: low prior probabilities, small effect sizes, flexible designs, and conflicts of interest combine to make most published findings false. We do not dispute these accounts. We locate a distinct class of failure in mechanism reference. **?** explains why a statistically significant finding may not replicate; we explain why a methodologically sound finding may not generalize. These are not competing accounts—they address different failure modes.

? distinguish three concepts that are often conflated: *replicability* (same result with new data), *robustness* (same result with different analytic choices), and *reproducibility* (same result from same data and analysis). Reference failure is, in their taxonomy, a failure of *generalizability*: the mechanism term does not refer in the new context, so the finding cannot be expected to replicate even if the original study and the replication are both methodologically sound. The original study is reproducible and may be robust; what fails is the implicit claim that the mechanism term denotes the same object in both contexts.

? makes a complementary argument: psychological claims are routinely stated at a level of abstraction that far exceeds what the experimental operationalization licenses. A study tests whether "ego depletion" reduces "self-control" as operationalized by a specific handgrip task in a specific population, but the claim is stated as a general law about ego depletion and self-control. This is structurally identical to claim laundering: evidence admissible under a narrow operationalization (analogous to the Role View) is used to support claims at a level of abstraction (analogous to the Structural or Process View) that the evidence cannot reach.

We do not claim that reference failure accounts for all replication failures—statistical power, publication bias, and p-hacking are real and well-documented. We claim that a distinct class of replication failures—those

where both the original and replication are methodologically sound but the mechanism term does not refer across the system boundary—is better diagnosed as reference failure than as statistical error.

6.4 Against Naive Realism: Van Fraassen, Stegenga, and the Limits of Empirical Adequacy

The framework challenges the implicit assumption that mechanism terms directly denote mind-independent objects the way "the electron" does. Mechanism reference is view-dependent: what "the induction head" refers to depends on the operative view. This does not mean that induction heads are not real; it means that their reality is mediated by the identity criterion that determines what counts as "the same" induction head. This position has affinities with **?**'s account of natural kinds as "self-vindicating."

?'s constructive empiricism poses a deeper challenge. Van Fraassen holds that "science aims to give us theories which are empirically adequate; and acceptance of a theory involves as belief only that it is empirically adequate" (p. 12). A theory is empirically adequate when what it says about observables is true, without any commitment to the reality of unobservables. If mechanism terms refer to unobservable internal structures of neural networks, and if constructive empiricism is correct, then the question of whether "the IOI circuit" refers is ill-posed—there is nothing for it to refer *to*, beyond the observable behavioral predictions it generates.

We reject this move. The mimic mechanism problem demonstrates why. Unconstrained DAS is empirically adequate in van Fraassen's sense: it saves the phenomena (100% IIA). But it has zero referential content. If constructive empiricism were correct, there would be no principled difference between constrained and unconstrained DAS—both are empirically adequate, and that is all science should aim for. But there *is* a principled difference: constrained DAS finds real features of the model; unconstrained DAS does not. The referential content is what separates mechanism science from mere prediction.

? raises a complementary challenge: the evidentiary force of convergent findings depends on assumptions about independence that are typically unexamined. We agree. The scaffold-path decomposition (Section **??**) provides a constructive response for weight-space vs. activation-space evidence, which are independent by construction. But Stegenga's challenge applies in full force to cases where multiple activation-based methods converge: such convergence may reflect shared blind spots rather than independent access to the phenomenon.

6.5 Epistemic Iteration and Reference Stabilization

? argues that scientific knowledge develops through *epistemic iteration*: "a process in which successive stages of knowledge, each building on the preceding one, are created in order to enhance the achievement of certain epistemic goals." The key insight is that inquiry does not begin from a secure foundation. Scientists start from inherited, imperfect practices and use the results of inquiry to refine the very starting point they relied on. Chang argues this circularity is progressive, not vicious, because self-correction is built into the procedure.

The contraction algorithm formalizes one form of epistemic iteration for mechanism claims. The scientist starts with a claim, encounters anomalous evidence, contracts to a weaker claim, and the contracted claim becomes the starting point for the next round of investigation. Each contraction improves the term's referential precision—not by converging on a fixed truth, but by iteratively pruning the over-commitments while preserving the empirical core. The process by which "knowledge neurons" was contracted to "MLP neurons involved in distributed factual recall" is a textbook case of Changian epistemic iteration: the refined term is more precise, more stable, and better supported, even though it was reached through a process of loss rather than discovery.

This framing also clarifies the relationship between mechanism *discovery* and mechanism *knowledge*. Discovery is the first iteration: establishing that a term refers within a system. Knowledge is the stabilization that results from multiple iterations of cross-system reference testing, contraction under anomalous evidence, and refinement. A term that has survived several rounds of contraction without losing its hard core has earned a kind of referential stability that a freshly coined term has not.

6.6 Forward Predictions

The framework generates testable predictions about mechanism terms currently in circulation.

Prediction 1: IOI circuit cross-architecture (partially confirmed). We predicted that "the IOI circuit" applied across architectures would default on Structural View transport. This prediction is now partially confirmed: **?** show that cross-family congruence drops to 0.13, while **?** demonstrate that even within a single model, prompt-specific circuits are distinct. Role View transport holds—**?** confirm that the same functional roles persist across scale—but the structural referent does not transport. The remaining prediction is that the field will not converge on view-explicit claims within two years, and reference debt will continue to accumulate.

Prediction 2: SAE feature universality. Claims that SAE features are "universal units of computation" are in a degenerating research programme in Lakatos's sense. Each new anomaly (**?**'s incompleteness and non-atomicity results; basis-dependence across dictionary sizes) is accommodated by protective-belt modifications ("features are universal at the right dictionary size," "features are universal up to linear transformation") that do not predict novel facts. The framework predicts that the term "canonical SAE feature" will zombie: the term will persist in the literature but its referent will have been dissolved.

Prediction 3: Cross-species mTOR transport. The 90% clinical trial failure rate suggests that pharmacological mechanism transport is failing systematically. The framework predicts that studies explicitly specifying the view under which cross-species transport is claimed, and validating the transport conditions for that view, will have substantially higher success rates than studies that assume transport from behavioral similarity alone. This is a testable prediction about study design, not about any particular drug.

6.7 Practical Implications for Interpretability Papers

An interpretability paper following this framework would make three additions to its methods section, none requiring new experiments: (1) declare the mechanistic view under which claims are made, specifying the identity criterion that determines what counts as "the same mechanism" across systems; (2) when transporting a mechanism term from a prior study, state which transport conditions have been verified and which are assumed; (3) when evidence reveals that a mechanism term does not refer as expected, apply rational contraction rather than abandoning the entire claim—give up the least entrenched conjuncts first.

These are not onerous requirements. The first is a sentence in the methods section. The second is a paragraph in the discussion. The third is a revision practice, not a reporting requirement. The cost of adopting them is minimal; the cost of not adopting them is the continued accumulation of mechanism terms with unexamined reference conditions.

6.8 Limitations

The framework has several limitations. First, it presupposes that mechanistic views can be made explicit. Second, the admissible transport maps may be computationally difficult to construct (**?**). Third, the framework treats views as fixed during analysis; a full theory of view revision is beyond the scope of this paper. Fourth, the five reference failure modes may not be exhaustive. Fifth, the forward predictions in Section **??** are derived from the framework's analysis of current reference debt, not from empirical study of reference failure rates; they could be wrong.

7 Conclusion

We have argued that cross-system mechanistic reference is not a free lunch. When a mechanism term, validated in one system, is applied to another, the application presupposes a transport map whose existence is not guaranteed and whose failure is detectable. We have defined five modes of reference failure, shown that standard belief revision frameworks (Bayesian updating, AGM contraction) are structurally inadequate for mechanism claims, provided a Levi-rational contraction algorithm grounded in Lakatos's distinction between hard core and protective belt, proved that inferential reach is strictly ordered, and shown that weight-space and activation-space evidence address formally complementary aspects of mechanism structure.

Mechanism discovery is the first iteration: the establishment that a term refers within a system. Mechanism knowledge is the result of multiple iterations of cross-system reference testing, contraction under anomalous evidence, and refinement. A term that has survived several rounds of contraction without losing its hard core has earned a kind of referential stability that a freshly coined term has not. The process by which "knowledge neurons" contracted to "MLP neurons involved in distributed factual recall" is a textbook case of Changian epistemic iteration (**?**): the refined term is more precise, more stable, and better supported, even though it was reached through a process of loss rather than discovery.

The present paper complements three companion works. **?** provides the view framework that generates the identity criteria this paper's transport conditions presuppose. **?** provides the evidence standards against which individual claims are evaluated. **?** addresses the horizontal composition problem—when individually valid claims compose into system-level understanding. Together, these four papers provide per-view, per-claim, per-term, and per-cluster evaluation frameworks for mechanistic science.

The central message is this: *what travels is what refers.*

A Full Proof of Theorem ??

Proof. Let $\sigma_i = (O, \sim_i, E_i, F_i, T_i)$ and $\sigma_j = (O, \sim_j, E_j, F_j, T_j)$ with $\sigma_i < \sigma_j$.

Step 1. By construction, $T_j \not\subseteq T_i$: there exist σ_j-characteristic inferences $\tau \in T_j \setminus T_i$.

Step 2. By view coherence, E_j distinguishes $\sim_j$-equivalence classes. A σ_j-characteristic inference τ depends on distinctions that $\sim_j$ draws but $\sim_i$ does not.

Step 3. The commitment hierarchy is cumulative: if $m \sim_j m'$ then $m \sim_i m'$, but not conversely. There exist $\sim_i$-equivalent objects that $\sim_j$ treats as distinct.

Step 4. E_i distinguishes $\sim_i$-classes but cannot distinguish $\sim_i$-equivalent objects. Since some $\sim_j$-distinct objects are $\sim_i$-equivalent, E_i cannot tell them apart.

Step 5. C supported by E_i is consistent with multiple $\sim_j$-classes, so it does not license τ.

This holds regardless of quantity: more evidence of the same type cannot resolve distinctions the type is constitutively unable to detect. □

B Worked Contraction: Fourier Features and Grokking

We apply Algorithm **??** to the Fourier/grokking case.

Original claim. $C = c_1 \wedge c_2 \wedge c_3 \wedge c_4$ where c_1: Fourier features develop during training; c_2: their development correlates with grokking; c_3: Fourier development causally drives grokking; c_4: the Fourier circuit is sufficient to explain grokking's timing.

Entrenchment. $c_4 \prec c_3 \prec c_2 \prec c_1$. c_1 is most entrenched (direct observation, multiply confirmed). c_4 is least entrenched (strongest inference, single line of support).

Anomaly. A: The Fourier circuit reaches completeness ∼3,000 steps before grokking. Timing is controlled by weight decay (**?**).

Step 1. A is inconsistent with c_4 directly. Minimal subsets: $\mathcal{S} = \{\{c_4\}, \{c_3, c_4\}\}$.

Step 2. $\mathcal{S} \neq \emptyset$; view revision not required.

Step 3. $S^* = \{c_4\}$ (minimizes entrenchment).

Step 4. $C' = c_1 \wedge c_2 \wedge c_3$, with c_3 interpreted weakly.

Lakatos diagnostic. This is a *progressive* contraction: c_1 and c_2 (hard core) are preserved, c_4 (protective belt) is discarded, and the contracted claim generates a novel prediction—that the timing mechanism is separate from the computational mechanism, which can be tested by manipulating weight decay independently.