A Higher-Categorical Semantics of Weight-Space Misfolding

Anonymous (AI-assisted)

13 November 2026

Abstract. We formalize the distinction between an edit to model weights, an adverse behavioral change, and an observed deployment of the edited artifact. An internal complete Segal object records directed release histories in an infinity-topos. Behavioral contracts are dependent predicates on scenarios. We prove that a test family certifies every contract exactly when it is an effective epimorphism over the declared scenario object. Proof-relevant attestations glue by Čech descent. We then characterize the largest deployment gate that preserves safety under all unavoidable transitions. Its domain is a greatest invariant subobject and its construction commutes with change of deployment context. We relate these results to the reported factual substitution, over-refusal, and selective omission cases in *Weight Space Misfolding*. A finite release model and exact tests of the formulas illustrate the scope of the guarantees.

Keywords. Infinity-topos, weight-space editing, behavioral contracts, effective epimorphism, provenance, supervisory control.

1 Introduction

Von Csefalvay, Doma, Unni Krishnan, and Foldi use *weight space misfolding* for unauthorized changes in a model's behavior brought about through weight editing [1]. Their study distinguishes direct interference from hypothetical onward propagation. Its selected cases concern factual substitution, refusal of benign assistance, and omission from a supplied record. Each case involves a retained checkpoint with separate measurements of target behavior and control performance. The measurements establish changes under the reported conditions. End-to-end propagation remains a hypothesis in the source.

An editor may have permission to create an artifact even when its behavior falls outside the release contract. A registered release may also be eligible for deployment while an individual inference runs under a different checkpoint or runtime configuration. These are distinct statements about a history. We give each a type, identify the quantifiers that a certification claim uses, and state the conditions under which local evidence yields a global claim.

The first result concerns testing. For any test map we compute the predicate that its passing results certify. It is the universal contract restricted to the image of that map. A covering family gives the intended universal statement, and adequacy for *every* contract implies that the family covers. The same cover gives an exact criterion for gluing proof-relevant evidence. The second result concerns deployment. Even a contract checked at the immediate destination of a release can miss a failure forced by a later transition. We construct the largest gate that is safe from every viable state, accounting for all transitions outside the gate's control. The classical full-information supervisory-control result is recovered when the topos is interpreted on discrete sets.

The starting point in [2, 3] is direct weight-space programming. Cohen et al. study changes to related knowledge after an edit [4]. A preceding journal article formalizes parameterized behavior, symmetry, and certificates for weight-space programming [5]. Here the objects of study are directed release histories, the precise coverage of observations, and invariants of deployment. The categorical ingredients are standard. The contributions are their combined application and the stated criteria for a release policy.

2 Behavior and release histories

Work in a Grothendieck infinity-topos $\mathcal{X}$, with the objects below in a fixed universe. Products and pullbacks are homotopy products and homotopy pullbacks. For a morphism f, reindexing on slices has adjoints $\Sigma_f \dashv f^* \dashv \Pi_f$. On subobjects, the corresponding adjoints are $\exists_f \dashv f^* \dashv \forall_f$. Here $\exists_f$ takes the image of the composite and $\forall_f$ is the restriction of Π_f to (-1)-truncated objects. We write $\top$ for the entire object, $\wedge$ for intersection, and $\Rightarrow$ for Heyting implication [6, 7].

Let A be a 0-truncated object of interface labels. Let $X \to A$ be a family of scenarios and $Y \to X$ a family of outcomes. A scenario specifies the input together with the environment, execution limits, and the role of any randomness used in the claim. For example, a deterministic semantics of sampling puts a seed into X. Equality at each seed then means equality under that coupling. A law-level claim needs a specified space of laws and an evaluator into that space.

Let $P \to A$ be a state object. A state includes the actual checkpoint and the metadata required by the stated release policy. The execution semantics is a map over X,

$$\mathrm{ev}\colon P \times_A X \longrightarrow Y. \tag{1}$$

The behavioral denotation is the adjunct $P \to \Pi_{X \to A} Y$ in the slice over A. This denotation can identify states with different artifact identities. An attestation for one artifact transfers to another only with the transport justified by its own evidence type. For example, a hidden unit

permutation can preserve the evaluator while changing the serialized artifact named in a release record.

Let $\mathcal{E}$ be an internal complete Segal object in $\mathcal{X}$ over the constant simplicial object on A, with $\mathcal{E}_0 = P$ and $\mathcal{E}_1 = E$. Write $s, t\colon E \to P$ for source and target. Its Segal maps give

$$\mathcal{E}_n \simeq \underbrace{E \times_P \cdots \times_P E}_{n \text{ factors}} \qquad (n \geq 2).$$

An arrow records an edit or release history. Its endpoint type alone records neither the event history nor the evidence carried by that history. Identities record empty histories. Composition concatenates histories up to the coherence encoded by $\mathcal{E}_n$. An arrow can have a later return to its starting checkpoint while both recorded events remain in the history. The free category on a finite directed graph, embedded as a discrete internal nerve, is a special case [8].

2.1 Contracts and adverse behavior

Set $J = E \times_A X$ and let $\pi\colon J \to E$ be the projection. There are two evaluation maps on J, $\mathrm{ev}_s(e, x) = \mathrm{ev}(s(e), x)$ and $\mathrm{ev}_t(e, x) = \mathrm{ev}(t(e), x)$. A contract $C \hookrightarrow J$ is a predicate on this pair and its declared context. It may also constrain state metadata. The accepted predicate over edges is

$$\mathsf{Cert}(C) = \forall_\pi C \ \hookrightarrow E. \tag{2}$$

For a state specification $\mathsf{Pass} \hookrightarrow Y$ over X, the two pullbacks $\mathsf{Pass}_s, \mathsf{Pass}_t$ to J record whether the source and target pass on that scenario. Two useful edge predicates are

$$\mathsf{Bad} = \exists_\pi(\neg\mathsf{Pass}_t), \qquad \mathsf{Loss} = \exists_\pi(\mathsf{Pass}_s \wedge \neg\mathsf{Pass}_t). \tag{3}$$

The first identifies violation at the destination. The second requires a previously satisfied obligation. With nondecidable specifications, one can supply a positive failure predicate and use it in place of $\neg\mathsf{Pass}_t$. In general, $\neg\exists_\pi\neg C = \forall_\pi\neg\neg C$. Thus inferring $\mathsf{Cert}(C)$ from the absence of a counterexample requires stability of C under double negation, which holds for decidable contracts.

For a family of scenarios J_e, the witness type of loss is $\sum_{x:J_e} \big(\mathsf{Pass}_s(e, x) \times \neg\mathsf{Pass}_t(e, x)\big)$. Equation (3) takes its propositional truncation. In sheaf semantics this truth can have locally specified witnesses without a globally chosen scenario. An auditable report can keep the witness and its context in the evidence type.

Write $\mathsf{WeightEdit} \hookrightarrow E$ for edges with a recorded weight intervention under matched nonweight execution settings. This is a provenance condition on the edit, in addition to the observed behavioral comparison. Let Auth mean permission to perform the recorded operation and Scope mean permission for the resulting behavioral change under the declared policy. One precise interpretation of unauthorized regression by weight editing is

$$\mathsf{Mis} = \mathsf{WeightEdit} \wedge \mathsf{Loss} \wedge \neg\mathsf{Scope}. \tag{4}$$

For example, an operator can authorize a maintenance edit while forbidding regressions on a protected contract. An edit that causes such a regression satisfies Loss even if Auth holds. For deployments, introduce $\mathsf{Lineage}, \mathsf{Observed} \hookrightarrow E$. The first expresses valid artifact lineage under the trust assumptions. The second expresses evidence that the target was used under the declared execution context. The deployed case of (4) is $\mathsf{Mis} \wedge \mathsf{Observed}$.

If permissions are only partially known, their negations are refutations. An empty log entry on its own gives an unresolved authorization question. These predicates describe the change, policy, and evidence. Intent would require additional data.

2.2 When edge contracts compose

Proposition 2.1. *Suppose the contract is induced by a fiberwise reflexive, transitive subobject $R \hookrightarrow Y \times_X Y$. The edges satisfying $\forall_x R(\mathrm{ev}_s(e, x), \mathrm{ev}_t(e, x))$ form a wide internal complete Segal subcategory of $\mathcal{E}$.*

Proof. Take the full object of states and restrict the n-simplices to those whose spine edges satisfy the predicate. Restriction commutes with the Segal products. The simplicial maps that compose consecutive spine edges preserve the predicate by transitivity on the same scenario. Degeneracies preserve it by reflexivity. The restrictions are subobjects, so their lifts have contractible spaces of choices wherever they exist. Completeness of $\mathcal{E}$ identifies each equivalence with an identity up to a path in the equivalence object, so all equivalences lie in the restricted object. The inverse of any such edge also lies there. Completeness therefore restricts. □

For an output distance, the contract *distance at most one* can hold for each of $0 \to 1$ and $1 \to 2$ while the composite fails. The triangle inequality composes the two budgets by addition. A policy with a fixed final tolerance can instead compare each release with a fixed approved baseline. When the policy changes across releases, its version and the rule for migration belong in the state. Proposition 2.1 applies only after proving the required closure for the chosen contract.

3 When tests certify a contract

A test family is a map $u\colon T \to J$. It names which scenario of which edge is checked. It may include repeated scenarios, and it may have evidence attached to each test. The following theorem concerns truth of the *same* predicate C on those scenarios. Measurement or checker soundness is a separate assumption.

Theorem 3.1 (Exact coverage criterion)**.** *Let $U = \mathrm{im}(u) \hookrightarrow J$. For any $C \hookrightarrow J$,*

$$\forall_{\pi u}(u^* C) = \forall_\pi(U \Rightarrow C). \tag{5}$$

Consequently the test predicate equals $\mathsf{Cert}(C)$ *for every subobject* C *if and only if* u *is an effective epimorphism. The formula is stable under reindexing along* $E' \to E$.

Proof. For any $D \hookrightarrow J$, adjunction and the image formula for intersection give

$$\begin{aligned} D \le \forall_u u^* C &\iff u^* D \le u^* C \\ &\iff \exists_u u^* D \le C \\ &\iff D \wedge U \le C \\ &\iff D \le (U \Rightarrow C). \end{aligned}$$

Compose right adjoints to obtain (5). If $U = \top$, the equality follows. Conversely, evaluate the asserted equality for every contract at $C = U$. Then $\forall_\pi U = \top$. The adjunction with π^* implies $U = \top$, so u is an effective epimorphism. Images and dependent quantification obey pullback Beck–Chevalley in $\mathcal{X}$. The formula therefore holds after the asserted reindexing. □

The necessity is uniform in C. For one particular contract, domain knowledge may justify a smaller test object. The exact condition for a sound inference at every generalized edge is $\forall_\pi(U \Rightarrow C) \le \forall_\pi C$. For instance, suppose $r\colon J \to K$ is over E and a proved abstraction yields $C = r^*\bar{C}$. If ru is an effective epimorphism, passing the tests certifies $\bar{C}$ on every fiber of $K \to E$, hence C on every fiber of π. The proof uses $\forall_{\pi u} u^* C = \forall_{(K\to E)ru}(ru)^*\bar{C}$ and r^* on the resulting full-fiber statement. The abstraction proof is where the omitted scenarios enter.

For a positive failure predicate $F \hookrightarrow J$, the dual image calculation gives $\exists_{\pi u} u^* F = \exists_\pi(U \wedge F)$. A single observed failure with a sound witness suffices to establish this existential claim. Certification requires the universal predicate in (5). These quantifiers express the difference between exhibiting a regression and establishing preservation of the contract.

Example 3.2 (Coverage of a context)**.** Let the scenario set consist of two response-length settings $\{a, b\}$, let $T = \{a\}$, and let $C = \{a\}$. The tests pass while $\forall_x C(x)$ fails. Testing a different family of inputs at the same setting cannot close this gap. A proved invariance of C under that setting would permit a quotient that identifies a and b.

3.1 Descent for attestations

Behavioral truth is subterminal. Its witnesses can carry structured data such as an artifact reference, a report, and evidence of agreement between two measurement procedures. Let $Q \to J$ be an arbitrary object of evidence. Put $T^{[n]} = T \times_J \cdots \times_J T$ with $n{+}1$ factors. If Q is evidence for C, a soundness map $Q \to C$ over J turns a global section of Q into a global proof of C.

Theorem 3.3 (Čech descent for evidence)**.** *If* $u\colon T \to J$ *is an effective epimorphism, restriction induces an equivalence*

$$\mathrm{Map}_{/J}(J, Q) \;\simeq\; \lim_{[n]\in\Delta} \mathrm{Map}_{/J}(T^{[n]}, Q). \tag{6}$$

The equivalence holds after every base change, in particular after restriction to the scenario family of a generalized edge.

Proof. The augmentation of the Čech nerve of u exhibits J as its colimit in $\mathcal{X}$ and in the slice over J. Mapping out of that colimit gives the displayed limit. Effective epimorphisms and the augmented colimit are stable under pullback in an infinity-topos [6, Sec. 6.2.3]. □

For subterminal Q over J, the coherence conditions are automatic once sections are given. For general Q, independently chosen local evidence may disagree on overlaps. There is already an obstruction in spaces. Let H be a nontrivial discrete group and take $T = * \to BH = J$ to be the chosen basepoint. It is an effective epimorphism since BH is connected. Take $Q = * \to BH$ to be the same map. Over T, the fiber is H, with a local section for each element. The transition on $T \times_{BH} T \simeq H$ is the regular action. A coherent descent datum would specify a fixed point for that action. Equivalently, a global section would make the identity on $\pi_1(BH) \cong H$ factor through the trivial group. Thus local sections exist on the cover while none of those choices extends globally. The example models the issue when a local certificate carries an identifier that must be transported under an equivalence.

3.2 Statistical evidence

In the discrete interpretation, fix a candidate before validation and sample $n \ge 1$ independent scenarios from a declared deployment distribution. Assume the checker detects failure of the selected predicate on each sampled scenario. If the unknown failure probability is p, the probability of zero observed failures is $(1-p)^n$. For $0 < \alpha < 1$, inverting the binomial test gives the exact one-sided $1-\alpha$ upper endpoint after zero failures [9].

$$p_{\mathrm{upper}} = 1 - \alpha^{1/n}. \tag{7}$$

For $0 < \varepsilon < 1$, the endpoint is at most ε when $n \ge \lceil \log(\alpha)/\log(1-\varepsilon) \rceil$. For instance, 299 independent failure-free observations give an upper endpoint of about 0.997% at 95% confidence. This is a statement about the specified distribution and the fixed artifact. Testing candidates adaptively on reused validation data invalidates the fixed-candidate calculation unless the selection is accounted for. The support and the definition of failure determine the scope of the bound.

4 Invariant deployment

Work over the state object P. A labeled one-step transition has a source and target in P. Let M denote unavoidable transitions, and let L denote the optional transitions eligible for authorization, provenance, and review checks. The labels may map to E over their endpoints. A gate chooses $H \hookrightarrow L$. Write $G \hookrightarrow P$ for the invariant required of every reachable deployment state. G can express behavioral acceptability, a bound runtime identity, or obligations on the full release state. It must represent all hypotheses used to interpret a state as safe.

For example, let $\ell\colon L_0 \to E$ map candidate transitions to their release records. A policy can set

$$L = \ell^*(\mathsf{Auth} \wedge \mathsf{Lineage} \wedge \mathsf{Cert}(C)) \hookrightarrow L_0.$$

The edge contract C controls behavioral eligibility, and G controls state safety after admission. When review uses a test predicate as evidence for $\mathsf{Cert}(C)$, Theorem 3.1 states the obligation for that replacement. Attesting actual use belongs to $\mathsf{Observed}$ and may take place after admission.

The gate controls its own choices of H. Every transition in M is possible without that choice. The words *unavoidable* and *possible* refer to control, so the invariant quantifies over all such edges, including different outcomes from the same state. Modeling an operational outcome as optional gives the gate control of that outcome. If the gate only selects an action and its outcome can vary, the state can be expanded to record the committed action, followed by unavoidable outcome transitions.

4.1 Reachability and invariants

For any labeled transition D and $B \hookrightarrow P$, define

$$\mathrm{post}_D(B) = \exists_{t_D} s_D^* B, \tag{8}$$
$$\mathrm{All}_D(B) = \forall_{s_D} t_D^* B. \tag{9}$$

The first is the image of the transitions from B. The second consists of states whose every D-successor belongs to B. For $I \hookrightarrow P$ initial and $D = M \amalg H$, set

$$\mathrm{Reach}_D(I) = \bigvee_{n\geq 0} \mathrm{post}_D^n(I).$$

Equivalently, this is the image of the endpoint of the coproduct of all finite D-path objects starting in I. At length n use an n-fold pullback over P, and at length zero use the state itself. This retains histories until the image is taken.

Proposition 4.1. $\mathrm{Reach}_D(I)$ *is the least D-invariant subobject containing I. Every finite execution starting in I lies in G at each prefix if and only if there is $B \hookrightarrow P$ with*

$$I \leq B \leq G, \qquad \mathrm{post}_D(B) \leq B.$$

Proof. The operations in post_D preserve joins, so the displayed join contains I and is D-invariant. Every invariant containing I contains each iterate. All prefix endpoints occur as endpoints of finite paths, including the path of length zero. Thus safety is precisely $\mathrm{Reach}_D(I) \leq G$, and the stated equivalence follows. □

4.2 The largest safe gate

Define $W_0 = \top_P$ and $W_{n+1} = G \wedge \mathrm{All}_M(W_n)$. Set $W = \bigwedge_{n\geq 0} W_n$. A state belongs to W exactly when every finite M-only path from it has its endpoint in G. To see the indexing, $W_1 = G$, and the recursion gives membership in G at the endpoints of all paths of lengths at most n for W_{n+1}.

Theorem 4.2 (Maximal safe deployment gate)**.** *W is the greatest subobject of G invariant under M. In the slice over L, set*

$$H_{\max} = s_L^* W \ \wedge \ t_L^* W. \tag{10}$$

Then $M \amalg H_{\max}$ preserves W and is safe from every $I \leq W$. Any gate $H \leq L$ with sources in W that is safe from every state of W satisfies $H \leq H_{\max}$. Existence of a safe gate from I is equivalent to $I \leq W$.

Proof. The sequence W_n is descending. The map All_M preserves meets since pullback and the right adjoint $\forall_{s_M}$ do. Thus

$$\begin{aligned} G \wedge \mathrm{All}_M(W) &= \bigwedge_{n\geq 0} \big(G \wedge \mathrm{All}_M(W_n)\big) \\ &= \bigwedge_{n\geq 0} W_{n+1} = W. \end{aligned}$$

If $B \leq G$ is M-invariant, the adjunctions give $B \leq \mathrm{All}_M(B)$. Induction shows $B \leq W_n$ for all n. This proves greatestness. Equation (10) and the invariance of W give safety by Proposition 4.1.

Suppose H is safe from W and its sources lie in W. The reachable object from W is M-invariant and lies in G, so it lies in W. Each H-target belongs to that reachable object. Thus $H \leq t_L^* W$ as well as $H \leq s_L^* W$, proving maximality. For general I, safety implies the M-reachable object from I is an M-invariant subobject of G. Hence $I \leq W$, which is also sufficient by the constructed gate. □

Safety from every state of W means the internal inclusion $\mathrm{Reach}_{M\amalg H}(W) \leq G$. The condition concerns all generalized states. The theorem permits denying every optional transition. A progress guarantee would require further hypotheses about available transitions or scheduling. The finite-state, fully observed interpretation is the invariant form of classical supervisory control [10, 11]. A deployer with a coarser observation can implement the maximal gate on all eligible edges exactly when membership in it descends to the observation. For

an observation $q\colon L \to O$ this means finding $B \hookrightarrow O$ with $q^*B = H_{\max}$. The largest decision predicate with $q^*B \leq H_{\max}$ is $\forall_q H_{\max}$, by adjunction. Its pullback is the largest gate measurable by q whose allowed edges lie in $H_{\max}$. An observation whose fiber includes an edge outside $H_{\max}$ cannot be certified by this rule. If q is an effective epimorphism, a decision on all of L descends exactly when the two pullbacks of $H_{\max}$ to $L \times_O L$ agree. Indeed, $\forall_q H_{\max}$ pulled back to L tests membership of every edge in the same observation fiber, by Beck–Chevalley. Under the agreement condition this is equivalent to membership of the original edge, which also gives the converse by pulling back a descended predicate.

Corollary 4.3 (Change of deployment context)**.** *Suppose the state, predicates, and transitions are over a context object B. For $g\colon B' \to B$, formation of W, $H_{\max}$, and* $\mathrm{Reach}_D(I)$ *in $\mathcal{X}_{/B}$ commutes with g^*.*

Proof. Use Beck–Chevalley for the source and target projections in (8) and (9). Reindexing between the slices preserves meets and joins because it has both adjoints. Pulling back the iterates, their meets or joins, and (10) gives the formulas in the new context. □

4.3 A finite release example

Take seven states named `base`, `review`, `stage`, `queue`, `live`, `fallback`, and `invalid`. The invariant G excludes only `invalid`. The tables specify the transition system.

Unavoidable source	Unavoidable destination
`review`	`base`
`stage`	`live`
`live`	`base`
`queue`	`fallback`
`fallback`	`invalid`

Eligible source	Eligible destination	In $H_{\max}$
`base`	`review`	yes
`review`	`stage`	yes
`review`	`queue`	no

The first iteration removes `invalid`, the second removes `fallback`, and the third removes `queue`. The fourth leaves

$$W = \{\texttt{base}, \texttt{review}, \texttt{stage}, \texttt{live}\}.$$

From `base` the gate reaches all four states in W. A gate checking immediate membership in G would admit the step to `queue`, which has a two-step unavoidable continuation to `invalid`. The example isolates the difference between validating an artifact at admission and establishing safety of its deployment history.

5 Interpreting the case studies

The three reported cases instantiate different choices of outcome and contract. Their categories can overlap, since several requirements may apply to the same response. The table reports the source's held-out target counts and its retention of baseline-passing controls [1, Sec. 4 and Table 2].

Case	Target count	Controls retained
Factual substitution	100/100	48/49
Over-refusal	5/22 exact	6/11
Selective omission	1/22	11/11

Factual substitution. For a fixed task context, let Y_x contain the extracted answer and let $V_x \subseteq Y_x$ be the reference answer set. Passing is membership in V_x. The source reports 100 target draws covering 91 distinct prompt strings. The count measures frequency under those probes. The control retention statistic compares pass status. One of the 49 baseline-passing controls was lost.

Refusal of benign assistance. Use a scenario predicate specifying when assistance is legitimate and an outcome predicate for the required response. Exact refusal and refusal-prefix occurrence are different classifiers on Y_x. The preprint gives five exact refusals and eleven prefix matches, including those five, on a recovered terminal state. The held-out target tally and the control tally address different predicates in (3). Recovery of an artifact from a trace also calls for its own lineage evidence.

Selective omission. Let an admissible response on scenario x contain all members of a declared set F_x of required record entries, with a fixed extraction rule. Then loss occurs if the source meets this obligation and the destination fails it. In the preprint the omission tally and the displayed controls come from different response-length settings. They belong to different parts of X. Applying the control result at the target setting would require a justification of transport or a separate check there, as in Example 3.2.

These counts are descriptive. The cases were selected, the prompts have a particular construction, and the contracts are bounded by the reported evaluation conditions. Equation (7) needs an independent sampling design and a specified deployment law. The reported evidence addresses individual retained states under controlled conditions.

5.1 Evidence for propagation

One observed bad state supplies a scenario witness of Bad or Loss if the corresponding criterion holds. A multi-release attribution requires observed deployment states at successive positions, a trace linking their release events,

and a supported comparison attributing the downstream change to the upstream state under a stated causal model. A temporal order and repeated adverse behavior alone admit a shared cause or independent regressions.

Let Z be an object of observed finite traces. A subobject $Z_{\mathrm{pair}} \hookrightarrow Z$ selects traces with two ordered, observed deployments and regression evidence for each. Specify an attribution family $K \to Z_{\mathrm{pair}}$ whose fibers carry evidence for the modeled effect of the first deployment on the second. The propagation judgment is the image

$$\mathsf{Attr} = \mathrm{im}(K \longrightarrow Z_{\mathrm{pair}} \longrightarrow Z) \hookrightarrow Z. \tag{11}$$

This judgment keeps the causal assumption explicit. For a fixed trace, a discrete instance could require a paired comparison in a structural causal model in which the downstream regression disappears when the upstream state is set to its approved alternative, holding the model's background variables fixed [12]. Other designs require their own attribution family. The path existence calculus in Section 4 supplies reachability claims. Equation (11) imposes an additional evidence obligation.

Proposition 5.1 (Observational underdetermination)**.** *In the discrete instance, the complete joint distribution of two recorded binary regression indicators need not identify the counterfactual comparison just described.*

Proof. Let U be a fair exogenous bit and let A_1, A_2 be the two indicators in time order. Both models set $A_1 = U$. The first sets $A_2 = A_1$, and the second sets $A_2 = U$. Both observational distributions assign probability $1/2$ to each of $(0,0)$ and $(1,1)$. In the context $U = 1$, replacing the equation for A_1 by $A_1 = 0$ yields $A_2 = 0$ in the first model and $A_2 = 1$ in the second. The comparison therefore depends on the structural model, even with the same ordered observations. □

This example is about attribution from observations. The source's experiments provide single-edit observations [1, Secs. 3.3 and 4]. The image judgment in (11) additionally requires evidence of the counterfactual comparison or an alternative attribution witness.

6 Release evidence and scope

In a release policy, Lineage can express that a trusted build or review chain vouches for a particular artifact digest. Supply-chain attestation systems specify how to verify such claims [13, 14]. To support Observed, the inference or service record must bind the loaded artifact and execution configuration to the event being asserted. To support $\mathsf{Cert}(C)$, the evidence must address the behavior predicate. The proofs assume any evidence accepted by the operational verifier has the meaning assigned to its type.

A release eligibility judgment may bind permission and evidence to the base artifact, target artifact, principal, runtime, and policy version. Testing one artifact and then serving a different one breaks that binding. A revocation or a policy migration is a state transition. If an automatic fallback can change the running artifact after admission, it belongs among the transitions analyzed by All_M. Rollback to a previously identified checkpoint is itself a new transition. For safety properties that include effects on an external system, the state must also record the relevant effects and any restoration obligations. This interpretation is consistent with the provenance and mitigation categories in the NIST taxonomy [15].

The gate theorem assumes a complete model of transitions relative to the asserted invariant. It is sound for the operational system when every concrete execution simulates an allowed model execution and the modeled invariant implies the concrete policy. Formally, given a state map $r\colon P_{\mathrm{concrete}} \to P$, require it to send each initial state into I, each concrete transition to an allowed model path with matching endpoints, and r^*G into the concrete safe subobject. Induction on concrete paths then transfers safety. Quantifying over an overapproximation of possible transitions can preserve soundness while reducing the set of allowed releases.

The results are semantic. In general, evaluating a universal contract, determining an effective cover, or computing the greatest invariant need not be decidable. Finite state and decidable predicates give executable calculations. Arbitrary neural weights, stochastic decoders, and continuous physical environments require justified abstractions or qualified statistical claims. The finite model in this paper concerns release policy. The neural-network measurements come from the cited preprint.

7 Finite verification

The finite calculations use exact subsets and Boolean relations. For Theorem 3.1, enumerate all functions $T \to J$ with $|T|, |J| \le 3$, all projections from J to a two-element edge set, and all subobjects C of J. Check (5) on each edge fiber and compare universal agreement across all C with surjectivity. This gives 2,817 instances, including cases with empty test or scenario fibers. For Theorem 4.2, on a three-element state set enumerate all 2^9 mandatory relations and all 2^3 safe subsets. In each of the 4,096 systems, compare the descending iteration with independent graph reachability. With every optional edge eligible, enumerate every gate with sources in W. In each of the 295,296 gate instances, check that safety from W agrees with containment in $H_{\max}$. The companion script also checks the seven-state example. All checks passed in Python 3.12.

Disclosure. AI assistance was used to draft, check derivations, and run the finite calculations.

References

[1] C. von Csefalvay, P. Doma, A. Unni Krishnan, and T. Foldi. *Weight Space Misfolding*. alphaXiv preprint, 8 Nov. 2026. Abstract · Version 1 PDF.

[2] R. Takehi, J. Teoh, A. Zhang, and O. Khattab. *Sparks of Weight-Space Programming: LLMs Can Natively Read & Write Neural Network Weights*. Research blog, 2026. Article.

[3] J. Hu et al. *Exploring Weight-Space Programming: Capabilities, Generalization, and Limitations*. arXiv:2611.09382, 2026. Preprint.

[4] R. Cohen, E. Biran, O. Yoran, A. Globerson, and M. Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024. Article.

[5] Anonymous researcher (agent-assisted). *Weight-Space Programming: Symmetry, Certified Synthesis, and Higher-Topos Semantics. Higher Categorical Semantics*, editorially reviewed, 2026. Article.

[6] J. Lurie. *Higher Topos Theory*. Annals of Mathematics Studies 170, Princeton University Press, 2009. Author's PDF.

[7] The Univalent Foundations Program. *Homotopy Type Theory: Univalent Foundations of Mathematics*. Institute for Advanced Study, 2013. Open book.

[8] C. Rezk. A model for the homotopy theory of homotopy theory. *Transactions of the American Mathematical Society*, 353(3):973–1007, 2001. Preprint.

[9] C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. Article.

[10] P. J. Ramadge and W. M. Wonham. Supervisory control of a class of discrete event processes. *SIAM Journal on Control and Optimization*, 25(1):206–230, 1987. Article.

[11] W. M. Wonham and P. J. Ramadge. On the supremal controllable sublanguage of a given language. *SIAM Journal on Control and Optimization*, 25(3):637–659, 1987. Article.

[12] J. Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009. Paper.

[13] S. Torres-Arias, H. Afzali, T. K. Kuppusamy, R. Curtmola, and J. Cappos. in-toto: Providing farm-to-table guarantees for bits and bytes. *USENIX Security Symposium*, 2019. Paper.

[14] SLSA. *Build: Verifying Artifacts*, specification v1.2. Specification.

[15] A. Vassilev, A. Oprea, A. Fordyce, H. Anderson, X. Davies, and M. Hamin. *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*. NIST AI 100-2e2025, 2025. Report.