Structural limits of observational learning for interventional competence

Marcus A. Thomas*
Memorial Sloan Kettering Cancer Center
New York, NY, USA

Abstract

Whether observational training can in principle support interventional competence is an open question. We establish that in general it cannot when mechanisms violate locality and autonomy and the observational distribution rewards the resulting spurious dependencies.

Realizability. We prove that interventional competence imposes a necessary condition on a system's internal states: post-intervention outputs must be invariant across all internal configurations consistent with the same intervention target.

Obstruction. We prove that observational learning drives systems away from this condition when mechanisms violate computational locality and autonomy, with gradient descent on predictive loss producing directions that strictly increase interventional error. The obstruction sharpens in agentic settings: under endogenous interventions, violations prevent interventional targets from being represented in an invariant, implementation-independent way. A separate geometric lower bound applies when the hypothesis class does not contain the true causal model: interventional error is bounded below by the L^2 distance from the class to the true interventional mapping, independent of access mode.

Amplification. We prove that local violations compound under sequential composition whenever context sensitivity persists, and apply the framework to interpret large language models.

We separately characterize minimal additional structure sufficient to overcome each obstruction: disagreement-maximizing criticism for within-class ambiguity in the exogenous case; architectural or external constraints in the endogenous case; and hypothesis-class expansion when the class is inadequate. None is available to observational training, so the gap between observational training and interventional deployment cannot be closed by additional data or optimization alone.

1 Introduction

Modern learning systems are trained almost exclusively on observational data: passively collected records of co-occurring variables under fixed data-generating processes. They are

*Corresponding author. Email: thomm15@mskcc.org.

then deployed in settings that demand reliable behavior under intervention—experimental manipulation, altered operating conditions, or the system's own actions. Whether training restricted to observational data can, in principle, support interventional competence is the question addressed in this paper.

We distinguish two notions of competence. *Observational competence* concerns performance with respect to a fixed empirical data distribution, whether obtained through passive observation or structured data collection, under which input variations do not constitute targeted manipulations of underlying causal mechanisms and the data-generating processes remain fixed throughout. *Interventional competence* is stronger: it refers to the capacity to support correct behavior under interventions that alter the data-generating process in mechanism-targeted ways, including interventions not realized in the observed data, together with invariance to changes in causally irrelevant aspects of context. Formal definitions appear in Section 3 and Appendix A. The key distinction is that observational competence may be silent about behavior under changes to the data-generating process, whereas interventional competence requires correctness precisely under such changes.

Our results rest on a single structural principle, the Locality–Autonomy Principle (LAP), formalized in Section 3. LAP is not a novel assumption; it makes explicit a condition implicit in standard interventionist frameworks [1, 2]. We treat it as a condition under which interventional targets are well-defined for a learning system, and distinguish domain-level LAP from model-level locality and autonomy (L&A). The central claim is that when L&A is violated at the computational level, observational learning cannot repair the mechanisms required for interventional competence: computational violations generically propagate to the semantic level, and observational training provides no process by which they can be detected and corrected.

To connect these results to contemporary practice, we interpret large language models as compositions of parametric mechanisms whose behavior under input perturbation is well-defined even without assuming causal correctness. Each token position constitutes a mechanism; the autoregressive structure defines a directed chain.

2 Relationship to Existing Approaches

The methods most directly related to this work divide into two groups according to whether they operate within the observational regime or make genuine use of interventional signal.

Observational and multi-environment methods. Invariant risk minimization [3], causal representation learning from observational data [4], out-of-distribution generalization methods [5], and invariance-based causal inference [6] all operate within the observational regime. They seek to extract invariant structure from observed data, whether from single or multiple environments. Our results characterize the structural ceiling of that entire regime. These methods may improve observational robustness, but the gradient obstruction in Theorem 3 applies to any procedure driven by observational loss, and therefore bounds their interventional reach regardless of the sophistication of the invariance objective.

The deeper limitation of multi-environment methods such as IRM is not merely that they exploit observational loss. It is that they never instantiate a commitment to a specific

causal abstraction that could be refuted by a targeted intervention. Environments in IRM are sources of distributional heterogeneity; the method extracts predictors that are statistically stable across them. Statistical stability across training environments is not the same as a falsifiable prediction about what will happen under a novel do(X=x). The system makes no commitment that could be exposed to disagreement-maximizing criticism, and therefore has no mechanism for detecting and revising causal errors of the kind Theorem 3 identifies. Formal guarantees for IRM and its variants are severely limited in nonlinear and latent-variable settings [5], a finding consistent with this analysis.

Methods that use genuine interventional signal. A separate line of work uses actual interventional data rather than multi-environment observational proxies. Ahuja et al. [7] prove identifiability of latent causal representations from interventional distributions, without structural assumptions on the graph. CITRIS [8] and its extension iCITRIS use temporal sequences in which intervention targets are *observed*, achieving identifiability for multidimensional causal factors. Brehmer et al. [9] use paired pre- and post-intervention observations—without known targets—and prove identifiability under perfect interventions. Ke et al. [10] combine observational and interventional data with unknown targets for structure discovery. Because these methods receive genuine interventional signal, the gradient obstruction in Theorem 3 does not apply to intervention patterns covered in their training distributions.

We acknowledge this boundary explicitly. Our impossibility results concern systems trained exclusively on observational data; they do not apply to procedures that incorporate interventional training signal of the kind these methods require.

Two limitations nonetheless prevent these methods from closing the gap identified by our results. First, their identifiability guarantees depend on conditions that do not generalize to open-world settings: [7] requires interventional distributions over known targets; [8] requires that intervention targets be observed at each time step; [9] requires perfect interventions covering all variables; and [10] requires access to interventional data with unknown but structurally constrained targets. None of these conditions is available in the open-world deployment settings—natural language, agentic interaction, real-world planning—that motivate the present work.

Second, and more fundamentally, identifiability from interventional training data is a property of the training procedure relative to the training distribution. It establishes that the learned representation correctly encodes causal structure for the intervention patterns seen during training. It does not instantiate the commitment-refutation structure that Conjecture 1 identifies as necessary for reliable interventional competence at deployment time. The gap between identifiability from training data and interventional competence under deployment is precisely the gap between encoding a causal abstraction statically and participating in a process by which failures of that abstraction become detectable and revisable.

Causal abstraction. The literature on formal abstraction between causal models at different levels of description [11–14] is discussed alongside the alignment condition in Section 4.

The Pearl Causal Hierarchy and neural learnability. The Causal Hierarchy Theorem [15] establishes that observational (Layer 1) data generically underdetermines interventional (Layer 2) and counterfactual (Layer 3) quantities: the subset of structural causal models for which the hierarchy collapses has measure zero. Xia et al. [16] show that this ceiling is not an artifact of limited model capacity—it persists for arbitrarily expressive neural models, so that no observationally trained network, however large, recovers interventional quantities from Layer 1 data alone; restoring identifiability requires injecting a structural inductive bias that encodes the necessary constraints. Our results are situated beneath this ceiling rather than in competition with it. The Causal Hierarchy Theorem is a statement about *information*: it characterizes what is recoverable from a distribution, with the learner abstracted away and consistency of estimation assumed. The obstruction we identify is a statement about *dynamics*: granting the underdetermination the hierarchy describes, we ask what observational optimization does to a learner's computational mechanisms as training proceeds, and show that gradient descent on predictive loss moves the system *away* from the realizability condition of Theorem 1 rather than merely failing to reach it. The two results compose: our obstruction operates within the space the Causal Hierarchy Theorem certifies as underdetermined, and neither implies the other. The distinction is sharpest at the point where the inductive bias of [16] is the proposed remedy. That construction restores identifiability when the encoded structural constraints are correct, i.e. when the true model lies in the hypothesis class. Our realizability-gap analysis (Theorem 6) addresses the complementary regime: when the class does not contain the true interventional mapping, the residual error is bounded below independently of data-access mode, and closing it is a problem of hypothesis-class *generation* rather than selection—a regime the learnability results, which fix the model class in advance, do not reach.

3 Formal Setup

This section establishes the conceptual and formal foundations on which the main results depend, beginning with the conditions that make interventional targets well-defined in a domain and proceeding to the corresponding conditions we impose on a learning system's internal structure. Extended formulations appear in Appendix B.

3.1 Parametric mechanisms

A learning system can only be evaluated against an interventional standard if its internal components are understood as representations of causal processes rather than as the processes themselves. The gap between representation and reality is precisely what the main results exploit: a mechanism that predicts correctly under observation may nonetheless fail to isolate the causal structure that governs behavior under intervention. We model the internal components of a learning system as *parametric mechanisms*: parameterized mappings that may be intended to represent autonomous causal processes, but whose causal adequacy is a substantive question rather than a consequence of predictive accuracy.

Definition 1 (Parametric Mechanism)**.** A parametric mechanism associated with variable X_i is a function

$$\mathcal{M}_i : \mathcal{X}_{\mathrm{PA}(i)} \times \mathcal{U}_i \times \Theta_i \to \mathcal{X}_i,$$

where $\mathcal{X}_{\mathrm{PA}(i)}$ is the state space of the parents of X_i, $\mathcal{U}_i$ represents exogenous influences, Θ_i is a parameter space, and $\mathcal{X}_i$ is the state space of X_i.

The inclusion of parameters reflects the fact that mechanisms are represented rather than directly observed. Causal adequacy is determined by whether the represented mechanism respects the invariances implied by intervention. If it obtains, predictive adequacy follows, but the converse is not true in general.

3.2 The Locality–Autonomy Principle

Interventions are only meaningful against a background of what does not change. When one aspect of a system is altered, the alteration acquires causal content precisely because other governing constraints persist. Without this background of invariance, the distinction between targeting one mechanism and disrupting the entire system collapses, and the interventional target loses its identity as a well-defined object. The Locality–Autonomy Principle makes this background requirement explicit as a condition on the domain being modeled.

Principle 1 (Locality–Autonomy Principle (LAP))**.** For interventions to be well-defined in a domain, the domain's causal structure must satisfy two conditions. *Locality* requires that each causal mechanism depends only on the states of its causal parents, so that the functional form of the mechanism governing a variable is not deformed by changes to the states of variables outside its direct causes. *Autonomy* requires that ideal interventions replace one mechanism without altering the others, so that the functional form of each unaffected mechanism is not deformed by changes to the parameters defining any other mechanism, and the rest of the system constitutes a stable background against which the targeted change is interpretable.

LAP is not an empirical assumption about particular domains, but a coherence condition on the semantics of intervention itself. It specifies the structural prerequisites under which the question of what would happen under a given intervention has a determinate answer. When LAP holds, the interventional targets in the competence definitions of Section 3 are well-defined objects that a model can in principle be evaluated against. When LAP fails, the notion of intervening on one mechanism while holding others fixed loses its referent, because the outcome depends on how the intervention was carried out rather than solely on what it targeted. In this sense LAP is prior to any particular causal framework: it is the condition that makes interventional semantics coherent at all, and it is shared by Pearl's structural causal models [1] and Woodward's interventionist account [2], both of which define causal relationships in terms of invariance under idealized interventions specified independently of the system under study.

Given that LAP holds in a domain, a further question arises about the learning system that is meant to operate in that domain: do the system's internal mechanisms reflect the same structural organization? We use *locality and autonomy* (L&A) to refer to this model-level condition, reserving LAP for the domain-level principle. L&A applies at two distinct levels within a model. At the *computational level*, it asks whether the system's internal

mechanisms depend only on the appropriate variables and parameters, in the sense that each mechanism responds to its computational parents and is not deformed by interventions targeting unrelated mechanisms. At the *semantic level*, it asks whether the model's outputs, construed as claims or assertions about causal structure in the world, track genuine causal organization rather than merely reflecting statistical regularities in the training data. The theorems directly concern computational L&A; semantic L&A is the deeper target that computational violations undermine. The relationship between these levels is developed in Appendix C.

In smooth parametric systems, computational L&A admits an analytic characterization via Lie derivatives along intervention-generating flows. Let $W_j := \partial/\partial X_j$ denote perturbation of state variable X_j, and Ξ_A the vector field on parameter space generated by interventions on θ_A. Locality and autonomy require, respectively,

$$\mathcal{L}_{W_j}\mathcal{M}_i = 0 \;\; (j \notin \mathrm{PA}(i)), \qquad \mathcal{L}_{\Xi_A}\mathcal{M}_i = 0 \;\; (i \neq A).$$

These serve as diagnostic witnesses in differentiable systems, not as definitions of LAP. The coordinate-free formulation accommodates systems with shared parameterization where no canonical decomposition of parameters into mechanism-specific blocks is given in advance. See Appendix B.1.

The three-level picture that emerges from this discussion organizes the rest of the paper. LAP is what makes the interventional target well-defined in the domain. Computational L&A is a sufficient condition for a model's internal mechanisms to represent that target correctly, though semantic L&A can in principle be maintained through active error correction even when computational L&A is violated. Semantic L&A is what must hold for the model's outputs to constitute genuine causal knowledge about the world rather than patterns in the usage of causal language. The theorems establish that observational training actively undermines computational L&A, and that violations at the computational level propagate generically to the semantic level, closing off the possibility of passive semantic correction within the observational regime.

3.3 Observational and interventional competence

Definition 2 (Observational Competence)**.** A model f is observationally competent relative to P_{obs} if it minimizes expected loss under the observational distribution:

$$f \in \arg\min_{g} \; \mathbb{E}_{(x,y)\sim P_{\text{obs}}}\big[L(g(x), y)\big].$$

Definition 3 (Interventional Competence)**.** Fix a task with causal variables (X, Y), an observational training distribution P_{obs}, and a class $\mathfrak{E}$ of admissible interventions. A model f supports interventional competence relative to a nontrivial intervention domain $\mathfrak{E}_{\text{int}} \subseteq \mathfrak{E}$ if: (i) $\mathfrak{E}_{\text{int}}$ contains interventions not realized in the observational regime; (ii) for all $\mathrm{do}(X{=}x) \in \mathfrak{E}_{\text{int}}$ and admissible contexts c, the model's output agrees with the true interventional target $f(x, c) = \mathbb{E}[Y \mid \mathrm{do}(X{=}x), c]$; and (iii) the model is invariant to context variation that is interventionally irrelevant for the task on $\mathfrak{E}_{\text{int}}$.

This definition makes competence explicitly domain-relative: it is a claim about correct and invariant behavior over a specified nontrivial set of interventions, not an absolute

guarantee over all conceivable manipulations. We further distinguish exogenous interventional competence, in which interventions are specified externally in the classical sense of Pearl's do(·) operator, from endogenous interventional competence, in which the system itself generates interventions through its own actions, policies, or internal mechanisms. The endogenous setting is strictly more demanding: it requires correctness under a family of policy substitutions over intervention-generating mechanisms. Both notions are formalized in Appendix A.

4 Main Results

4.1 Internal Realizability of Interventional Mappings

Interventional competence is defined at the task level, but it must be realized by a system's internal dynamics. That an interventionally competent system implements an intervention-invariant mapping is a near-immediate consequence of the definitions. The substantive question is whether such a mapping is well-defined within the system's computation. Theorem 1 answers this by identifying a necessary and sufficient condition on the system's internal states: post-intervention outputs must be invariant across all internal configurations consistent with the same intervention target. This alignment condition is the structural target that Theorem 3 then shows observational learning drives systems away from.

Theorem 1 (Internal Realizability of Interventional Mappings)**.** *Let (X, Y) be variables related by a causal process, and suppose a model f is interventionally competent on a task involving interventions* $\mathrm{do}(X{=}x)$ *over a nontrivial domain.*

(a) Necessity. *There exists a mapping $\psi : \mathcal{X} \to \mathcal{Y}$ such that, for all x in the intervention domain, the model's output under* $\mathrm{do}(X{=}x)$ *is given by $Y = \psi(x, c)$, where the functional form of ψ is invariant to the mechanism by which X is set.*

(b) Alignment. *Let $\mathcal{M}$ be the system's computational mechanism over internal states Z, inducing output distribution $P_{\mathcal{M}}$, with projections $\pi_X : Z \to \mathcal{X}$ and $\pi_Y : Z \to \mathcal{Y}$, and let $I_x : Z \to Z$ be an internal intervention map satisfying $\pi_X(I_x(z)) = x$. The task-level interventional distribution* $P(Y \mid \mathrm{do}(X{=}x))$ *is well-defined via the system's internal dynamics if and only if, for all x and all z_1, z_2 with $\pi_X(z_1) = \pi_X(z_2) = x$,*

$$(\pi_Y)_* P_{\mathcal{M}}(\cdot \mid I_x(z_1)) = (\pi_Y)_* P_{\mathcal{M}}(\cdot \mid I_x(z_2)).$$

Proof sketch. For part (a): the interventional distribution $P(Y \mid \mathrm{do}(X{=}x), c)$ depends only on (x, c), not on which protocol sets X to x. A model giving different outputs for the same (x, c) under different protocols would fail to match this unique target under at least one, contradicting interventional competence. The mapping $\psi(x, c) := f(x, c)$ is therefore well-defined and intervention-invariant. For part (b): if post-intervention outcome distributions differ across internal states sharing the same projected value x, the distribution $P(Y \mid \mathrm{do}(X{=}x))$ depends on task-irrelevant internal degrees of freedom and is not well-defined as a function of x alone. The converse is immediate. □

Full proofs appear in Appendix D (Proofs 14 and 22). Part (a) is architecture-independent but close to definitional. The nontrivial content is the alignment condition in part (b): it characterizes when a system's internal dynamics can represent a coherent interventional

mapping at all. The alignment condition is violated whenever post-intervention behavior depends on internal features not determined by the intervention target. Such dependence is invisible to observational evaluation but constitutes interventional failure. Together, parts (a) and (b) establish the realizability criterion that Theorems 3 and 5 then show observational learning cannot meet.

The X-fiber formulation. The alignment condition in part (b) admits a geometric reformulation that makes the locus of failure explicit. For each $x \in \mathcal{X}$, define the *X-fiber* over x as the set of internal states projecting to x:

$$F_x := \{z \in Z : \pi_X(z) = x\}.$$

Elements of F_x are internal states that are task-level indistinguishable with respect to X: they may differ arbitrarily in internal degrees of freedom, but all correspond to the same external value of the intervention target. The alignment condition then states that the post-intervention output distribution must be constant across each X-fiber:

$$(\pi_Y)_* P_{\mathcal{M}}(\cdot \mid I_x(z_1)) = (\pi_Y)_* P_{\mathcal{M}}(\cdot \mid I_x(z_2)) \qquad \text{for all } z_1, z_2 \in F_x.$$

Failure of the alignment condition corresponds to post-intervention behavior varying along directions within a fiber, i.e., along directions that do not alter the intervention target.

In smooth systems, the fiber condition admits a local counterpart. The tangent space $T_z Z$ at each internal state decomposes with respect to the projection π_X, and the subspace $\ker(d\pi_X)$ consists of infinitesimal directions that leave X unchanged. These are the directions tangent to the X-fiber through z. Failure of the alignment condition is then witnessed by non-vanishing of the Lie derivative

$$\mathcal{L}_v\big((\pi_Y)_* P_{\mathcal{M}}(\cdot \mid I_x(\cdot))\big) \neq 0 \qquad \text{for some } v \in \ker(d\pi_X).$$

The following corollary connects the fiber condition to the computational L&A violations that Theorem 3 analyzes.

Corollary 2 (L&A Violations Imply Alignment Failure)**.** *Suppose the model exhibits an autonomy violation ($\varepsilon^{A,i}_{\text{aut}} > 0$) or a locality violation ($\varepsilon^{A,i}_{\text{loc}} > 0$) with respect to a pair (A, i) where $i \notin \text{Desc}(A)$. Then the X-fiber condition fails: the task-level interventional distribution $P(Y \mid \text{do}(X_A{=}a))$ is not well-defined via the system's internal dynamics.*

Proof sketch. L&A violations produce nonzero uniform spurious dependence of $\mathcal{M}_i$ on θ_A or X_A. In smooth systems, these manifest as non-vanishing Lie derivatives along directions that are task-irrelevant for X_i, which lie in $\ker(d\pi_{X_i})$ by construction. The fiber condition fails along precisely these directions. Full proof in Appendix D (Corollary 23). □

Corollary 2 establishes that the alignment condition is not merely a convenient target but the exact structural property that L&A violations destroy. Theorem 3 then shows that observational learning drives systems toward the violations identified here, closing the chain from training dynamics to failure of interventional realizability.

Relationship to causal abstraction. The alignment condition in part (b) is related to conditions developed in the causal abstraction literature. Rubenstein et al. [11] introduced exact transformations between structural equation models at different levels of description, and Beckers and Halpern [12] developed a hierarchy of abstraction notions requiring that interventions commute with the mapping between low-level and high-level causal models. Geiger et al. [13, 14] operationalized this framework for neural networks via interchange interventions, in which internal activations are swapped between inputs that agree on a high-level variable and the output is checked for invariance.

The alignment condition differs from these in three respects. First, it does not presuppose two complete causal models. The projections π_X and π_Y define task variables, but no separate high-level SCM with its own structural equations is required. The condition asks whether the system's internal dynamics support *any* coherent interventional semantics at the projected level. Second, in the causal abstraction framework the abstraction relation is a diagnostic or a training objective. Here, it is a necessary condition whose violation is then shown to be actively reinforced by observational learning (Theorem 3). Third, the condition is stated for probabilistic systems without requiring determinism, whereas the core results of Beckers and Halpern [12] apply to deterministic causal models, with probabilistic extensions treated separately [17].

4.2 Observational gradients are misaligned with interventional error

We now show that observational optimization generically drives systems away from the alignment condition that Theorem 1 identifies. The obstruction is not merely that interventional errors are unobserved, but that observational gradients actively point toward increased interventional error.

Autonomy violations arise when a mechanism $\mathcal{M}_i$ depends spuriously on parameters θ_A associated with an unrelated mechanism $\mathcal{M}_A$. Locality violations arise when $\mathcal{M}_i$ depends spuriously on the state of a variable X_A that is not among its true causal parents. In either case, the observational distribution may contain correlations (e.g., $\mathrm{Cov}(X_A, X_i) \neq 0$) that reward these spurious dependencies by reducing observational loss. Yet under the true causal semantics, interventions on X_A do not affect non-descendants $X_i \notin \mathrm{Desc}(A)$. The model therefore incurs interventional error under $\mathrm{do}(X_A{=}a)$, and observational gradients generically point in the direction that increases this error.

Theorem 3 (Gradient Misalignment Under Computational L&A Violations)**.** *Consider a causal system with true graph $G = (V, E)$ and a model with parametric mechanisms $\{\mathcal{M}_i : i \in V\}$. Fix a variable $A \in V$ and a non-descendant $i \notin \mathrm{Desc}(A)$, so that under the true causal semantics interventions on X_A do not affect X_i.*

Suppose the model violates (L&A) at the computational level with respect to the pair (A, i), i.e. the model's mechanisms fail to reflect the structural conditions that LAP requires of the domain, in at least one of the following senses:

*(I) (*Autonomy violation*) The mechanism $\mathcal{M}_i$ depends spuriously on a parameter block θ_A associated with $\mathcal{M}_A$, with directional spurious dependence magnitude $\varepsilon^{A,i}_{\mathrm{aut}}(\hat{v}) > 0$ along a unit direction $\hat{v}$ in parameter space.*

*(II) (*Locality violation*) The mechanism $\mathcal{M}_i$ depends spuriously on the state X_A, with directional spurious dependence magnitude $\varepsilon^{A,i}_{\text{loc}}(\hat{w}) > 0$ along a unit direction $\hat{w}$ in $\mathcal{X}_A$.*

Assume further that the observational distribution exhibits correlations that allow these spurious dependencies to reduce squared-error observational loss.

Then the set of directions along which interventional error grows at least linearly in displacement magnitude forms an open cone in parameter space (autonomy case) or state space (locality case). Moreover, in the small-violation regime, the observational gradient has nonzero projection onto this cone except on a measure-zero set of parameter configurations: gradient descent on L_{obs} decomposes into a component lying in the violation cone, along which interventional error strictly increases, and a component outside it. The directional claim along the specific $\hat{v}$ (autonomy) or along the parameter direction v coupled to $\hat{w}$ (locality) instantiates this decomposition: along these directions, the directional derivatives of observational loss and interventional error have strictly opposite signs.

Remark 1 (Lie derivatives as diagnostic witnesses). The magnitudes $\varepsilon^{A,i}_{\text{aut}}(\hat{v})$ and $\varepsilon^{A,i}_{\text{loc}}(\hat{w})$ admit a coordinate-free characterization as directional derivatives, coinciding with Lie derivatives along intervention-generating flows. They are diagnostic witnesses of L&A violations in differentiable systems, not definitions. LAP itself requires no differentiability. See Appendix B.1.

Proof sketch. The argument has two parts. First, the interventional error bound: under an autonomy violation, the model's output for node i under $\text{do}(X_A = a)$ depends on θ_A through the spurious coupling, while the true output does not. The fundamental theorem of calculus applied along the segment $\theta_A(t) = t\phi\hat{v}$ yields $|\mathcal{M}_i(\cdot; \phi\hat{v}) - \mathcal{M}_i(\cdot; 0)| \geq |\phi| \cdot \varepsilon^{A,i}_{\text{aut}}(\hat{v})$. Under a locality violation, the analogous bound along $\hat{w}$-displacements is $|s| \cdot \varepsilon^{A,i}_{\text{loc}}(\hat{w})$, where $a = a_0 + s\hat{w}$ and a_0 is a calibration value at which the model is correct.

Second, the gradient misalignment: the observational correlation $\text{Cov}(X_A, X_i) \neq 0$ means that gradient descent on squared-error loss drives parameters in a direction that exploits this correlation. For autonomy violations, this direction is $\hat{v}$, and movement along $\hat{v}$ increases interventional error by part one. For locality violations, the direction is a $v \in \Theta_i$ whose action increases the directional sensitivity of $\mathcal{M}_i$ to X_A along $\hat{w}$, so movement along v likewise increases interventional error. Along these directions, the directional derivatives of observational loss and interventional error have strictly opposite signs, establishing structural misalignment. □

The full proofs—separately treating autonomy and locality violations—appear in Appendix D (Proofs 15 and 17). The directional-derivative formulation provides the coordinate-free witness: the non-vanishing of $\langle \partial_{\theta_A} \mathcal{M}_i, \hat{v} \rangle$ along the intervention-generating direction is what makes the obstruction geometric rather than coordinate-dependent.

Remark 2 (Cone structure of the obstruction). The violation-amplifying directions identified in Theorem 3 form an open cone, not a single ray. For autonomy violations, the cone consists of all unit directions $\hat{v}$ in parameter space satisfying $\varepsilon^{A,i}_{\text{aut}}(\hat{v}) > 0$; for locality violations, the analogous cone in state space is defined by $\varepsilon^{A,i}_{\text{loc}}$. Step 1 of the proof establishes the linear lower bound on interventional error along every direction in the cone, not only along the specific $\hat{v}$ or $\hat{w}$ named in the statement.

The substantive content of the obstruction is therefore geometric: the observational gradient decomposes into a projection onto the violation cone and a projection onto its

complement. The first projection strictly increases interventional error by Step 1; the second may reduce both observational loss and interventional error. Generically the first projection is nonzero. The exceptional case in which it vanishes corresponds to the cone lying entirely in the orthogonal complement of ∇L_{obs}, which requires fine-tuning between the spurious sensitivity structure of $\mathcal{M}_i$ and the residual covariance structure of the observational distribution. Under the non-degeneracy condition in assumption (vii), this exceptional case is excluded.

The conclusion does not depend on identifying a single descent direction that maximally increases interventional error. It depends only on the violation cone having nonempty intersection with the half-space of observational descent directions, which holds generically. Observational optimization therefore provides a pathway to increased interventional error along a positive-measure set of directions, of which the specific $\hat{v}$ (autonomy) and the parameter direction v coupled to $\hat{w}$ (locality) are representative instances.

4.3 Endogenous Interventions and Implementation-Dependent Semantics

Theorem 3 treats exogenous interventions. We now turn to the endogenous case (Appendix Section A, Definition 7), in which the obstruction changes character: L&A violations induce implementation dependence. Distinct policy realizations equivalent at the intervention target may yield different downstream outcomes, so interventional quantities that are invariant in the exogenous setting need not be invariant under endogenous realization absent additional structural constraints.

Theorem 4 (Endogenous L&A Violations Induce Implementation Dependence)**.** *Consider an endogenous SCM with mechanisms $\{\mathcal{M}_i : i \in V\}$ and admissible policy substitutions Π_A acting on variable $A \in V$. For each $\pi_A \in \Pi_A$, let $\mathcal{M}^{\pi_A}$ denote the counterfactual system obtained by replacing the original A-generating mechanism with π_A while leaving all other mechanisms fixed.*

The following results show that, under L&A violations, there exist policy substitutions equivalent at the target level whose downstream outputs diverge. This does not imply that interventional targets are ill-defined in an absolute sense, but rather that they are not invariant with respect to the system's internal realization of interventions.

(a) Autonomy violations. *Fix $i \neq A$ and suppose $\mathcal{M}_i$ depends spuriously on a parameter block θ_A associated with $\mathcal{M}_A$. Suppose there exists a unit vector $\hat{v}$ in parameter space and a scalar $\varepsilon^{A,i}_{\text{aut}}(\hat{v}) > 0$ such that the directional derivative of $\mathcal{M}_i$ along $\hat{v}$ is uniformly bounded below in magnitude by $\varepsilon^{A,i}_{\text{aut}}(\hat{v})$. If Π_A contains policies π_A, π'_A with $\theta_A(\pi_A) - \theta_A(\pi'_A)$ aligned with $\hat{v}$, then*

$$\left|X_i^{\mathcal{M}^{\pi_A}} - X_i^{\mathcal{M}^{\pi'_A}}\right| \;\geq\; \varepsilon^{A,i}_{\text{aut}}(\hat{v}) \cdot \|\theta_A(\pi_A) - \theta_A(\pi'_A)\|.$$

(b) Locality violations. *Fix $i \notin \text{Desc}(A)$ and suppose $\mathcal{M}_i$ depends spuriously on the state X_A. Suppose there exists a unit vector $\hat{w}$ in $\mathcal{X}_A$ and a scalar $\varepsilon^{A,i}_{\text{loc}}(\hat{w}) > 0$ such that the directional derivative of $\mathcal{M}_i$ along $\hat{w}$ is uniformly bounded below in magnitude by $\varepsilon^{A,i}_{\text{loc}}(\hat{w})$. If Π_A contains policies π_A, π'_A with $X_A^{\pi_A} - X_A^{\pi'_A}$ aligned with $\hat{w}$, then*

$$\left|X_i^{\mathcal{M}^{\pi_A}} - X_i^{\mathcal{M}^{\pi'_A}}\right| \;\geq\; \varepsilon^{A,i}_{\text{loc}}(\hat{w}) \cdot \|X_A^{\pi_A} - X_A^{\pi'_A}\|.$$

In both cases, distinct policy substitutions that should leave X_i unchanged instead produce different outputs, with discrepancies bounded below by the violation magnitude along the relevant direction.

Proof sketch. In both parts, the fundamental theorem of calculus applied along the line segment between the two parameter (or state) values yields a lower bound proportional to the aligned component of the policy difference and the directional spurious dependence magnitude. Since $i \neq A$ in (a) and $i \notin \mathrm{Desc}(A)$ in (b), the discrepancy reflects an L&A violation rather than a legitimate intervention effect. □

Full proofs appear in Appendix D (Proofs 20 and 21). Theorem 4 sharpens the obstruction. In the exogenous setting, $P(Y \mid \mathrm{do}(X_A{=}a))$ is a fixed target the model is wrong about. In the endogenous setting, the target itself fails to be well-defined: $P(Y \mid \mathrm{do}(X_A{=}a))$ is not represented as a function of a alone, because its value depends on which policy produced a.

This matters for deployment. The obstruction in Theorem 3 could in principle be addressed by interventional feedback that exposes the model's errors relative to the true target. The obstruction in Theorem 4 cannot be addressed this way, because there is no unique target against which to measure the error. Policies that should be interventionally equivalent produce different outputs, and no external signal is available to adjudicate among them.

4.4 Error amplification under sequential composition

Theorems 1 and 3 concern individual mechanisms. In modern architectures, especially autoregressive and staged systems, mechanisms are applied sequentially. Even small local deviations can have large downstream effects because later mechanisms operate on contexts that may already be perturbed. The following result shows that this amplification is structural.

Theorem 5 (Error Amplification in Causal Chains)**.** *Consider a sequence of parametric mechanisms*

$$X_t = \mathcal{M}_t(X_{<t}, U_t), \qquad t = 1, \ldots, T,$$

with compact state spaces and bounded exogenous noise, and assume each mechanism is Lipschitz so that small perturbations in its inputs induce proportionally small perturbations in its outputs. Let X_t^* denote a reference trajectory. At each step t, the mechanism may introduce a fresh deviation $\delta_t \geq 0$ from the reference, with $\delta_t = 0$ when no new error occurs. Define the per-step fresh conditional deviation magnitude $\varepsilon_t := \mathbb{E}[\delta_t \mid \delta_t \neq 0]$.

Assume that a fresh deviation introduced at step t persists to influence downstream steps with probability at least $p_{\min} > 0$ per step, and that along any surviving deviation lineage, propagated magnitude contracts by a factor between β' and β per step, with $0 < \beta' \leq \beta < 1$. In the small-error regime, total deviation at each step decomposes additively into contributions from distinct origination times. Then the cumulative error $D_{\mathrm{total}} := \sum_{s=1}^{T} d(X_s, X_s^*)$ satisfies

$$\mathbb{E}[D_{\mathrm{total}}] \;\geq\; \sum_{t=1}^{T} \Pr(\delta_t \neq 0) \cdot \varepsilon_t \cdot \frac{1 - (p_{\min}\beta')^{T-t+1}}{1 - p_{\min}\beta'}.$$

In particular, even rare early deviations can accumulate and dominate total error whenever the effective persistence factor $p_{\min}\beta'$ is bounded away from zero.

Proof sketch. Each fresh deviation is tracked from its point of origin through subsequent steps. On the event that the deviation lineage survives from step t to step $s > t$, propagated magnitude is bounded below by $(\beta')^{s-t}$ (contraction) and survival probability is bounded below by $(p_{\min})^{s-t}$ (persistence). The expected contribution of a deviation originating at t to the error at step s is therefore at least $\Pr(\delta_t \neq 0) \cdot (p_{\min}\beta')^{s-t} \cdot \varepsilon_t$. Summing over all downstream positions and all origination points yields the geometric series bound. □

The contraction assumption ($\beta < 1$) ensures that each mechanism is individually well-behaved: along any single lineage, propagated magnitude shrinks per step. The interest of Theorem 5 lies in the subtler regime where each step is locally contractive, yet early deviations still propagate forward rather than vanishing because new fresh deviations compound with the residual influence of earlier ones. When $p_{\min}\beta' \to 0$, deviations dissipate quickly and errors remain localized. When $p_{\min}\beta'$ is bounded away from zero, even infrequent early deviations contribute disproportionately to total error. The full proof, including the detailed additive decomposition and conditional independence assumptions, appears in Appendix D (Proof 19).

Remark 3 (Connection to Theorems 1 and 3). Autonomy and locality violations force downstream mechanisms to remain sensitive to upstream deviations, yielding a persistence factor $p_{\min}\beta'$ bounded away from zero. Interventional errors created by the gradient obstruction of Theorem 3 are therefore amplified under sequential composition: Theorem 5 converts local misalignment into global error accumulation.

4.5 Hypothesis space expansion

The results in Sections 4 assume the true causal model/mechanism $M^* \in \mathcal{H}$ throughout. Theorems 3 and 4 concern errors within the observational equivalence class $\mathcal{E}(P_{\text{obs}}) \cap \mathcal{H}$ and obstructions to resolving that class from observational data. These results are silent on a distinct question: what happens when no member of $\mathcal{H}$ corresponds to M^* at all?

Realizability is a strong assumption. It presupposes that the hypothesis class is expressive enough to contain the true causal mechanism, so that the learning problem reduces to selection within $\mathcal{H}$. Hypothesis classes are in practice chosen for tractability, prior knowledge, or architectural constraint, with no guarantee that they contain M^*. A fixed $\mathcal{H}$ reflects the system's current representational capacity, and whether that capacity suffices is itself part of the learning problem.

When $M^* \notin \mathcal{H}$, the analysis changes character. No selection procedure within $\mathcal{H}$ can produce an interventionally competent model, because the target is not a member of the search space. The residual error is not a function of how $\mathcal{H}$ is navigated but of how far $\mathcal{H}$ lies from M^*. Observational data cannot reveal this distance: distinct $\mathcal{H}$-members closest to M^* in an interventional sense may be observationally indistinguishable, and none carries a signature announcing its inadequacy. Interventional access does reveal the distance but cannot close it: refutations of $\mathcal{H}$-members establish that $\mathcal{H}$ is inadequate, not how to repair it.

This section formalizes the irreducibility of this gap. The key object is a geometric distance from M^* to $\mathcal{H}$ measured in the space of interventional mappings. This distance lower-bounds the interventional error of every $\mathcal{H}$-member, and therefore of every protocol restricted to

outputs in $\mathcal{H}$. The bound is tight, independent of data access mode, and independent of optimization procedure.

Setup. Fix a task with causal variables (X, Y), intervention domain $\mathfrak{E}_{\text{int}}$, and context space $\mathcal{C}$. For a parametric causal model M satisfying the alignment condition of Theorem 1, the task-level interventional mapping $\psi_M : \mathfrak{E}_{\text{int}} \times \mathcal{C} \to \mathcal{Y}$ is well-defined by

$$\psi_M(x, c) = \mathbb{E}_M[Y \mid \text{do}(X = x), c].$$

The restriction to parametric $\mathcal{H}$ is adopted for convenience and ensures the standard regularity conditions under which $\Psi_{\mathcal{H}} := \{\psi_M : M \in \mathcal{H}\}$ is closed in the relevant function space.

Let μ be a probability measure on $\mathfrak{E}_{\text{int}} \times \mathcal{C}$ with support covering the intervention–context pairs on which competence is evaluated. Equip $F := L^2(\mu)$ with its standard norm and define the interventional metric

$$d_{\text{int}}(M, M') := \|\psi_M - \psi_{M'}\|_{L^2(\mu)}.$$

The measure μ encodes the evaluation regime: distinct μ induce distinct geometries. Choices of μ concentrated on interventions where $\mathcal{H}$ happens to be adequate yield small realizability gaps; choices reflecting genuine deployment demands yield gaps that measure the system's exposure to $\mathcal{H}$-inadequacy. We state results at the level of μ and note where the choice matters.

Definition 4 (Realizability Gap)**.** The *realizability gap* of $\mathcal{H}$ relative to M^* under μ is

$$\Delta(M^*, \mathcal{H}; \mu) := \inf_{M \in \mathcal{H}} \|\psi_M - \psi_{M^*}\|_{L^2(\mu)}.$$

Dependence on μ is suppressed where unambiguous.

Theorem 6 (Access-Mode Independence of the Realizability Gap)**.** *Let $\mathcal{H}$ be a parametric class of causal models satisfying the alignment condition of Theorem 1, and let M^* denote the true causal model. For any protocol Π that receives observational data, interventional queries, or any combination thereof, and outputs a model $M_\Pi \in \mathcal{H}$,*

$$\mathbb{E}\big[\|\psi_{M_\Pi} - \psi_{M^*}\|_{L^2(\mu)}\big] \;\geq\; \Delta(M^*, \mathcal{H}; \mu).$$

No protocol restricted to outputs in $\mathcal{H}$ can achieve expected interventional error below the realizability gap, regardless of data access mode.

Proof. $M_\Pi \in \mathcal{H}$ almost surely, so Definition 4 gives $\|\psi_{M_\Pi} - \psi_{M^*}\|_{L^2(\mu)} \geq \Delta(M^*, \mathcal{H}; \mu)$ almost surely. Taking expectations preserves the inequality. □

Remark 4 (Convex case and decomposition). When $\Psi_{\mathcal{H}}$ is convex and closed in $L^2(\mu)$, the metric projection of ψ_{M^*} onto $\Psi_{\mathcal{H}}$ is the unique element minimizing distance. Denote it $\psi_{\mathcal{H}^*}$. Since $L^2(\mu)$ is a Hilbert space, the projection satisfies the orthogonality relation

$$\langle \psi_{M^*} - \psi_{\mathcal{H}^*},\, \psi_M - \psi_{\mathcal{H}^*} \rangle \;\leq\; 0 \qquad \text{for all } M \in \mathcal{H},$$

yielding the decomposition

$$\|\psi_M - \psi_{M^*}\|^2 \;\geq\; \|\psi_M - \psi_{\mathcal{H}^*}\|^2 + \|\psi_{\mathcal{H}^*} - \psi_{M^*}\|^2,$$

with equality iff $\psi_M = \psi_{\mathcal{H}^*}$. The first term on the right is the within-$\mathcal{H}$ suboptimality of M; the second is $\Delta(M^*, \mathcal{H})^2$. This parallels the approximation–estimation decomposition in statistical learning theory. For causal-model classes, $\Psi_{\mathcal{H}}$ is typically non-convex, and the decomposition need not hold; the bound in Theorem 6(a) is unaffected.

Corollary 7 (Access Modes Relative to the Gap)**.** *Within $\mathcal{H}$:*

(i) Observational protocols cannot discriminate among members of $\mathcal{E}(P_{\text{obs}}) \cap \mathcal{H}$. Observationally consistent models can differ in d_{int} by up to the d_{int}-diameter of $\mathcal{E}(P_{\text{obs}}) \cap \mathcal{H}$, and this ambiguity cannot be reduced by additional observational data. Worst-case interventional error of an observational protocol therefore has a component uncontrolled by observational signal, in addition to the realizability gap $\Delta(M^*, \mathcal{H})$.

(ii) Criticism protocols in the sense of Proposition 13 identify $\psi_{\mathcal{H}^*}$ in finitely many interventions, reducing within-$\mathcal{H}$ suboptimality to zero. Total interventional error equals $\Delta(M^*, \mathcal{H})$.

(iii) No protocol restricted to outputs in $\mathcal{H}$ can achieve total interventional error below $\Delta(M^*, \mathcal{H})$.

Corollary 8 (Expansion Necessity)**.** *Achieving interventional error below $\Delta(M^*, \mathcal{H})$ requires enlarging the hypothesis class to some $\mathcal{H}' \supset \mathcal{H}$ with $\Delta(M^*, \mathcal{H}') < \Delta(M^*, \mathcal{H})$. No selection procedure operating within a fixed $\mathcal{H}$ achieves this.*

Theorem 6 separates two distinct sources of interventional failure. Theorem 3 concerns within-$\mathcal{H}$ suboptimality: observational learning cannot reach $\psi_{\mathcal{H}^*}$. The realizability gap concerns residual error that remains when $\psi_{\mathcal{H}^*}$ is reached exactly. The two are governed by different structures, and neither subsumes the other.

Criticism protocols address the first. They identify $\psi_{\mathcal{H}^*}$ by exposing disagreement among $\mathcal{H}$-members under intervention. They do not address the second: disagreement among $\mathcal{H}$-members is bounded by the diameter of $\Psi_{\mathcal{H}}$ and carries no information about the distance from $\Psi_{\mathcal{H}}$ to ψ_{M^*}. A refuted $\mathcal{H}$-member is not ψ_{M^*}; that fact does not indicate whether any remaining $\mathcal{H}$-member is.

The within-$\mathcal{H}$ problem and the gap-reduction problem differ in kind. Within-$\mathcal{H}$ identification is a search problem: the target lies in a specified space and the task is to locate it. Classical complexity theory applies. Gap reduction is a generation problem: the target lies outside the specified space and the task is to construct candidate extensions. The hypothesis space is not fixed at problem-definition time, and no classical complexity class characterizes the resulting problem as far as we are aware. The question is whether an algorithm can, from interventional refutations alone, propose extensions to $\mathcal{H}$ that reduce $\Delta(M^*, \mathcal{H})$. This is a structural capacity, not an efficiency property. A protocol either has the means to generate genuinely new hypotheses from refutation signals or it does not; compute does not substitute for the absence of this capacity.

Conjecture-and-criticism combines within-$\mathcal{H}$ refutation with structural proposals for expansion. Whether such proposals can be reliably generated from interventional refutation signals, and what structural information $\psi_{M^*} - \psi_{\mathcal{H}^*}$ carries about the missing content of $\mathcal{H}$, is the content of the remaining open question. Theorem 6 establishes only that without such expansion, interventional competence is bounded by $\Delta(M^*, \mathcal{H})$, and no protocol operating within $\mathcal{H}$ can do better.

5 Structural Interventional Failure in Large Language Models

The parametric mechanism framework makes the impossibility results apply to large language models as a direct consequence rather than informal analogy. This section develops the connection explicitly.

5.1 Computational structure

Large language models generate sequences autoregressively. At each position t, a token X_t is produced as a function of the preceding context and stochastic noise,

$$X_t = \mathcal{M}_t(X_{<t}; \theta, U_t),$$

where U_t denotes exogenous randomness and the parameters θ are shared across positions. Unrolling inference over time yields a directed chain $X_1 \to X_2 \to \cdots \to X_T$, so that perturbations or interventions on earlier tokens propagate forward.

Each mapping $\mathcal{M}_t$ is a parametric mechanism in the sense of Definition 1: a parameterized input–output mapping whose behavior under input perturbation is well-defined when parameters are held fixed. We identify the token context $X_{<t}$ with the parent variables of the mechanism at position t, reflecting the model's computational dependency structure rather than any claim about task-level causal structure.

Under this interpretation, the alignment condition of Theorem 1 generically fails. Attention makes each token position sensitive to the full preceding context, including elements that are not causal parents of the task-relevant output. Post-intervention behavior therefore depends on task-irrelevant internal features, violating the X-fiber condition: two internal states that agree on the intervention target can produce different outputs because they differ in causally irrelevant context. Theorems 3, 4, and 5 then apply at the computational level. The observational learning objective does not enforce that these mechanisms correspond to autonomous causal processes or depend only on variables that are causal parents of task-relevant outputs. Interventions on task-irrelevant context can alter behavior, consistent with locality violations in the LAP sense. In language models, such context includes prompt formatting and stylistic cues; in agentic systems, it includes task-irrelevant features of the environment state, observation ordering, or interaction history that carries no causal information about the target variable.

Autonomy may fail as well due to parameter sharing. Mechanisms at different positions are coupled through a shared parameter vector, so deviations introduced at one position are evaluated by downstream mechanisms with the same internal parameterization. Under such coupling, observational gradients can increase interventional error and amplify downstream deviations under composition.

The application of Theorem 5 to the LLM setting requires care on three points. The violation magnitudes $\varepsilon^{A,i}_{\text{aut}}$ and $\varepsilon^{A,i}_{\text{loc}}$ are evaluated pointwise via Lie derivatives and vary with context across token positions rather than remaining uniform as the theorem assumes. The persistence and contraction factors $p_{\min}$ and β' are likewise assumed constant across steps, but in a system whose mechanisms are conditioned on the full preceding sequence these

factors inherit the same context-dependence. The additive decomposition of deviations across origination times requires contributions from distinct steps to interact weakly, a condition that full-context attention does not support in general. These considerations limit the quantitative precision of the amplification bound in the LLM setting. They do not, however, provide grounds for expecting that LAP-related errors diminish along the token sequence rather than accumulate. Theorem 3 establishes that observational training actively produces violations at each position, and no mechanism within the observational regime counteracts their forward propagation. The qualitative conclusion of accumulation therefore stands even when the assumptions of Theorem 5 are not jointly satisfied with precision.

5.2 Interaction structure and error propagation

The token-level chain analyzed above is interrupted in practice by inputs originating outside the LLM's parametric mechanism. Any such input, whether a human turn, a tool return, or an environment observation in an agentic setting, constitutes a genuine $\mathrm{do}(X_t)$ intervention: it sets the value of a position in the sequence exogenously rather than through the LLM's learned mechanism. These exogenous interventions partition the full token sequence into segments, bounding within-segment accumulation by resetting the mechanism chain at each boundary.

This reset operates at the mechanistic level rather than at the level of content. If a segment accumulates errors due to LAP violations, those errors shape the substance of what passes to the next exogenous input, whether that is the human's understanding of the problem, a query constructed for a tool, or a state passed to a downstream agent. The subsequent segment therefore begins from a starting point that is itself downstream of prior accumulation. Across segment boundaries, errors propagate through the content of exogenous inputs in exactly the structural sense that Theorem 5 describes within segments, producing a higher-order accumulation process at the semantic level. The segment boundary resets the mechanism chain without resetting the error state carried in the content.

This two-level structure connects the token-level applicability of Theorem 5 to the semantic-level analysis developed below. Within each segment, LAP violations accumulate through the parametric mechanism chain. Across segments, semantically corrupted content propagates errors forward through exogenous inputs despite mechanistic resets, with segments playing the role of tokens and semantic errors playing the role of local deviations in the amplification structure.

How frequently and reliably these resets occur depends on the structure of deployment. In turn-based interaction, human turns provide exogenous resets at conversational boundaries, bounding within-turn accumulation while leaving the higher-order cross-segment process unconstrained by anything other than the quality of human evaluation. In tool-augmented settings, tool returns provide additional resets that may occur at higher frequency and at more precisely targeted points in the reasoning chain, and tool outputs may introduce genuinely corrective information rather than merely partitioning the sequence. In agentic settings without human oversight, exogenous interventions arise from environment interactions alone, and the frequency and quality of those interactions in checking accumulated semantic error determines how severely the higher-order process compounds across the full task horizon.

The higher-order propagation across segments is especially consequential when outputs

function as causal claims, assertions about which worldly variables influence which others, rather than merely as text. Training corpora contain descriptions of causal structure produced by agents with partial and imperfect causal knowledge, but from the model's perspective this text is observational data like any other, and learning to predict text about causation is not equivalent to learning from interventions on the causal structures being described. Computational locality and autonomy violations therefore propagate to the semantic level not incidentally but through the same gradient dynamics that Theorem 3 identifies, shaping internal representations without regard to whether they support intervention-stable causal content, as formalized in Appendix C (Proposition 9).

6 Scope and the Sufficiency Question

The preceding results are structural, not universal. This section identifies the conditions under which they apply and the minimal additions that suffice to overcome them.

6.1 When does LAP hold approximately?

Many empirical domains exhibit approximate locality and autonomy across practically relevant intervention ranges, even when strict LAP fails at some level of description. In engineering and physical systems whose transition dynamics are fixed and invariant, interventions correspond to well-defined, repeatable changes in the data-generating process. In such settings, interaction provides genuine interventional signal rather than merely additional data, and the gradient obstruction of Theorem 3 does not arise.

Systems such as AlphaZero [18] operate within this regime. Although trained by gradient-based optimization, they repeatedly encounter genuine interventions under a single invariant causal law. The transition dynamics take the form $s_{t+1} = T(s_t, a_t)$, where actions are treated as exogenous inputs to a stable transition law. Incorrect action–outcome hypotheses are refutable through experience, which renders interventional error visible to the learning dynamics.

Closed-world interaction does not guarantee mechanistically faithful internal representations. Behavior may converge to match the environment's interventional distribution without the system encoding an explicit or unique causal abstraction. What such interaction provides is access to genuinely interventional signal under a stable causal law, which removes the gradient obstruction.

These conditions do not extend to open-world settings. Outside closed environments, there is no single invariant transition law: states are partially observed, multiple mechanisms interact, and action effects vary across contexts. Unless feedback introduces invariant intervention semantics and refutability of incorrect causal hypotheses, the gradient obstruction persists.

6.2 When does the gradient obstruction not activate?

The obstruction in Theorem 3 requires three conditions to activate: (i) a LAP violation with uniformly nonzero magnitude, (ii) observational correlations that reward the spurious

dependence, and (iii) the system is trained by observational loss minimization. If any of these conditions fails, the theorem does not apply. In particular, if the model happens to satisfy locality and autonomy then observational gradients are not necessarily misaligned with interventional error.

Similarly, the amplification in Theorem 5 requires the effective persistence factor $p_{\min}\beta'$ to be bounded away from zero. If early deviations are rapidly absorbed (e.g., through architectural bottlenecks or explicit correction mechanisms), cumulative error remains bounded independently of sequence length.

6.3 Active criticism and disagreement-maximizing interventions

This subsection identifies the minimal additional structure that suffices to overcome the limits established above. Distinct causal models may agree on the observational distribution while differing in their responses to intervention. A learner restricted to passive data cannot distinguish them even in the infinite-sample limit (formalized in Appendix A). Resolving this underdetermination requires information not contained in P_{obs}.

Rather than framing intervention selection in terms of expected information gain, we adopt a criticism-based perspective aligned with elimination rather than confirmation. At any time t, let $\mathcal{C}_t$ denote the set of hypotheses consistent with all data observed so far. A disagreement-maximizing policy selects interventions that maximize worst-case divergence among candidates:

$$\Delta(e; \mathcal{C}_t) = \sup_{\mathcal{M}_1, \mathcal{M}_2 \in \mathcal{C}_t} D(P_{\mathcal{M}_1}(X \mid e) \| P_{\mathcal{M}_2}(X \mid e)).$$

Under mild separability assumptions, iteratively applying disagreement-maximizing interventions eliminates false hypotheses in finitely many steps (Appendix D, Proposition 13).

The role of active criticism can be understood epistemically as introducing commitment and refutation into the learning process. Commitment consists in treating a specific causal hypothesis as governing interventional behavior. Refutation consists in exposing that commitment to interventions under which it may fail. This introduces an essential asymmetry among observationally equivalent hypotheses: once a commitment is made, error becomes discoverable under intervention. Purely observational learning lacks this operation. It can eliminate hypotheses incompatible with the observed distribution, but cannot privilege one hypothesis over another within an observational equivalence class.

Exogenous versus endogenous criticism. The sufficiency result above applies to the exogenous setting. Competing hypotheses produce different predictions under intervention, and the true interventional distribution adjudicates among them.

The endogenous setting is different. Theorem 4 shows that L&A violations cause policy substitutions π_A, π'_A that should be interventionally equivalent to produce different outputs for non-descendants of A. The ambiguity is internal to the model's representation, not a mismatch with the world. No external target at the model's abstraction level is available to adjudicate.

Criticism retains a role, but a weaker one. Applying π_A and π'_A and observing $X_i^{\mathcal{M}^{\pi_A}} \neq X_i^{\mathcal{M}^{\pi'_A}}$ detects the L&A violation but does not correct it.

Correction requires conditions criticism alone does not supply. Either enforce L&A architecturally, so inconsistent outputs cannot arise, or reintroduce external signal through environment feedback or constraints on the admissible policy class. The first route eliminates the violation; the second supplies the adjudication the endogenous setting lacks.

Criticism is therefore sufficient for interventional identification in the exogenous case and necessary but insufficient in the endogenous case. Conjecture 1 should be read accordingly: a refutation-exposing process is a prerequisite, but in the endogenous setting it must be combined with architectural or external conditions that make refutation actionable.

The semantic level. The same structure applies at the semantic level developed in Appendix C. A system that generates causal claims is implicitly committed to the factoring conjecture $\sigma = \iota \circ \pi_Y$: that worldly semantic content is fully determined by task-level output. This commitment is exactly the kind that disagreement-maximizing criticism can expose. Two causal hypotheses that agree on the observational distribution and on task-level outputs may nonetheless differ in the worldly content their outputs express under intervention—and a disagreement-maximizing policy that selects interventions maximizing divergence among consistent hypotheses will, when applied at the semantic level, expose precisely those cases where the factoring conjecture fails. Observational training provides no analogue of this operation. It can eliminate factoring commitments incompatible with the observed distribution, but cannot distinguish among the many factoring conjectures that survive observational consistency, including those that would be immediately refuted by a single well-chosen intervention.

6.4 Conjecture: sufficiency of LAP-satisfying structure

The impossibility results establish what observational learning *cannot* achieve. We complement them with a conjecture about what *would* be sufficient.

Conjecture 1 (Interventional Competence)**.** *Interventional competence requires that (i) the target domain admits intervention semantics satisfying the Locality–Autonomy Principle, and (ii) the system participates in a process capable of detecting and revising failures of the factoring conjecture $\sigma = \iota \circ \pi_Y$—that is, a process that exposes the system's implicit semantic commitments to genuine interventional refutation risk, rather than merely encoding a causal abstraction that satisfies locality and autonomy statically.*

This conjecture is not a premise of any theorem in this paper; it is an interpretive claim about what reliable interventional competence must ultimately require. It is not enough that a valid causal abstraction exist; the system must participate in a process by which failures of its semantic commitments become detectable and revisable under interventional pressure. Observational learning alone cannot provide this, as established by the main results. Active error correction of the kind described in Section 6.3 can, because disagreement-maximizing criticism is precisely the mechanism that exposes factoring commitments to refutation.

The conjecture is falsifiable in two directions. It would be refuted by demonstrating sustained interventional competence in a system that provably participates in no process capable of exposing its semantic commitments to refutation—showing that static encoding

suffices after all. It would be supported by showing that systems lacking such a process systematically fail under novel interventions in ways that correlate with the violation magnitudes $\varepsilon^{A,i}_{\text{aut}}$ and $\varepsilon^{A,i}_{\text{loc}}$ identified in Theorem 3. The second direction connects the conjecture to empirical investigation in a way the impossibility results themselves, being structural, cannot.

References

[1] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[2] James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, 2003.

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[4] Bernhard Schölkopf et al. Toward causal representation learning. *Proceedings of the IEEE*, 109:612–634, 2021.

[5] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021.

[6] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78:947–1012, 2016.

[7] Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 372–407. PMLR, 2023.

[8] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 13557–13603. PMLR, 2022.

[9] Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 38319–38331, 2022.

[10] Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael C Mozer, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

[11] Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*, 2017.

[12] Sander Beckers and Joseph Y. Halpern. Abstracting causal models. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, volume 33, pages 2678–2685, 2019.

[13] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[14] Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–63, 2025.

[15] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On Pearl's hierarchy and the foundations of causal inference. In Hector Geffner, Rina Dechter, and Joseph Y. Halpern, editors, *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 507–556. ACM Books, 2022.

[16] Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

[17] Sander Beckers, Frederick Eberhardt, and Joseph Y. Halpern. Approximate causal abstraction. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, 2019.

[18] David Silver et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362:1140–1144, 2018.

[19] Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, pages 220–227, 1990.

[20] Steen A Andersson, David Madigan, and Michael D Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics*, 25(1):505–541, 1997.

[21] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

[22] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

[23] Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012.

[24] Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments sufficient and in the limit necessary to identify all causal relations among N variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 178–184, 2005.

[25] Tom M. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2):203–226, 1982. doi: 10.1016/0004-3702(82)90040-6.

[26] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory (COLT)*, pages 287–294. ACM, 1992. doi: 10.1145/130385.130417.

[27] Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959. doi: 10.1214/aoms/1177706205.

[28] Simon Tong and Daphne Koller. Active learning for structure in Bayesian networks. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 863–869, 2001.

[29] Yang-Bo He and Zhi Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9(84):2523–2547, 2008.

A Extended Background: Conceptual Foundations of Observational and Interventional Learning

This section collects conceptual material underlying the distinction between observational and interventional competence used throughout the main text. Its purpose is to clarify definitions, scope, and status of the structural principles employed, without expanding the claims or introducing additional results.

A.1 Causality, mechanisms, and counterfactual constraints

Causality is invoked in this work as a structural commitment about the organization of the world: that there exist law-like relations among variables such that interventions on one variable impose counterfactual constraints on others. A purely statistical description characterizes how variables co-vary under a fixed data-generating process; a causal description characterizes how outcomes would change under hypothetical manipulations.

Formally, we treat causality as the claim that for certain variables (X, Y), there exist mappings that remain stable under interventions on X, in the sense that the distribution of Y under $\mathrm{do}(X{=}x)$ is governed by the same underlying mechanism across admissible contexts. These stability requirements distinguish causal relations from mere correlations.

Causal claims are inherently counterfactual. They concern how a system would respond were a variable set to a value not necessarily realized in the observed data. This counterfactual character gives rise to the central difficulty: observational data alone may be insufficient to constrain counterfactual behavior.

To express causal hypotheses formally, we make use of *parametric causal mechanisms*: parameterized mappings from inputs and exogenous factors to outputs that are intended to represent autonomous causal processes in the world. Such mechanisms may be instantiated by learned models, but their causal adequacy depends on whether they respect the invariances implied by intervention, rather than on their predictive accuracy under observation.

A.2 Pearl, Woodward, and endogenous interventions

Both Pearl's structural causal models and Woodward's interventionist theory define causation using interventions that are specified externally to the system being modeled. In Pearl's framework, interventions are formal regime changes represented by the $\mathrm{do}(\cdot)$ operator; in Woodward's framework, interventions are idealized manipulations whose independence from the system under study is required to ground causal interpretation. Neither framework provides a general account of systems that must internally generate and reason about interventions as part of their operation.

For Pearl, fixing intervention semantics prior to inference or learning allows structural causal models to maintain clean counterfactual semantics and support powerful results concerning identifiability, transportability, and causal effect estimation.

Woodward's interventionist account similarly relies on interventions external to the system under study, though for philosophical reasons. Interventions serve as idealized probes of causal structure: to test whether X causes Y, one considers whether Y would change under

suitable interventions on X that break X's usual causes while leaving other mechanisms intact. This requirement of independence is essential to Woodward's semantic project. Interventions must not be causally entangled with the mechanisms they are intended to test, or causal interpretation collapses. As a result, Woodward intentionally excludes interventions that arise endogenously from feedback, policy choice, or adaptive behavior within the same system.

The Locality–Autonomy Principle (LAP) together with the theorems presented in this work are designed to address this gap in the AI context while largely adhering to the same philosophical view of causation. Our goal is not to offer a new account of causal semantics, but a principle characterizing when intervention semantics can be realized by a learning system. We extend beyond both frameworks by identifying the representational and architectural conditions under which intervention semantics can be realized endogenously, and by characterizing the structural failure modes that arise when those conditions are violated. Whereas exogenous violations create interventional errors that observational gradients cannot reach, endogenous violations induce implementation dependence in the system's realization of interventional semantics, preventing interventional targets from being represented in an invariant manner without additional structural constraints.

A.3 Formal definitions of competence

The main text distinguishes observational competence from interventional competence. We record here the full formal definitions.

Definition 5 (Observational Competence)**.** A model f is observationally competent relative to P_{obs} if it minimizes expected loss under the observational distribution:

$$f \in \arg\min_{g} \; \mathbb{E}_{(x,y)\sim P_{\text{obs}}}\big[L(g(x), y)\big].$$

Observational competence concerns adequacy with respect to a fixed empirical joint distribution, whether obtained through passive observation or through interaction that does not alter the underlying intervention semantics. It may be silent about behavior under interventions not realized in that distribution.

Definition 6 (Interventional Competence (Exogenous; domain-relative))**.** Fix a task with causal variables (X, Y), an observational training distribution P_{obs}, and a class $\mathfrak{E}$ of admissible exogenous interventions interpreted in the classical sense: applying $e \in \mathfrak{E}$ surgically replaces the generating mechanism for X while holding all other mechanisms fixed.

A model f supports exogenous interventional competence relative to a nontrivial intervention domain $\mathfrak{E}_{\text{int}} \subseteq \mathfrak{E}$ if:

1. **Beyond-support reach.** $\mathfrak{E}_{\text{int}}$ contains interventions not realized in the observational regime.
2. **Correct interventional behavior on the domain.** For all $e = \text{do}(X{=}x) \in \mathfrak{E}_{\text{int}}$ and admissible contexts c,
$$f(x, c) = \mathbb{E}[Y \mid \text{do}(X{=}x), c].$$
3. **Invariance to nuisance context.** If two contexts c, c' are interventionally equivalent for the task on $\mathfrak{E}_{\text{int}}$, then $f(x, c) = f(x, c')$ for all interventions in the domain.

This definition makes competence explicitly domain-relative: it is a claim about correct and invariant behavior over a specified nontrivial set of interventions, not an absolute guarantee. It excludes accidental correctness and memorization by requiring correctness beyond the observational regime together with invariance to task-irrelevant context variation.

Definition 7 (Interventional Competence (Endogenous; agentic and domain-relative))**.** Fix a task with causal variables (X, Y) and consider systems in which the mechanism that sets X is itself generated and varied within the system. Let Π be a class of admissible intervention-generating mechanisms for X, understood as alternatives the system can represent as counterfactual substitutions. For each $\pi \in \Pi$, let $\mathcal{M}^{\pi}$ denote the counterfactual system obtained by replacing the original X-generating mechanism with π while leaving all other mechanisms fixed.

A model supports endogenous interventional competence relative to a nontrivial policy domain $\Pi_{\text{int}} \subseteq \Pi$ if, for all $\pi \in \Pi_{\text{int}}$ and admissible contexts c,

$$f(\pi, c) = \mathbb{E}[Y \mid \mathcal{M}^{\pi}, c],$$

and the model is invariant to changes in components of c that are irrelevant to Y under that substitution.

Endogenous interventional competence strengthens the exogenous notion: classical do(X=x) operations can be viewed as degenerate substitutions in which the X-generator is replaced by a constant mechanism. In the endogenous regime, locality and autonomy are not merely sufficient conditions for good generalization; they are what make the relevant counterfactual semantics well-defined.

A.4 Underdetermination by observation and observational equivalence

A central obstacle to interventional learning from passive data is *underdetermination by observation*. Distinct causal models may induce the same observational distribution while differing in their responses to intervention.

Definition 8 (Observational Equivalence Class)**.** Let $\mathcal{H}$ be a class of causal models and P_{obs} an observational distribution. The *observational equivalence class* of P_{obs} within $\mathcal{H}$ is

$$\mathcal{E}(P_{\text{obs}}) := \{\mathcal{M} \in \mathcal{H} : P_{\mathcal{M}}(X) = P_{\text{obs}}(X)\}.$$

Members of $\mathcal{E}(P_{\text{obs}})$ are indistinguishable on the basis of observational data alone, even in the infinite-sample limit. When $|\mathcal{E}(P_{\text{obs}})| > 1$ and members differ in their interventional distributions, observational data cannot uniquely determine interventional behavior.

This underdetermination applies equally to empirical risk minimization and Bayesian posterior concentration with fixed inductive bias, unless that bias collapses the equivalence class to a singleton.

B Structural and Differential Framework

This section develops the full structural and differential framework underlying the main text. It formalizes the notions of locality, autonomy, and mechanism independence stated informally in the main paper.

B.1 Locality and Autonomy

In smooth parametric systems, LAP violations admit a convenient analytic characterization. Interventions or learning updates can be idealized as vector fields generating flows in state or parameter space, and failures of locality or autonomy are witnessed by non-vanishing Lie derivatives of a mechanism $\mathcal{M}_i$ along vector fields corresponding to interventions on other mechanisms. When $\mathcal{M}_i$ is a scalar-valued function, these Lie derivatives reduce to directional derivatives; when $\mathcal{M}_i$ is itself a vector field, they reduce to Lie brackets (commutators).

The Lie-derivative formulation is a diagnostic witness of LAP violations in differentiable systems, not a definition of LAP.

Locality (state-modularity). Locality requires that the mechanism governing variable X_i depend only on the states of its causal parents. Interventions on non-parent variables may change the values of variables throughout the system, but they must not deform the mechanism governing X_i.

Let $W_j := \partial/\partial X_j$ denote the infinitesimal intervention that perturbs state variable X_j. Locality requires invariance of $\mathcal{M}_i$ under all non-parent interventions:

$$\mathcal{L}_{W_j}\mathcal{M}_i = 0 \qquad (j \notin \mathrm{PA}(i)).$$

If $\mathcal{M}_i$ is represented as a scalar function, this reduces to $\partial_{X_j}\mathcal{M}_i = 0$. If $\mathcal{M}_i$ is represented as a vector field, it is equivalent to the commutation condition $[\mathcal{M}_i, W_j] = 0$.

Autonomy (mechanism-modularity). Autonomy requires that each causal mechanism be independently modifiable. Interventions that replace the mechanism governing variable A may alter the values of downstream variables (since their inputs change), but they must not deform any other mechanism.

Let Ξ_A denote the vector field on parameter space generated by interventions on the mechanism-defining parameters θ_A. Autonomy requires invariance of all other mechanisms under these parameter interventions:

$$\mathcal{L}_{\Xi_A}\mathcal{M}_i = 0 \qquad (i \neq A).$$

When $\mathcal{M}_i$ is represented as a vector field, this is equivalent to $[\mathcal{M}_i, \Xi_A] = 0$.

Locality restricts sensitivity to non-parent state interventions. Autonomy restricts sensitivity to mechanism interventions on any other variable. Together, these invariance conditions ensure that interventions can be interpreted as localized modifications without unintended global effects on other mechanisms.

C Semantic Structure, Error Correction, and Causal-Claim Generation

This section addresses systems that generate causal claims or assertions: linguistic or symbolic outputs that purport to describe interventional relationships among worldly variables.

C.1 The gap between internal task variables and worldly causal structure

The machinery developed so far operates entirely within the model's own variable framework. Projections $\pi_X : Z \to \mathcal{X}$ and $\pi_Y : Z \to \mathcal{Y}$ map internal states to task variables, and the X-fiber condition of Theorem 22 demands that post-intervention output distributions be constant across internal states agreeing on those variables. This is a purely internal condition: it concerns coherence within the model's own causal abstraction, not correspondence between that abstraction and the world.

For systems whose outputs are interpreted as claims about worldly causal structure, a further question arises. Even a system satisfying the X-fiber condition perfectly may generate causal claims that bear no reliable relationship to genuine causal organization in the world, if the task variables $\mathcal{X}$ and $\mathcal{Y}$ do not themselves track real causal structure. Observational training does not in general provide pressure toward such correspondence. The training objective rewards predictive accuracy over P_{obs}, which constrains internal representations to reflect statistical regularities in the data rather than the causal organization of the processes that generated it.

This gap motivates a second level of analysis. We introduce a *semantic projection* $\sigma : Z \to \mathcal{W}$, where $\mathcal{W}$ is a space of worldly states, to formalize the relationship between internal states and the worldly content that outputs are interpreted as describing. The role of σ is not to introduce new causal structure but to represent the interpretation of model outputs as claims about the world. The key point, developed below, is that σ is not an independent formalism but is constrained to factor through the existing π machinery. Semantic correctness then places demands on this factoring that the X-fiber condition is necessary but not sufficient to satisfy.

C.2 Implemented versus represented causal knowledge

A system may have causal knowledge in two senses. In the first, the system's input-output behavior accords with some causal structure, but this structure is not explicitly encoded in any separable internal representation. The causal regularities are *implemented* but not *represented*. In the second, the system maintains an explicit, inspectable model of causal structure that can be queried, compared against alternatives, and revised.

The key criterion is whether the system's causal commitments can be locally revised in response to detected errors, or whether correction requires retraining wholesale. Systems with merely implemented causal knowledge have no locus for targeted correction. Systems with represented causal knowledge admit, in principle, separation between the representational

vehicle and the causal content, enabling errors to be localized, attributed, and repaired without global disruption.

C.3 Computational and semantic L&A

We use "locality and autonomy" (L&A) for model-level assessments and reserve "LAP" exclusively for the domain-level structural principle.

A system satisfies *computational L&A* if its internal mechanisms satisfy locality and autonomy with respect to the system's own computational state variables and parameters. The X-fiber condition of Theorem 22 formalizes the alignment requirement that computational L&A is intended to support: post-intervention outcome distributions must be constant across internal states sharing the same projected task variable value.

Semantic L&A requires a further condition.

A system that produces outputs interpreted as causal claims or assertions is implicitly committed to a factoring conjecture: that worldly semantic content is fully determined by task-level output, so that there exists an interpretation map $\iota : \mathcal{Y} \to \mathcal{W}$ such that $\sigma = \iota \circ \pi_Y$. We treat this not as a stipulated condition that holds in some well-behaved regime, but as a conjecture in the Popperian sense—a commitment the system makes, implicitly, that is in principle exposed to refutation through interventional probing, but for which observational training provides no refutation mechanism. A system whose outputs change under interventions on causally irrelevant context has thereby refuted its own factoring conjecture, because semantic content has been shown to depend on internal degrees of freedom not captured by π_Y. The significance of this framing is that the factoring conjecture need not be established as true in order for the paper's impossibility results to apply: it suffices that the system is committed to it, that the commitment is the kind that can in principle be refuted by intervention, and that observational training provides no process by which such refutation can occur. When the factoring holds as a conjecture that has survived interventional pressure, semantic content is fully determined by the model's task-level output, and the X-fiber condition on π_Y is sufficient to guarantee that semantic content is invariant across internal states agreeing on X. Semantic locality then reduces to the same fiber-invariance requirement already expressed by Theorem 22, applied at the level of worldly content rather than task-variable output. That the factoring has survived interventional pressure does not guarantee correspondence with true causal structure; it guarantees only internal consistency under the interventions so far applied.

When the factoring fails, semantic content depends on internal degrees of freedom not captured by π_Y. Two internal states z_1, z_2 may agree on $\pi_Y(z_1) = \pi_Y(z_2)$ while yielding different worldly content under σ. This is a semantic locality violation: the causal claim generated by the system depends on internal features that are not determined by the task output, and therefore not constrained by the X-fiber condition.

Stated in the language of Theorem 22, semantic locality requires σ to be constant across X-fibers: for all $z_1, z_2 \in F_x$,

$$\sigma(I_x(z_1)) = \sigma(I_x(z_2)).$$

In smooth systems, failure of this condition is witnessed by non-vanishing of $\mathcal{L}_v(\sigma \circ I_x)$ along directions $v \in \ker(d\pi_X)$, the same fiber-tangent directions already identified in Theorem 22

Level	What it is	Can be wrong?	Role of L&A / LAP
World / Domain	True causal structure and intervention semantics	N/A	LAP is a structural property of the domain
Semantic	Model's interpreted causal structure about the world	Yes	Must respect domain LAP; L&A required for interventional competence
Computational	Internal states, parameters, and computation	Yes	L&A may or may not hold; not constrained by LAP

Table 1: Distinction between world, semantic (causal abstraction), and computational levels.

as the locus of alignment failure. The logical relationships are: Interventional competence $\Rightarrow$ semantic L&A, but semantic L&A $\not\Rightarrow$ computational L&A. Nevertheless,

$$\neg\,\text{computational L\&A} \xRightarrow[\text{generically under learning}]{} \neg\,\text{semantic L\&A},$$

indicating a typical dynamical failure rather than a logical entailment.

C.4 Generic propagation of computational violations to the semantic level

When the factoring $\sigma = \iota \circ \pi_Y$ holds and π_Y satisfies the X-fiber condition, semantic L&A is guaranteed. Computational L&A violations undermine both requirements simultaneously: they introduce sensitivity along directions in $\ker(d\pi_X)$, and the same spurious sensitivities that cause the X-fiber condition to fail for π_Y cause it to fail for $\sigma \circ I_x$ as well. The following corollary makes this precise; it is a direct consequence of Corollaries 23 and 24.

Corollary 9 (Structural Instability of Semantic L&A)**.** *Let a system's computational dynamics violate locality or autonomy at the level of internal mechanisms. By Corollary 23, this implies that the X-fiber condition fails for $(\pi_Y)_* P_{\mathcal{M}}(\cdot \mid I_x(\cdot))$. Maintaining semantic L&A despite this failure would require the factoring map ι to compensate: specifically, $\sigma \circ I_x$ would need to be invariant across X-fibers even though $(\pi_Y)_* P_{\mathcal{M}}(\cdot \mid I_x(\cdot))$ is not. This requires fine-tuned alignment between the kernel of $d\iota$ and the subspace of computational sensitivities induced by learning. Absent architectural constraints enforcing L&A or an active process of error detection and correction, this alignment is not structurally stable. Semantic L&A violations therefore arise generically whenever computational L&A is violated, and by Corollary 24, observational learning actively deepens rather than corrects this misalignment.*

Remark 5 (Measure-zero alignment and generic failure). Let $\mathcal{V}_{\text{irr}} \subset T_z Z$ denote the subspace of perturbation directions at an internal state z that influence computational dynamics but lie in $\ker(d\sigma)$, corresponding to no worldly content. Semantic L&A requires $\mathrm{d}\sigma(v) = 0$ for all $v \in \mathcal{V}_{\text{irr}}$. Absent architectural constraints, this invariance selects a lower-dimensional subset of

admissible factoring maps within the space of smooth maps $Z \to \mathcal{W}$. Small perturbations of the learned representation generically break this invariance. Semantic L&A in the presence of computational L&A violations therefore holds only on a measure-zero set of representations.

C.5 Error correction and the maintenance of semantic L&A

The generic propagation argument establishes that semantic L&A cannot be passively maintained in systems with computational L&A violations. This leaves open the possibility of active maintenance through error correction.

Active maintenance requires preserving the factoring condition $\sigma = \iota \circ \pi_Y$ under intervention-induced change. In differential-geometric terms, this corresponds to a form of parallel transport: as the system moves along an intervention-induced path in internal state space, its causal commitments are adjusted so that σ continues to factor through π_Y appropriately.

For such active correction to be possible, several capacities are required. First, the system must represent its causal commitments in a form that admits inspection and local revision. If causal knowledge is merely implemented, distributed across weights with no separable structure, there is no locus for targeted correction. Second, the system must detect when its causal claims depend on features they should not depend on, which requires counterfactual probes that vary putatively irrelevant features and check whether semantic content changes. Third, once a violation is detected, the system must modify the offending commitment without inducing widespread collateral changes, a requirement that parallels the autonomy condition at the level of correction rather than mechanism.

Human cognition may instantiate an approximate version of this process. Human causal representations are incomplete, frequently wrong, and neurally distributed in ways that almost certainly violate computational L&A. Yet humans exhibit a degree of interventional competence that purely observational learners cannot match. A plausible explanation is that human interventional competence is maintained through ongoing error discovery and correction. Causal conjectures that fail under intervention are attributed to specific commitments and revised with some degree of locality. This does not require computational modularity; it requires only that the revision process operate at the level of represented commitments rather than distributed weights.

C.6 Large language models as causal-claim generators

When prompted appropriately, large language models produce outputs that function as causal claims. The relationship between these outputs and worldly causal structure is mediated by a proxy chain: the world has causal structure; humans form partial representations of that structure; humans produce text expressing those representations; this text enters the training corpus as observational data. Each link introduces potential for slippage. What the model learns is not causal structure per se, but patterns in how humans talk about causal structure.

The failure mode is two-layered. At the computational level, LLMs violate L&A in both senses developed in the main text. Attention mechanisms allow each token position to depend on the full context, including features that are not causal parents of the worldly content being described. Parameter sharing across positions and domains means that mechanisms generating causal claims about different topics cannot be independently modified. As discussed

in Section 5, these violations are contextually distributed rather than uniform: violation magnitudes $\varepsilon_{\text{aut}}^{A,i}$ and $\varepsilon_{\text{loc}}^{A,i}$ vary across token positions and contexts. Where the conditions of Theorems 15 and 17 are met, the theorems apply: observational training creates interventional errors that the training objective cannot reach.

At the semantic level, the factoring condition $\sigma = \iota \circ \pi_Y$ faces additional pressure. Even if an LLM's internal task variables were perfectly aligned in the sense of Theorem 22, there is no guarantee that those task variables correspond to genuine worldly causal organization. The training objective rewards prediction of next tokens, not correspondence with the causal structure of the processes generating those tokens. Where computational L&A violations occur, Corollary 9 characterizes how they propagate to the semantic level: causal claims across domains are coupled through shared parameters, so updates in one domain may alter claims in another, and perturbations in early context positions propagate through the autoregressive chain in ways that do not track worldly relevance.

LLMs also lack the capacities identified as necessary for error correction. Causal knowledge in LLMs is implemented rather than represented; there is no explicit causal model that can be inspected, queried, or locally revised. The model does not engage in self-directed interventional probing during inference, and the global entanglement of parameters precludes localized revision.

D Complete Proofs

This section contains full proofs of all formal results stated in the main text. Theorems are restated for completeness.

D.1 Passive Bias Insufficiency for Causal Identification

Let $\mathcal{M}^* \in \mathcal{H}$ denote the true model, and assume $\mathcal{M}^* \in \mathcal{B}$. Define the observationally consistent subset of the bias as

$$\mathcal{A} := \mathcal{B} \cap \mathcal{E}(P_{\text{obs}})$$

where $\mathcal{E}(P_{\text{obs}})$ is the observational equivalence class (Section A.4, Definition 8).

Theorem 10 (Passive Bias Insufficiency)**.** *Assume $|\mathcal{A}| > 1$, and that there exist distinct $\mathcal{M}', \mathcal{M}'' \in \mathcal{A}$ and an intervention $e \in \mathfrak{E}$ such that*

$$P_{\mathcal{M}'}(X \mid e) \neq P_{\mathcal{M}''}(X \mid e).$$

Then no passive learner restricted to $\mathcal{B}$ can uniquely identify the interventional behavior of $\mathcal{M}^*$ from observational data alone.

Proof. Both $\mathcal{M}'$ and $\mathcal{M}''$ lie in $\mathcal{A}$, so $P_{\mathcal{M}'}(X) = P_{\mathcal{M}''}(X) = P_{\text{obs}}$. A passive learner trained on samples from P_{obs} assigns identical likelihood to both hypotheses and cannot distinguish them. By assumption, the two models disagree under intervention e, so their interventional behaviors differ. The learner therefore cannot identify the interventional behavior of $\mathcal{M}^*$. □

Theorem 11 (Existence of an Identifying Intervention Sequence)**.** *Assume $|\mathcal{A}| < \infty$, and that for every distinct pair $\mathcal{M}_1, \mathcal{M}_2 \in \mathcal{A}$ there exists $e \in \mathfrak{E}$ with $P_{\mathcal{M}_1}(X \mid e) \neq P_{\mathcal{M}_2}(X \mid e)$. Assume further that interventional distributions are observed without noise. Then there exists a finite sequence of interventions, of length at most $|\mathcal{A}| - 1$, whose induced distributions identify $\mathcal{M}^*$ within $\mathcal{B}$.*

Proof. Proceed by induction on $|\mathcal{A}|$. If $|\mathcal{A}| = 1$, the claim holds with the empty sequence. Suppose $|\mathcal{A}| = k > 1$. Pick any $\mathcal{M} \in \mathcal{A}$ with $\mathcal{M} \neq \mathcal{M}^*$. By pairwise separability, some $e \in \mathfrak{E}$ satisfies $P_{\mathcal{M}}(X \mid e) \neq P_{\mathcal{M}^*}(X \mid e)$. Observing the true interventional distribution under e eliminates $\mathcal{M}$, yielding an updated candidate set of size at most $k - 1$. Apply the inductive hypothesis. The total length is at most $|\mathcal{A}| - 1$. □

Corollary 12 (Extension to Structural Causal Models)**.** *Let $\mathcal{H}$ be the class of SCMs over endogenous variables $(X_1, \ldots, X_n)$, with interventions defined by $\text{do}(X_S = x_S)$ for arbitrary subsets S. If there exist multiple observationally equivalent SCMs differing under at least one intervention, then the conclusions of Theorem 10 apply unchanged. If in addition the pairwise separability and finiteness conditions of Theorem 11 hold, so does its conclusion.*

Remark 6 (Relationship to existing underdetermination results). The impossibility claim in Theorem 10 relates to classical results on Markov equivalence [19, 20] and to the underdetermination arguments in [21, 22]. Those results establish that observational data identifies causal structure only up to a Markov equivalence class, under causal Markov and

faithfulness assumptions over DAGs. Theorem 10 differs in three respects: it assumes no graph structure, imposes neither faithfulness nor causal Markov conditions, and states the result for an abstract model class $\mathcal{H}$ with an explicit inductive bias $\mathcal{B}$. The existence claim in Theorem 11 relates to [23, 24], which characterize the number of interventions needed to orient all edges in a DAG. The present bound of $|\mathcal{A}| - 1$ is loose by comparison, reflecting the absence of graph structure: without a combinatorial object to exploit, each intervention is credited with eliminating only a single hypothesis.

D.2 Sufficiency of Disagreement-Maximizing Criticism

Theorem 11 establishes that some finite intervention sequence identifies $\mathcal{M}^*$ within $\mathcal{B}$. It does not specify how to choose one. This subsection gives a constructive criterion.

Let $\mathcal{C}_t \subseteq \mathcal{A}$ denote the candidate set at time t, with $\mathcal{C}_0 = \mathcal{A}$. For an intervention $e \in \mathfrak{E}$, define the disagreement measure

$$\Delta(e; \mathcal{C}_t) := \big|\{\, P_{\mathcal{M}}(X \mid e) : \mathcal{M} \in \mathcal{C}_t \,\}\big| - 1,$$

the number of distinct interventional distributions induced by $\mathcal{C}_t$ under e, minus one. A *disagreement-maximizing policy* selects, at each step, an intervention $e_t \in \arg\max_{e \in \mathfrak{E}} \Delta(e; \mathcal{C}_t)$.

Proposition 13 (Sufficiency of Disagreement-Maximizing Criticism)**.** *Assume the hypotheses of Theorem 11: $|\mathcal{A}| < \infty$, pairwise interventional separability on $\mathcal{A}$, and noise-free observation of interventional distributions. Then any disagreement-maximizing policy identifies $\mathcal{M}^*$ within $\mathcal{B}$ in at most $|\mathcal{A}| - 1$ steps.*

Proof. By Theorem 11, the hypotheses guarantee that some identifying sequence exists. It remains to show that the disagreement-maximizing policy realizes this.

If $|\mathcal{C}_t| > 1$, pick any distinct $\mathcal{M}_1, \mathcal{M}_2 \in \mathcal{C}_t$. By pairwise separability, some $e \in \mathfrak{E}$ satisfies $P_{\mathcal{M}_1}(X \mid e) \neq P_{\mathcal{M}_2}(X \mid e)$, so $\Delta(e; \mathcal{C}_t) \geq 1$. The policy selects such an intervention. Observing the true interventional distribution under e_t eliminates every $\mathcal{M} \in \mathcal{C}_t$ whose induced distribution under e_t disagrees with the observed one. Since $\Delta(e_t; \mathcal{C}_t) \geq 1$, at least one hypothesis is eliminated, giving $|\mathcal{C}_{t+1}| \leq |\mathcal{C}_t| - 1$.

The candidate set strictly shrinks at each step and remains a subset of $\mathcal{A}$. Termination occurs when $|\mathcal{C}_t| = 1$, at which point the sole remaining hypothesis is $\mathcal{M}^*$. The bound $|\mathcal{A}| - 1$ follows. □

Remark 7 (On the choice of disagreement measure). The measure Δ counts distinct interventional distributions. Alternative measures, such as worst-case or average pairwise divergence between candidates, yield policies with the same termination guarantee under the stated hypotheses. The guarantee depends only on the policy selecting an intervention that separates at least one pair whenever $|\mathcal{C}_t| > 1$. Finer measures matter when observations are noisy or when the cost of interventions varies, neither of which is modeled here.

Remark 8 (Relationship to existing active learning and sequential design literature). Proposition 13 restates a standard elimination guarantee in the setting of this paper. The mechanism—maintain a candidate set, select queries that induce disagreement among candidates, eliminate hypotheses inconsistent with observations—appears in the version-space framework of [25]

and in the query-by-committee algorithm of [26]. Sequential selection of experiments to distinguish hypotheses under a finite candidate set dates to [27]. The termination bound $|\mathcal{A}| - 1$ under noise-free observation follows from the trivial elimination argument common to this line of work.

In the causal setting, [28] applied active learning to Bayesian network structure discovery via interventions, and [29] developed intervention selection strategies that minimize the size of the interventional Markov equivalence class. [23] gave optimal active learning strategies under graph structure. These results operate on DAGs over a finite variable set, from which finiteness of the hypothesis space follows automatically.

Proposition 13 differs in its setting rather than its mechanism. The candidate set $\mathcal{A}$ is an abstract subset of an inductive bias $\mathcal{B}$, not a Markov equivalence class. Without graph structure, finiteness of $\mathcal{A}$ does not follow from any ambient combinatorial object and must be stated directly. The disagreement measure Δ correspondingly counts distinct interventional distributions rather than exploiting graph-theoretic features of the candidate set. The bound $|\mathcal{A}| - 1$ is loose compared to the $O(\log n)$ and $O(n)$ bounds available under graph structure [23, 24], reflecting the cost of the abstraction: each intervention is credited with eliminating only one hypothesis because no combinatorial structure licenses stronger progress. The contribution here is conceptual rather than algorithmic—the proposition shows that the standard elimination guarantee transfers to the graph-free setting established in Theorem 10.

D.3 Causal Mechanism Necessity (Theorem 1 of main text)

Theorem 14 (Task-General Applicability of Causal Mechanisms)**.** *Let (X, Y) be variables in a causal system, and let $\mathfrak{E} = \{\mathrm{do}(X{=}x) : x \in \mathcal{D}\}$ be a family of interventions indexed by a nontrivial domain $\mathcal{D} \subseteq \mathcal{X}$ with $|\mathcal{D}| \geq 2$.*

Suppose a model f is interventionally competent on a task involving these interventions, in the sense of Definition 6. Then there exists a mapping $\psi : \mathcal{D} \times \mathcal{C} \to \mathcal{Y}$ such that: (a) for all $x \in \mathcal{D}$ and $c \in \mathcal{C}$, the model's output is $f(x, c) = \psi(x, c)$; (b) ψ is intervention-invariant: for any two protocols π_1, π_2 both resulting in $X = x$, $\psi^{\pi_1}(x, c) = \psi^{\pi_2}(x, c) = \psi(x, c)$.

Proof. Step 1: Existence. By assumption, for each $x \in \mathcal{D}$ and $c \in \mathcal{C}$, the model produces $f(x, c) = \mathbb{E}[Y \mid \mathrm{do}(X{=}x), c]$. Define $\psi(x, c) := f(x, c)$. This is well-defined because f is a function.

Step 2: Intervention-invariance. Suppose toward contradiction that ψ depends on the protocol. Let π_1 and π_2 be two protocols both setting $X = x$. If $\psi^{\pi_1}(x, c) \neq \psi^{\pi_2}(x, c)$, the model's output depends on which protocol was used. But $P(Y \mid \mathrm{do}(X{=}x), c)$ is defined by the causal semantics and depends only on (x, c), so the model fails to match the unique interventional distribution under at least one protocol, contradicting interventional competence.

Step 3: Necessity of internal representation. The mapping ψ takes cause (x) and context (c) as inputs, produces effect (y) as output, and is invariant to the method by which the cause was established. This is the defining property of a causal mechanism in the interventionist sense. □

Remark 9 (Architectural Independence). The theorem places no constraints on how ψ is implemented. The requirement is functional (intervention-invariant input-output behavior), not architectural.

D.4 Autonomy Violations Create Unreachable Interventional Errors

Theorem 15 (Autonomy Violations Create Unreachable Interventional Errors)**.** *Consider a causal system with true graph $G = (V, E)$ and a model with parametric mechanisms $\{\mathcal{M}_i : i \in V\}$. Fix a variable $A \in V$ and a distinct variable $i \neq A$. Under the true causal semantics, interventions on $\mathcal{M}_A$ do not alter $\mathcal{M}_i$.*

Assume:

(i) *$\mathcal{M}_i$ depends spuriously on a parameter block θ_A associated with $\mathcal{M}_A$.*

(ii) *$\mathcal{M}_i$ is C^1 in θ_A with uniformly continuous derivative.*

(iii) (Directional violation magnitude) *There exist a unit vector $\hat{v}$ in parameter space and a scalar $\varepsilon^{A,i}_{\text{aut}}(\hat{v}) > 0$ such that*

$$\inf_{(x,u),\,\theta_A} \left|\langle \partial_{\theta_A} \mathcal{M}_i(x, u; \theta_A),\, \hat{v}\rangle\right| \;\geq\; \varepsilon^{A,i}_{\text{aut}}(\hat{v}).$$

(iv) $\mathrm{Cov}(X_A, X_i) \neq 0$ *with consistent sign across the support.*

(v) *Squared-error observational loss.*

(vi) (Causal baseline correctness) *There exists a reference value $\theta_A = 0$ at which the mechanism $\mathcal{M}_i$ is autonomous with respect to $\mathcal{M}_A$ and coincides with a causally correct baseline. In particular,*

$$\varepsilon^{A,i}_{\text{int}}(\theta; a)\big|_{\theta_A=0} = 0 \quad \textit{for all } a.$$

No assumption is made that $\theta_A = 0$ is observationally optimal.

(vii) (Local observational pressure along $\hat{v}$) *The directional derivative of the observational loss along $\hat{v}$ is strictly negative at baseline:*

$$D_{\hat{v}} L_{\text{obs}}(\theta)\big|_{\theta_A=0} < 0.$$

Observational training locally rewards movement away from the causally correct baseline in the direction $\hat{v}$.

Define the interventional error for node i under $\mathrm{do}(X_A = a)$ *as*

$$\varepsilon^{A,i}_{\text{int}}(\theta; a) := \mathbb{E}\left[\left|X_i^{\text{model}}(a; \theta) - X_i^{\text{true}}(a)\right|\right].$$

Then, restricting attention to parameter displacements along $\hat{v}$ (i.e., $\theta_A = \phi\hat{v}$ for $\phi \in \mathbb{R}$):

(a) For all sufficiently small $|\phi|$ and all a in a compact intervention domain, interventional error grows at least linearly in $|\phi|$:

$$\varepsilon_{\text{int}}^{A,i}(\phi\hat{v}; a) \;\geq\; |\phi| \cdot \varepsilon_{\text{aut}}^{A,i}(\hat{v}).$$

(b) The observational and interventional gradients along $\hat{v}$ are strictly misaligned:

$$\partial_\phi L_{\text{obs}}(\phi\hat{v}) \cdot \partial_\phi \varepsilon_{\text{int}}^{A,i}(\phi\hat{v}; a) \;<\; 0.$$

Gradient descent on observational loss therefore has a component along $\hat{v}$ that strictly increases interventional error.

Proof. We proceed in two steps: first the interventional error bound, then the directional gradient misalignment.

Step 1: Interventional error bound. Fix $(x_{\text{PA}(i)}, u) \in \mathcal{X}_{\text{PA}(i)} \times \mathcal{U}_i$. Under $\text{do}(X_A = a)$, the true mechanism is unchanged by autonomy in the true system:

$$X_i^{\text{true}}(a) = \mathcal{M}_i^{\text{true}}(x_{\text{PA}(i)}, u).$$

The model output is

$$X_i^{\text{model}}(a; \phi\hat{v}) = \mathcal{M}_i(x_{\text{PA}(i)}, u; \phi\hat{v}).$$

By assumption (vi), at $\phi = 0$,

$$\mathcal{M}_i(x_{\text{PA}(i)}, u; 0) = \mathcal{M}_i^{\text{true}}(x_{\text{PA}(i)}, u).$$

Parametrize the segment $\theta_A(t) = t\phi\hat{v}$ for $t \in [0, 1]$ and define

$$g(t) := \mathcal{M}_i(x_{\text{PA}(i)}, u; t\phi\hat{v}).$$

By assumption (ii), g is C^1 on $[0, 1]$ with derivative

$$g'(t) = \left\langle \partial_{\theta_A} \mathcal{M}_i(x_{\text{PA}(i)}, u; t\phi\hat{v}),\, \phi\hat{v} \right\rangle.$$

By assumption (iii),

$$|g'(t)| \;\geq\; |\phi| \cdot \varepsilon_{\text{aut}}^{A,i}(\hat{v}) \;>\; 0 \qquad \text{for all } t \in [0, 1],$$

and g' is continuous and non-vanishing, hence of constant sign. The fundamental theorem of calculus gives

$$|g(1) - g(0)| = \int_0^1 |g'(t)|\, dt \;\geq\; |\phi| \cdot \varepsilon_{\text{aut}}^{A,i}(\hat{v}).$$

Since $g(1) - g(0) = \mathcal{M}_i(\cdot; \phi\hat{v}) - \mathcal{M}_i^{\text{true}}$, we obtain

$$\left| X_i^{\text{model}}(a; \phi\hat{v}) - X_i^{\text{true}}(a) \right| \;\geq\; |\phi| \cdot \varepsilon_{\text{aut}}^{A,i}(\hat{v}).$$

Taking expectations establishes part (a).

Step 2: Directional gradient misalignment along $\hat{v}$. By assumption (vii),

$$\partial_\phi L_{\text{obs}}(\phi\hat{v})\big|_{\phi=0} < 0.$$

L_{obs} is differentiable, so $\partial_\phi L_{\text{obs}}$ is continuous in a neighborhood of 0. There exists $\delta > 0$ such that for all $\phi \in (0, \delta)$,

$$\partial_\phi L_{\text{obs}}(\phi\hat{v}) < 0.$$

By part (a), for $\phi > 0$,

$$\varepsilon^{A,i}_{\text{int}}(\phi\hat{v}; a) \;\geq\; \phi \cdot \varepsilon^{A,i}_{\text{aut}}(\hat{v}),$$

so $\varepsilon^{A,i}_{\text{int}}$ is strictly increasing to first order along $\hat{v}$. Hence for all sufficiently small $\phi > 0$,

$$\partial_\phi \varepsilon^{A,i}_{\text{int}}(\phi\hat{v}; a) > 0.$$

Combining,

$$\partial_\phi L_{\text{obs}}(\phi\hat{v}) \cdot \partial_\phi \varepsilon^{A,i}_{\text{int}}(\phi\hat{v}; a) < 0 \qquad \text{for sufficiently small } \phi > 0,$$

establishing part (b). □

Corollary 16 (Autonomy Violations Induce Persistent Context Sensitivity)**.** *Violations of autonomy force downstream mechanisms to remain sensitive to upstream deviations, yielding a persistence factor $p_{\min}\beta'$ bounded away from zero in Theorem 19. Interventional errors created by autonomy violations are therefore amplified under sequential composition.*

D.5 Locality Violations Create Unreachable Interventional Errors

Theorem 17 (Locality Violations Create Unreachable Interventional Errors)**.** *Consider a causal system with true graph $G = (V, E)$ and a model with parametric mechanisms $\{\mathcal{M}_i : i \in V\}$. Fix $A \in V$ and $i \notin \text{Desc}(A)$.*

Assume:

(i) *$\mathcal{M}_i$ depends spuriously on X_A, with $\mathcal{M}_i : \mathcal{X}_{\text{PA}(i)} \times \mathcal{X}_A \times \mathcal{U}_i \to \mathcal{X}_i$.*

(ii) *$\mathcal{M}_i$ is C^1 in X_A with uniformly continuous derivative.*

(iii) (Directional violation magnitude) *There exist a unit vector $\hat{w} \in \mathcal{X}_A$ and a scalar $\varepsilon^{A,i}_{\text{loc}}(\hat{w}) > 0$ such that*

$$\inf_{(x_{\text{PA}(i)}, x_A, u)} \big|\langle \partial_{X_A} \mathcal{M}_i(x_{\text{PA}(i)}, x_A, u),\, \hat{w}\rangle\big| \;\geq\; \varepsilon^{A,i}_{\text{loc}}(\hat{w}).$$

(iv) $\text{Cov}(X_A, X_i) \neq 0$ *with consistent sign.*

(v) *Squared-error observational loss.*

(vi) (Causal baseline correctness at a_0) *There exists a reference value $a_0 \in \mathcal{X}_A$ such that the interventional error vanishes at a_0:*

$$\varepsilon^{A,i}_{\text{int}}(a_0) = 0.$$

(vii) (Local observational pressure coupled to $\hat{w}$) *There exists a reference parameter value θ_i^0 (at which the causal baseline of (vi) holds) and a direction $v \in T_{\theta_i^0}\Theta_i$ such that the directional derivative of the observational loss along v is strictly negative at baseline:*

$$D_v L_{\text{obs}}(\theta_i)\big|_{\theta_i=\theta_i^0} < 0,$$

and moving along v increases the directional sensitivity of $\mathcal{M}_i$ to X_A along $\hat{w}$:

$$\frac{d}{d\phi}\Big|_{\phi=0} \big|\langle \partial_{X_A}\mathcal{M}_i(\theta_i^0 + \phi v),\, \hat{w}\rangle\big| \;>\; 0.$$

Then, restricting attention to intervention displacements along $\hat{w}$ (i.e., $a = a_0 + s\hat{w}$ for $s \in \mathbb{R}$):

(a) For all s with $|s|$ in a compact range,

$$\varepsilon_{\text{int}}^{A,i}(a_0 + s\hat{w}) \;\geq\; |s| \cdot \varepsilon_{\text{loc}}^{A,i}(\hat{w}).$$

(b) The observational and interventional gradients are misaligned along v:

$$(\nabla_{\theta_i} L_{\text{obs}} \cdot v) \cdot (\nabla_{\theta_i} \varepsilon_{\text{int}}^{A,i} \cdot v) \;<\; 0.$$

Proof. **Step 1: Interventional error bound.** Since $i \notin \text{Desc}(A)$, X_i^{true} is independent of a. The model output at $a = a_0 + s\hat{w}$ is

$$X_i^{\text{model}}(a) = \mathcal{M}_i(x_{\text{PA}(i)}, a_0 + s\hat{w}, u).$$

By assumption (vi), the interventional error vanishes at a_0.

Parametrize the segment $X_A(t) = a_0 + ts\hat{w}$ for $t \in [0,1]$ and define

$$h(t) := \mathcal{M}_i(x_{\text{PA}(i)}, a_0 + ts\hat{w}, u).$$

By assumption (ii), h is C^1 on $[0,1]$ with derivative

$$h'(t) = \big\langle \partial_{X_A}\mathcal{M}_i(x_{\text{PA}(i)}, a_0 + ts\hat{w}, u),\, s\hat{w}\big\rangle.$$

By assumption (iii),

$$|h'(t)| \;\geq\; |s| \cdot \varepsilon_{\text{loc}}^{A,i}(\hat{w}) \;>\; 0 \qquad \text{for all } t \in [0,1],$$

and h' is continuous and non-vanishing, hence of constant sign. The fundamental theorem of calculus gives

$$|h(1) - h(0)| = \int_0^1 |h'(t)|\, dt \;\geq\; |s| \cdot \varepsilon_{\text{loc}}^{A,i}(\hat{w}).$$

Since $h(1) - h(0) = \mathcal{M}_i(\cdot; a_0 + s\hat{w}) - \mathcal{M}_i(\cdot; a_0)$, taking expectations establishes part (a).

Step 2: Directional gradient misalignment along v.

Introduce a local coordinate ϕ along v with $\phi = 0$ at θ_i^0. By assumption (vii),

$$\partial_\phi L_{\text{obs}}(\phi)\big|_{\phi=0} < 0,$$

and by continuity there exists $\delta > 0$ such that $\partial_\phi L_{\text{obs}}(\phi) < 0$ for all $\phi \in (0, \delta)$.

Also by assumption (vii), moving along v increases the directional derivative of $\mathcal{M}_i$ along $\hat{w}$ in $\mathcal{X}_A$. Combined with part (a) applied at each ϕ, the interventional error along $\hat{w}$-displacements is strictly increasing along v for small $\phi > 0$:

$$\partial_\phi \varepsilon_{\text{int}}^{A,i}(a_0 + s\hat{w}; \theta_i^0 + \phi v) > 0.$$

Combining,

$$\partial_\phi L_{\text{obs}}(\phi) \cdot \partial_\phi \varepsilon_{\text{int}}^{A,i}(\cdot; \phi) < 0 \qquad \text{for sufficiently small } \phi > 0,$$

which establishes part (b). □

Remark 10 (Strength of the Directional Claim). The gradient misalignment in part (b) is directional along v in Θ_i combined with a corresponding direction $\hat{w}$ in $\mathcal{X}_A$. This does not preclude other directions along which gradient descent might reduce both losses. In the regime where the model is well-fit except for exploitation of spurious correlations, the locality-violation-increasing directions dominate the loss gradient. The directional claim suffices: observational optimization provides a pathway to increased interventional error that it actively rewards.

Corollary 18 (Locality Violations Induce Context Sensitivity)**.** *Spurious dependence on non-parent states forces downstream mechanisms to remain sensitive to upstream interventions, yielding a persistence factor bounded away from zero in Theorem 19.*

D.6 Error Amplification in Causal Chains (Theorem 3 of main text)

Theorem 19 (Cumulative Error Amplification in Causal Chains)**.** *Consider a sequence of parametric mechanisms $X_t = \mathcal{M}_t(X_{<t}, U_t)$, $t = 1, \ldots, T$, where each state space $\mathcal{X}_t \subset \mathbb{R}^{d_t}$ is compact and each mechanism is L-Lipschitz. Let X_t^* denote a reference trajectory.*

At each step t, the mechanism may introduce a fresh deviation $\delta_t \geq 0$, with $\delta_t = 0$ when no new error occurs. Define $\varepsilon_t := \mathbb{E}[\delta_t \mid \delta_t \neq 0]$.

Assume:

(i) Exogenous noise U_t is supported on a compact set.

(ii) Deviations lie in a small-error regime.

(iii) Origin-wise persistence: there exist $0 < \beta' \leq \beta < 1$ such that on the survival event $S_{t\to r+1}$,

$$\beta' \leq \frac{\alpha_{t\to r+1}}{\alpha_{t\to r}} \leq \beta, \qquad \alpha_{t\to t} = 1.$$

(iv) Lineage survival: $\Pr(S_{t\to r+1} \mid S_{t\to r}) \geq p_{\min}$ for all t, r.

(v) Additive decomposition: $d(X_s, X_s^*) = \sum_{t=1}^{s} \alpha_{t\to s}\delta_t$ *with nonnegative contributions.*

(vi) Conditional independence: $\mathbb{E}[\delta_t \mid \delta_t \neq 0, S_{t\to s}] = \mathbb{E}[\delta_t \mid \delta_t \neq 0] = \varepsilon_t$.

(vii) Nonnegativity: $\delta_t \geq 0$, $\alpha_{t\to s} \geq 0$ *a.s.*

Then

$$\mathbb{E}[D_{\text{total}}] \;\geq\; \sum_{t=1}^{T} \Pr(\delta_t \neq 0) \cdot \varepsilon_t \cdot \frac{1 - (p_{\min}\beta')^{T-t+1}}{1 - p_{\min}\beta'}. \tag{1}$$

Proof. Step 1: Decomposition into fresh contributions. By assumption (v), $D_{\text{total}} = \sum_{s=1}^{T} d(X_s, X_s^*) = \sum_{t=1}^{T}\sum_{s=t}^{T} \alpha_{t\to s}\delta_t$.

Step 2: Conditional propagation bounds. Fix origin time t with $\delta_t \neq 0$. On the survival event $S_{t\to s}$, by iterated application of assumption (iii): $\alpha_{t\to s} \geq (\beta')^{s-t}$.

Step 3: Probability of continued lineage influence. Define $S_{t\to s}$ as survival of the deviation lineage from t to s. By assumption (iv),

$$\Pr(S_{t\to s} \mid S_{t\to t}) = \prod_{r=t}^{s-1} \Pr(S_{t\to r+1} \mid S_{t\to r}) \geq (p_{\min})^{s-t}.$$

Step 4: Contribution of one fresh deviation. The contribution at time s from origin t is $C_{t\to s} := \alpha_{t\to s}\delta_t$. We have

$$\mathbb{E}[C_{t\to s}] \geq \Pr(\delta_t \neq 0) \cdot \Pr(S_{t\to s} \mid \delta_t \neq 0) \cdot (\beta')^{s-t} \cdot \varepsilon_t \geq \Pr(\delta_t \neq 0) \cdot (p_{\min}\beta')^{s-t} \cdot \varepsilon_t,$$

using Steps 2–3 and assumption (vi).

Step 5: Summation. Summing over $s \geq t$ and all t:

$$\mathbb{E}[D_{\text{total}}] \geq \sum_{t=1}^{T} \Pr(\delta_t \neq 0) \sum_{s=t}^{T} (p_{\min}\beta')^{s-t}\varepsilon_t = \sum_{t=1}^{T} \Pr(\delta_t \neq 0) \cdot \varepsilon_t \cdot \frac{1 - (p_{\min}\beta')^{T-t+1}}{1 - p_{\min}\beta'}.$$

□

Remark 11 (Regimes of Context Sensitivity). The effective persistence factor $p_{\min}\beta'$ distinguishes three regimes: (1) $p_{\min}\beta' \ll 1$: rapid decay, deviations remain localized; (2) $p_{\min}\beta' \lesssim 1$: slow decay, early deviations dominate cumulative error; (3) $p_{\min}\beta' \geq 1$: (excluded by assumptions) exponential blow-up. Autonomy or locality failures in sequential architectures may place systems in the second regime, where amplification arises from persistence rather than local instability.

Remark 12 (Single-Origination Bound Without Additivity). Assumption (v) (additive decomposition) can be dropped if the conclusion is weakened to a bound on the contribution of a single origination time. Fix any $t^* \in \{1, \ldots, T\}$. Without assuming that contributions from distinct origination times interact additively, Steps 2–4 of the proof apply to t^* alone: on the survival event $S_{t^*\to s}$, the propagated magnitude is bounded below by $(\beta')^{s-t^*}$ by assumption (iii), survival probability is bounded below by $(p_{\min})^{s-t^*}$ by assumption (iv), and

assumption (vi) gives $\mathbb{E}[\delta_{t^*} \mid \delta_{t^*} \neq 0, S_{t^* \to s}] = \varepsilon_{t^*}$. By nonnegativity (assumption (vii)), the contribution of the t^* lineage is a lower bound on D_{total}. Therefore, without assumption (v):

$$\mathbb{E}[D_{\text{total}}] \;\geq\; \Pr(\delta_{t^*} \neq 0) \cdot \varepsilon_{t^*} \cdot \frac{1 - (p_{\min}\beta')^{T-t^*+1}}{1 - p_{\min}\beta'}.$$

This bound grows geometrically in $T - t^*$ whenever $p_{\min}\beta'$ is bounded away from zero, establishing that a single early deviation can dominate cumulative error without any assumption about how multiple lineages interact.

The full multi-origination bound in (1) requires assumption (v), which fails in attention-based architectures: errors at distinct positions interact nonlinearly through the attention computation rather than propagating independently. The single-origination bound does not require additivity and is sufficient for the qualitative claim in Section 5 that early errors compound geometrically rather than dissipate. The full bound should not be applied to attention-based settings without further justification of assumption (v).

D.7 Autonomy Violations Destroy Endogenous Interventional Semantics

Theorem 20 (Autonomy Violations Preclude Endogenous Interventional Competence)**.** *Consider an endogenous SCM with mechanisms $\{\mathcal{M}_i : i \in V\}$ and admissible policy substitutions Π_A acting on variable $A \in V$. Fix $i \neq A$.*

Assume:

(i) *$\mathcal{M}_i$ depends spuriously on θ_A.*

(ii) *$\mathcal{M}_i$ is C^1 in θ_A with uniformly continuous derivative.*

(iii) (Directional violation magnitude) *There exist a unit vector $\hat{v}$ in parameter space and a scalar $\varepsilon^{A,i}_{\text{aut}}(\hat{v}) > 0$ such that*

$$\inf_{(x,u),\,\theta_A} \left|\langle \partial_{\theta_A} \mathcal{M}_i(x, u; \theta_A),\, \hat{v}\rangle\right| \;\geq\; \varepsilon^{A,i}_{\text{aut}}(\hat{v}).$$

(iv) (Policy richness) *Π_A contains a pair of policies π_A, π'_A with $\theta_A(\pi_A) - \theta_A(\pi'_A) = s\hat{v}$ for some $s \neq 0$.*

Then for any such pair,

$$\left|X_i^{\mathcal{M}^{\pi_A}} - X_i^{\mathcal{M}^{\pi'_A}}\right| \;\geq\; \varepsilon^{A,i}_{\text{aut}}(\hat{v}) \cdot \|\theta_A(\pi_A) - \theta_A(\pi'_A)\|.$$

Endogenous interventional competence with respect to Π_A is therefore impossible along the direction $\hat{v}$.

Proof. Fix $(x_{\text{PA}(i)}, u) \in \mathcal{X}_{\text{PA}(i)} \times \mathcal{U}_i$. Parametrize the segment between $\theta_A(\pi'_A)$ and $\theta_A(\pi_A)$ by

$$\theta_A(t) := \theta_A(\pi'_A) + ts\hat{v}, \qquad t \in [0, 1],$$

and define $g(t) := \mathcal{M}_i(x_{\mathrm{PA}(i)}, u; \theta_A(t))$. By assumption (ii), g is C^1 on $[0,1]$ with derivative

$$g'(t) = \big\langle \partial_{\theta_A} \mathcal{M}_i(x_{\mathrm{PA}(i)}, u; \theta_A(t)),\, s\hat{v} \big\rangle.$$

By assumption (iii),

$$|g'(t)| \;\geq\; |s| \cdot \varepsilon^{A,i}_{\mathrm{aut}}(\hat{v}) \;>\; 0 \qquad \text{for all } t \in [0,1],$$

and g' is continuous and non-vanishing on $[0,1]$, hence of constant sign by the intermediate value theorem. The fundamental theorem of calculus gives

$$|g(1) - g(0)| = \left| \int_0^1 g'(t)\, dt \right| = \int_0^1 |g'(t)|\, dt \;\geq\; |s| \cdot \varepsilon^{A,i}_{\mathrm{aut}}(\hat{v}) = \varepsilon^{A,i}_{\mathrm{aut}}(\hat{v}) \cdot \|\theta_A(\pi_A) - \theta_A(\pi'_A)\|.$$

Since $g(1) - g(0) = \mathcal{M}_i(\cdot; \theta_A(\pi_A)) - \mathcal{M}_i(\cdot; \theta_A(\pi'_A))$, the bound follows. □

D.8 Locality Violations Destroy Endogenous Interventional Semantics

Theorem 21 (Locality Violations Preclude Endogenous Interventional Competence)**.** *Consider an endogenous SCM with $i \notin \mathrm{Desc}(A)$ and admissible policy substitutions Π_A.*

Assume:

(i) $\mathcal{M}_i$ *depends spuriously on* X_A.

(ii) $\mathcal{M}_i$ *is* C^1 *in* X_A *with uniformly continuous derivative.*

(iii) (Directional violation magnitude) *There exist a unit vector* $\hat{w}$ *in* $\mathcal{X}_A$ *and a scalar* $\varepsilon^{A,i}_{\mathrm{loc}}(\hat{w}) > 0$ *such that*

$$\inf_{(x_{\mathrm{PA}(i)}, x_A, u)} \big| \langle \partial_{X_A} \mathcal{M}_i(x_{\mathrm{PA}(i)}, x_A, u),\, \hat{w} \rangle \big| \;\geq\; \varepsilon^{A,i}_{\mathrm{loc}}(\hat{w}).$$

(iv) (Policy richness) Π_A *contains a pair of policies* π_A, π'_A *with* $X_A^{\pi_A} - X_A^{\pi'_A} = r\hat{w}$ *for some* $r \neq 0$.

Then for any such pair,

$$\big| X_i^{\mathcal{M}^{\pi_A}} - X_i^{\mathcal{M}^{\pi'_A}} \big| \;\geq\; \varepsilon^{A,i}_{\mathrm{loc}}(\hat{w}) \cdot \|X_A^{\pi_A} - X_A^{\pi'_A}\|.$$

Endogenous interventional competence with respect to Π_A *is therefore impossible along the direction* $\hat{w}$.

Proof. Fix $(x_{\mathrm{PA}(i)}, u) \in \mathcal{X}_{\mathrm{PA}(i)} \times \mathcal{U}_i$. Parametrize the segment between $X_A^{\pi'_A}$ and $X_A^{\pi_A}$ by

$$X_A(t) := X_A^{\pi'_A} + tr\hat{w}, \qquad t \in [0,1],$$

and define $h(t) := \mathcal{M}_i(x_{\mathrm{PA}(i)}, X_A(t), u)$. By assumption (ii), h is C^1 on $[0,1]$ with derivative

$$h'(t) = \big\langle \partial_{X_A} \mathcal{M}_i(x_{\mathrm{PA}(i)}, X_A(t), u),\, r\hat{w} \big\rangle.$$

By assumption (iii),

$$|h'(t)| \;\geq\; |r| \cdot \varepsilon_{\text{loc}}^{A,i}(\hat{w}) \;>\; 0 \qquad \text{for all } t \in [0,1],$$

and h' is continuous and non-vanishing on $[0,1]$, hence of constant sign. The fundamental theorem of calculus gives

$$|h(1) - h(0)| = \int_0^1 |h'(t)|\, dt \;\geq\; |r| \cdot \varepsilon_{\text{loc}}^{A,i}(\hat{w}) = \varepsilon_{\text{loc}}^{A,i}(\hat{w}) \cdot \|X_A^{\pi_A} - X_A^{\pi'_A}\|.$$

Since $i \notin \text{Desc}(A)$, the true value of X_i is invariant under the policy substitution, so the discrepancy $h(1) - h(0)$ reflects an L&A violation rather than a legitimate intervention effect. □

D.9 Mechanism–Intervention Alignment

Theorem 22 (Mechanism–Intervention Alignment)**.** A system's internal dynamics define a coherent task-level interventional distribution if and only if states that are indistinguishable at the task level remain indistinguishable after intervention.

Formally, let a system implement a computational mechanism $\mathcal{M}$ over internal states Z, with projections $\pi_X : Z \to \mathcal{X}$ and $\pi_Y : Z \to \mathcal{Y}$. For each $x \in \mathcal{X}$, let $I_x : Z \to Z$ be an internal intervention map satisfying $\pi_X(I_x(z)) = x$. The set of all internal states sharing the same projected value x forms the X-fiber *over x:*

$$F_x := \{z \in Z : \pi_X(z) = x\}.$$

These are precisely the internal states that are task-level indistinguishable with respect to X: they may differ internally, but all correspond to the same external value.

The task-level interventional distribution $P(Y \mid \text{do}(X{=}x))$ is well-defined via the system's internal dynamics if and only if: for all $x \in \mathcal{X}$ and all $z_1, z_2 \in F_x$,

$$(\pi_Y)_* P_{\mathcal{M}}(\cdot \mid I_x(z_1)) = (\pi_Y)_* P_{\mathcal{M}}(\cdot \mid I_x(z_2)).$$

That is, post-intervention outcomes must be constant across each X-fiber: the outcome cannot depend on task-irrelevant internal degrees of freedom that distinguish states within the same fiber.

In smooth systems, failure of this condition is witnessed by non-vanishing of the Lie derivative $\mathcal{L}_v P_{\mathcal{M}}(\cdot \mid I_x(\cdot))$ along directions $v \in \ker(d\pi_X)$—the tangent directions that leave X unchanged but alter post-intervention outcomes. These are the local counterpart of the X-fiber: where the fiber is the global set of task-indistinguishable states, $\ker(d\pi_X)$ captures the infinitesimal directions along which that indistinguishability holds.

Proof. Necessity. Suppose $P(Y \mid \text{do}(X{=}x))$ is well-defined as a function of x alone. Fix any x and any $z_1, z_2 \in F_x$. Since $\pi_X(z_1) = \pi_X(z_2) = x$, both states lie in the same X-fiber. If $(\pi_Y)_* P_{\mathcal{M}}(\cdot \mid I_x(z_1)) \neq (\pi_Y)_* P_{\mathcal{M}}(\cdot \mid I_x(z_2))$, the post-intervention distribution of Y depends on which internal state within F_x was used, contradicting well-definedness of $P(Y \mid \text{do}(X{=}x))$ as a function of x alone.

Sufficiency. Suppose the X-fiber condition holds. For each x, define

$$P(Y \mid \mathrm{do}(X{=}x)) := (\pi_Y)_* P_{\mathcal{M}}(\cdot \mid I_x(z))$$

for any $z \in F_x$. The fiber condition guarantees this is independent of the choice of z, so the distribution is well-defined as a function of x alone.

Lie derivative witness. In smooth systems, suppose the fiber condition fails: there exist x and $z_1, z_2 \in F_x$ such that $(\pi_Y)_* P_{\mathcal{M}}(\cdot \mid I_x(z_1)) \neq (\pi_Y)_* P_{\mathcal{M}}(\cdot \mid I_x(z_2))$. Under mild regularity (specifically, path-connectedness of F_x), there exists a smooth path $\gamma : [0,1] \to F_x$ with $\gamma(0) = z_1$ and $\gamma(1) = z_2$. By the fundamental theorem of calculus, the map $t \mapsto (\pi_Y)_* P_{\mathcal{M}}(\cdot \mid I_x(\gamma(t)))$ must have nonzero derivative at some $t^* \in [0,1]$. The tangent vector $\dot{\gamma}(t^*) \in T_{\gamma(t^*)}Z$ satisfies $d\pi_X(\dot{\gamma}(t^*)) = 0$ since γ stays within F_x, so $\dot{\gamma}(t^*) \in \ker(d\pi_X)$. Setting $v = \dot{\gamma}(t^*)$ yields a direction in $\ker(d\pi_X)$ along which $\mathcal{L}_v P_{\mathcal{M}}(\cdot \mid I_x(\cdot))$ is nonzero. □

Corollary 23 (L&A Violations Imply Alignment Failure)**.** *Suppose the model exhibits an autonomy violation ($\varepsilon^{A,i}_{\mathrm{aut}} > 0$) or a locality violation ($\varepsilon^{A,i}_{\mathrm{loc}} > 0$) with respect to a pair (A, i) where $i \notin \mathrm{Desc}(A)$. Then the X-fiber condition of Theorem 22 fails: the task-level interventional distribution $P(Y \mid \mathrm{do}(X_A{=}a))$ is not well-defined via the system's internal dynamics.*

Proof. By Theorems 15 and 17, L&A violations imply nonzero uniform spurious dependence of $\mathcal{M}_i$ on θ_A or X_A respectively. In smooth systems, these manifest as non-vanishing Lie derivatives along directions that are task-irrelevant for X_i—directions lying in $\ker(d\pi_{X_i})$ by construction, since they do not alter the true causal parents of X_i. By the Lie derivative witness of Theorem 22, the X-fiber condition therefore fails, and the interventional distribution is not well-defined at the task level. □

Corollary 24 (Semantic Instability Under Observational Learning)**.** *Let a system's computational mechanisms violate L&A, and suppose the system is trained by observational loss minimization. Then the X-fiber condition of Theorem 22 fails generically, and semantic L&A is not maintained under observational learning dynamics.*

Proof. By Corollary 23, L&A violations imply alignment failure. By Theorems 15 and 17, observational gradient descent strictly increases the magnitude of L&A violations: it drives $\varepsilon^{A,i}_{\mathrm{aut}}$ and $\varepsilon^{A,i}_{\mathrm{loc}}$ upward rather than toward zero. Observational learning therefore actively deepens alignment failure rather than correcting it. The X-fiber condition is not satisfied at initialization (by assumption of L&A violation) and is driven further from satisfaction by gradient dynamics. Semantic L&A, which requires the X-fiber condition to hold at the semantic projection level (Section C), therefore fails generically under observational learning whenever computational L&A is violated. □

Remark 13 (Logical Chain). The results in this section form a connected logical chain. Theorems 15 and 17 establish that L&A violations create spurious sensitivities and that observational gradients deepen rather than correct them. Theorem 22 establishes that such sensitivities, when they lie along task-irrelevant directions, prevent the interventional distribution from being well-defined at the task level. Corollary 23 connects the two: L&A violations directly imply alignment failure. Corollary 24 completes the chain: under observational learning, alignment failure is not incidental but actively reinforced.