Neurosymbolic Artificial Intelligence
XX(X):1–36
©The Author(s) 2026
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Expressiveness, execution, and lifted recurrent closure in neurosymbolic intervention logic

Aliefya Mahalva Ayn[1]

Abstract

When can a neural state trained to answer logical intervention queries be reused after the system acts? A depth-one intervention query can express graph reachability, so shallow causal/modal syntax can demand deep local execution. For anonymous controls, a rank-preserving translation reduces depth-h query equivalence to first-order equivalence of response structures. Recurrent reuse is instead a bisimulation problem, decided by complete 1-types of the depth-$(h+1)$ response structure; for finite residual cores the decision needs only the next layer of hidden residuals and is independent of domain size. Ordinary anonymous closure need not imply stabilization, even for one known mechanism; six controls is minimal at positive horizon. For named monotone mechanisms, residual clutters give exact recurrent states. Five-seed experiments separate these effects: a local GNN hits the predicted receptive-field limit, and on a handwritten-digit task, replacing the depth-two query state by a bit state repairs long-horizon accuracy for both neural and symbolic updaters. Query adequacy and recurrent closure must therefore be tested separately.

Keywords

neurosymbolic artificial intelligence, first-order logic, modal logic, causal reasoning, expressiveness–computation trade-off, causal abstraction, recurrent representation

[1]Department of Computer Science, Rutgers University, New Brunswick, NJ, USA
ORCID iD: 0009-0001-2875-3376

Corresponding author:
Aliefya Mahalva Ayn, Department of Computer Science, Rutgers University, New Brunswick, NJ, USA.
Email: ama545@scarletmail.rutgers.edu

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

Introduction

Neurosymbolic AI inherits a tension familiar from logic and knowledge representation: richer languages make more distinctions available to a reasoner, but those distinctions need not be cheap to execute or compact to preserve inside a learned state. First-order quantification, modal transition structure, and causal intervention semantics stress different computational resources. A short causal/modal formula can require a long graph computation, and a representation that is sufficient for all queries up to depth h need not remain sufficient after the system acts. These are two forms of the same expressiveness–computation problem: what information does the logical language demand, and what state must a neural–symbolic architecture retain in order to execute that language repeatedly?

We study this question for a causal dynamic-logic view of interventions. The operator $[\mathrm{do}(\cdot)]$ is treated as an action modality whose semantics is mechanism replacement rather than arbitrary Kripke accessibility. The section "Shallow intervention syntax, deep causal execution" shows that shallow intervention syntax can hide deep causal execution: a depth-one intervention query realizes directed reachability, which is not first-order definable in the baseline edge signature on finite paths and requires message-passing depth proportional to graph distance under standard locality assumptions. This separates *logical depth* from *execution depth*.

The main finite-state theory then asks a different question. Suppose a neural interface is trained to answer every intervention query up to horizon h. May that same interface be fed back to a recurrent updater after an action? In general, no. Two histories can agree on every depth-h query while one more action sends them to states requiring different depth-h answers. This is the obstruction behind Moore/Myhill–Nerode refinement, L^* observation-table closure, predictive-state representations, MDP homomorphisms, and recurrent symbolic automata (Moore 1956; Nerode 1958; Angluin 1987; Littman et al. 2001; Ravindran and Barto 2003). Here we specialize it to causal intervention interfaces and connect it explicitly to do-commuting causal abstraction.

The paper is organized around three answers to the recurrent-reuse questions. *First, logic and lifting.* A reachability construction separates causal/modal syntax from graph execution, while a rank-preserving response-structure translation turns anonymous intervention equivalence into a first-order model-theory problem. This yields finite and countable query classifications and, crucially, a recurrent counterpart: anonymous reuse is a bisimulation problems decided by successor 1-types, with a finite-core one-hidden-layer decision procedure whose compilation is independent of ambient domain size, and a domain-size-uniform bounded-rank updater. *Second, constructive recurrent state.* Named monotone mechanisms admit residual-clutter states and a class-wide closure criterion; a finite-sample diagnostic tests whether a proposed quotient is transition-stable. Secondary quantitative extensions for real-valued responses and approximate transitions are collected in the supplement. *Third, neurosymbolic evidence.* Five-seed experiments connect the logical results to execution: a local GNN exhibits the predicted message-passing depth barrier, scalar probes separate exact quotient information from continuous encoder detail that survives even without query training, the Digits-grid

Table 1. Results at a glance. The conjunction with $m = 8$ is the running finite example in the experimental sections.

Result	Resource exposed	Neurosymbolic design consequence
Theorem 1 and the FO(TC) upper bracket	causal/modal syntax versus graph execution	shallow intervention syntax can require unbounded ordinary modal depth or long local message passing
Lemma 4 and Theorems 15, 14	quantifier rank and symmetry	anonymous query equivalence is a first-order response-structure problem; lifting is exact only under stated model-theoretic conditions
Theorems 30, 31, 34 and Corollary 32	anonymous actions and successor types	a lifted query state is recurrently reusable iff it is a bisimulation; finite cores expose the required hidden layer
Theorems 21 and 22	query sufficiency versus recurrent memory	a query quotient must be action-closed before it is fed back recurrently
Theorem 36 and the recurrent experiments	closure versus adequacy	test both transition closure and query fidelity; neither implies the other

run quantifies additional perception cost, and clutter learning under noisy perception separates adequacy from algebraic closure.

For a concrete thread through the paper, keep in mind the $m = 8$ conjunction. Its exact depth-two interface has 38 classes but its recurrent Moore machine has 256 states; this single example reappears in the structural lower bound, closure certificate, and neural ablations.

We do not set out to provide an unrestricted first-order theorem prover. The scope is more circumscribed: in our dynamic causal query language, quantifier rank and intervention depth are regarded as separate resources. When one works on countable pure domains, the query problem can be reduced to a question of elementary equivalence between response structures. Orbit equivalence is to be found in $\mathcal{L}_{\omega_1\omega}$, and finitarily so provided the finite-residual-core and homogeneous-threshold conditions are met, though this does not hold in general. Rather than a blanket classification of the countable case, the recurrent section puts forward finite-core decision results and sharp separations. In the matter of non-symmetric mechanisms, anonymous activation is nondeterministic for query classes; here the "Lifted recurrent closure under anonymous actions" has bisimulation in place of the deterministic Nerode congruence. A finite residual core will give a decision procedure and bounded rank a domain-size-uniform updater, but absent such cores type omission leads to a finitary failure that is quite definitive. As for stabilization and ordinary set-valued anonymous closure, they are not obliged to coincide at a positive horizon for a given mechanism, per Theorem 34(a). It is still an open question whether stabilization is compelled by type-determinism or by a multiplicity-sensitive graded closure

Related work and novelty boundaries

Intervention logics, causal abstraction, and action theories. Structural-equation interventions support formal languages and axiomatizations of causal counterfactual and intervention reasoning (Galles and Pearl 1998; Halpern 2000; Pearl 2009; Briggs

2012; Zhang 2013; Bareinboim et al. 2022); related work extends the logical setting to open universes, epistemic operators, and team or multiteam semantics (Ibeling and Icard 2020; Barbero et al. 2020; Barbero and Sandu 2021, 2024; Mossé et al. 2024). These works primarily ask which causal claim a language validates or how hard causal entailment is. Our question is orthogonal: after a query language induces an observational quotient, when does intervention execution descend to that quotient recurrently? Causal abstraction asks a closely related commuting-square question between model levels (Rubenstein et al. 2017; Beckers and Halpern 2019; Rischel and Weichwald 2021; Zennaro et al. 2023; Geiger et al. 2025); neural variants learn such maps from data (Xia et al. 2021; Xia and Bareinboim 2024). Situation-calculus work likewise uses refinement mappings and bisimulation between high- and low-level action theories (Banihashemi et al. 2017; Hofmann and Belle 2023). In Banihashemi et al., the refinement mapping is fixed as part of the abstraction relation; here the quotient is instead generated by the declared intervention-query family and may therefore fail recurrent closure. Lorenz and Tull's query-aware compositional abstractions provide a complementary categorical perspective (Lorenz and Tull 2026). Our deterministic named-action result specialize commuting abstraction to a query-induced quotient; the anonymous theory is its nondeterministic bisimulation counterpart.

Automata, predictive states, and sequential neurosymbolic systems. Moore/Myhill–Nerode refinement and L^* observation-table closure provide the classical finite-state background (Moore 1956; Nerode 1958; Angluin 1987). Predictive-state representations, MDP abstractions, state-abstraction taxonomies, bisimulation metrics, and reward machines all require action-stable state summaries (Littman et al. 2001; Ravindran and Barto 2003; Li et al. 2006; Givan et al. 2003; Ferns et al. 2004; Toro Icarte et al. 2022). NeSyA and DeepDFA combine neural perception with explicit automata (Manginas et al. 2025; Umili and Capobianco 2024; Umili et al. 2026); DeepProbLog, NeurASP, Logic Tensor Networks, and Scallop illustrate other neural–symbolic execution substrates (Manhaeve et al. 2018; Yang et al. 2020; Badreddine et al. 2022; Li et al. 2023). We do not replace these systems; we give an interface-level criterion for when a query-trained state can be fed back recurrently.

First-order symmetry, lifting, and graph locality. The anonymous response structures connect the intervention language to finite-model theory, Ehrenfeucht–Fraïssé games, Scott analysis, and orbit equivalence. Their count/orbit collapses are analogous to lifted inference through exchangeability and domain-liftable first-order counting (Niepert and Van den Broeck 2014; van Bremen and Kuželka 2023; Wang et al. 2024). Causal models with infinitely many variables provide a neighboring infinite-domain setting (Halpern and Peters 2022; Ibeling and Icard 2020). For neural execution, logical characterizations and distributed lower bounds for message-passing GNNs explain why local architectures inherit modal-depth limitations (Barceló et al. 2020; Grohe 2021; Loukas 2020).

Novelty boundary. We use classical tools openly: Gaifman locality, Moore/Nerode refinement, bounded morphisms, Möbius inversion, Scott's theorem, the Łoś–Vaught test, and elementary sampling bounds. The new technical core is instead the combination specific to intervention interfaces: a rank-preserving response-structure translation with

variable rebinding; the class-wide residual-clutter closure criterion; the anonymous-action bisimulation/type theory with a one-hidden-layer finite-core update whose explicit compilation has no ambient-domain factor; and sharp separation among closure, stabilisation, and type-determinism, including the single-mechanism family in Theorem 34(a). The real-valued minimax theorem below is a secondary quantitative extension.

Shallow intervention syntax, deep causal execution

Let $G = (D, E)$ be a finite directed acyclic graph and let Boolean endogenous variables satisfy

$$X(v) = \bigvee_{u:E(u,v)} X(u), \qquad \bigvee \varnothing = 0. \tag{1}$$

At the unmodified world define the depth-one causal/modal query

$$\mathrm{Influences}(x, y) := [\mathrm{do}(X(x) = 1)]X(y). \tag{2}$$

Theorem 1. Semantic depth and execution depth can separate. *For the mechanism in* (1)*:*

(i) $\mathrm{Influences}(x, y)$ *holds exactly when* y *is reachable from* x*, including the length-zero path.*

(ii) No single ordinary first-order formula $\phi(x, y)$ *over the baseline signature* $\{E, X\}$ *defines this binary relation on all finite directed acyclic graphs.*

(iii) Consider synchronous aggregate–combine message passing in which, after t *rounds, the target readout is a function only of the isomorphism type of its radius-*t undirected *neighbourhood with directed edges and node features retained as labels. If the only intervention information is a marker at the intervened node and there is no global pooling, virtual node, marker-relative positional encoding, precomputed reachability feature, or attention/readout over all nodes, then some directed chains with source–target undirected distance* L *require at least* L *rounds.*

Proof. After setting $X(x) = 1$, topological induction gives $X(y) = 1$ exactly at vertices reachable from x. For (ii), baseline X is identically zero. Gaifman locality for first-order formulas with free variables gives, for a hypothetical defining formula $\phi(x, y)$, a radius r such that sufficiently separated tuple neighbourhoods determine its truth value; see Libkin (Libkin 2004, Chapter 4). On a sufficiently long directed path, choose interior vertices x before y at Gaifman distance at least $2r + 2$. The radius-r neighbourhood of the ordered tuple (x, y) is then a disjoint union of two directed path segments, and it is isomorphic—with tuple coordinates respected—to the radius-r neighbourhood of (y, x). Reachability holds only for (x, y), contradicting locality.

For (iii), take the single chain

$$v_0 \to v_1 \to \cdots \to v_{2L} \tag{3}$$

with target v_L. Put the intervention marker at v_0 in one instance and at v_{2L} in the other. The marker is at undirected distance L from the target in both instances, but only the first has a directed path from marker to target. By induction on message-passing rounds, after t rounds the target state depends only on its radius-t initialized neighbourhood. Thus for every $t < L$ the target sees identical marked-feature neighbourhoods while the correct answers differ. This is the standard locality mechanism behind message-passing lower bounds (Barceló et al. 2020; Grohe 2021; Loukas 2020).

An exact symbolic executor evaluates one intervention in $O(|D| + |E|)$ time by a topological sweep; this upper bound is algorithmic rather than an additional expressiveness claim.

Message-passing locality check. To give the execution result an empirical anchor, we trained a four-layer directed aggregate–combine GNN on the paired-chain construction of Theorem 1(iii). The network uses separate incoming and outgoing message map and a target-only readout, with no global pooling, virtual node, positional encoding, or attention. Training examples place the intervention marker one to four edges on either side of the target; evaluation uses distances one through eight. Across five seeds, accuracy is 100% at distances 1–4 and exactly 50% at distances 5–8. More strongly, the mean absolute paired-logit gap is positive through distance four and identically zero thereafter. The latter equality is architectural rather than statistical: beyond four rounds the two target receptive fields are identical. Code and seed-level outputs are in the reproducibility archive.

Logical upper bracket. For each target parameter y, let P_y mark only y. The reachability query of Theorem 1(i) is equivalently $\langle E^* \rangle P_y$, the modal μ-calculus formula $\mu Z.(P_y \vee \langle E \rangle Z)$, or, under the standard non-reflexive transitive-closure convention, $x = y \vee [\mathrm{TC}_{u,v}\, E(u, v)](x, y)$. No fixed-depth basic-modal formula expresses the relation on all finite DAGs. More generally, for the acyclic OR mechanism in (1), every formula built from first-order atoms, Boolean connectives, quantifiers, and persistent positive interventions is FO(TC)-definable on the baseline DAG: carry the finite list of intervention roots in scope and translate $X(y)$ to a disjunction of reflexive-reachability predicates. The complete compilation, including bound-variable rebinding, is given in Supplementary Section S6.10.

The theorem and this upper bracket isolate the expressiveness–computation trade-off that motivates the rest of the paper: causal/modal depth one does not imply one-step neural execution, and compiling transitive execution into a richer logical operator or a precomputed transition relation moves rather than removes the computational cost. The remaining sections ask how much state must be retained once those transitions are executed repeatedly.

Finite named-control intervention interfaces

The main theory uses a finite, explicitly named control domain

$$D = C_m = \{c_1, \ldots, c_m\}. \tag{4}$$

There are Boolean controls $X_1, \ldots, X_m$ and an observed output

$$Y = F(X_1, \ldots, X_m). \tag{5}$$

Only persistent positive interventions $X_i := 1$ are allowed. A concrete state is a pair (F, S), where $S \subseteq [m]$ is the set of controls already activated, and action i maps S to $S \cup \{i\}$. Rigid predicates and equality may be queried, but neither S nor the hidden mechanism F is directly observable.

Terms are variables or constants c_i. Let $\mathcal{L}_h$ be the intervention-depth-h fragment generated by

$$\psi ::= R(\bar{t}) \mid t = t' \mid Y \mid \neg\psi \mid (\psi \wedge \psi) \mid \exists x\, \psi \mid [\mathrm{do}(X(t) = 1)]\psi. \tag{6}$$

Intervention depth is defined by $\mathrm{dd}(R(\bar{t})) = \mathrm{dd}(t = t') = \mathrm{dd}(Y) = 0$, $\mathrm{dd}([\mathrm{do}(X(t) = 1)]\psi) = 1 + \mathrm{dd}(\psi)$, quantifiers transparent, and the maximum over Boolean connectives. Thus $\mathcal{L}_h$ consists exactly of formulas with $\mathrm{dd} \leq h$. There are no ordinary Kripke modalities in this completeness result. The quantification domain is exactly C_m, so every intervention term denotes a control; controls outside the current set S have value zero. Because D is a fixed named finite set, quantification can be expanded into a finite disjunction. We therefore use the logical notation for compositional query specification rather than claiming an expressiveness advantage over propositional enumeration. The preceding execution section separated logical depth from causal execution depth; the present section isolates the state-information problem once actions themselves are available.

Define the response tensor

$$\mathsf{V}_h(F, S) = \big(F(1_{S \cup T})\big)_{T \subseteq [m],\ |T| \leq h}. \tag{7}$$

The constants c_i (or equivalent rigid singleton predicates) are essential: without names, permutation-symmetric controls need not be individually distinguishable.

Theorem 2. Query completeness and horizon consumption. *For two states (F, S) and (G, S') over the common named control domain and identical rigid predicates, write $(F, S) \equiv_{\mathcal{L}_h} (G, S')$ when the states agree on every* sentence *of $\mathcal{L}_h$. Then*

$$(F, S) \equiv_{\mathcal{L}_h} (G, S') \quad \Longleftrightarrow \quad \mathsf{V}_h(F, S) = \mathsf{V}_h(G, S'). \tag{8}$$

For $h \geq 1$, after action i the remaining-horizon tensor is obtained by exact reindexing:

$$\mathsf{V}_{h-1}(F, S \cup \{i\})_T = \mathsf{V}_h(F, S)_{T \cup \{i\}}, \qquad |T| \leq h - 1. \tag{9}$$

Proof sketch. Structural induction shows that a formula of intervention depth at most h depends only on the coordinates of V_h; a do-modality consumes one horizon unit and reindexes the tensor. Conversely, every tensor coordinate is read by the corresponding finite do-word. Full details are in Supplementary Section S6.11.

The theorem characterizes information for answering the declared query family. It does not imply that the quotient can be reused at the same horizon after an action.

Residual clutters on arbitrary control domains

The same notation will be used on finite and countable control domains. A *clutter* $H \subseteq [D]^{<\omega}$ is an antichain of finite subsets. It induces the monotone mechanism

$$F_H(S) = \mathbf{1}\{\exists E \in H : E \subseteq S\}, \qquad S \in [D]^{<\omega}. \tag{10}$$

For a finite current set S, define the residual clutter and its horizon truncation by

$$\mathcal{R}_H(S) = \min_{\subseteq}\{E \setminus S : E \in H\}, \tag{11}$$

$$\mathcal{R}_{H,h}(S) = \{R \in \mathcal{R}_H(S) : |R| \le h\}. \tag{12}$$

Because every residual $E \setminus S$ is finite, any nonempty family of residuals has a member of least cardinality; hence the inclusion-minimal operator is well defined even when H is infinite. When F is monotone on finite subsets, its minimal true sets form such a clutter $\mathcal{H}_F$ and $F = F_{\mathcal{H}_F}$.

Anonymous first-order completeness and lifted orbit closure

Remove the constants c_i and retain equality as the only rigid structure. The control domain D is a *pure* set, finite or countably infinite. States are pairs (F, S) with $S \subseteq D$ finite and F defined on finite subsets of D; we write $F(A)$ for $F(\mathbf{1}_A)$. The anonymous language $\mathcal{L}_h^{\mathrm{anon}}$ is generated by

$$\psi ::= x = y \mid Y \mid \neg\psi \mid (\psi \wedge \psi) \mid \exists x\, \psi \mid [\mathrm{do}(X(x) = 1)]\psi, \tag{13}$$

where all terms are variables, quantifiers range over D, and $(F, S), b \models [\mathrm{do}(X(x) = 1)]\psi$ iff $(F, S \cup \{b(x)\}), b \models \psi$. Intervention depth is defined by $\mathrm{dd}(x = y) = \mathrm{dd}(Y) = 0$, $\mathrm{dd}([\mathrm{do}(X(x) = 1)]\psi) = 1 + \mathrm{dd}(\psi)$, quantifiers transparent, and the maximum over Boolean connectives; $\mathcal{L}_h^{\mathrm{anon}}$ consists of the formulas of depth at most h. Quantifier rank qr is the usual first-order notion; no bound is imposed on it unless stated. For $q \in \mathbb{N} \cup \{\infty\}$ write $(F, S) \equiv_h^q (G, S')$ when the two states agree on all sentences of $\mathcal{L}_h^{\mathrm{anon}}$ of rank at most q, and $\equiv_h := \equiv_h^\infty$.

Response structures and the translation lemma

Definition 3. Response structure. Let $\sigma_h = \{R_0, R_1, \ldots, R_h\}$, where R_j is a j-ary relation symbol (R_0 nullary). For a state (F, S) over D, the *depth-h response structure* is the σ_h-structure

$$\mathfrak{M}_h(F, S) = (D, R_0, \ldots, R_h), \qquad R_j(a_1, \ldots, a_j) \iff F(S \cup \{a_1, \ldots, a_j\}) = 1,$$

with repeated arguments allowed. Each R_j is symmetric and depends only on the set $\{a_1, \ldots, a_j\}$.

A bijection $\sigma : D \to D$ is an isomorphism $\mathfrak{M}_h(F, S) \cong \mathfrak{M}_h(G, S')$ exactly when

$$F(S \cup T) = G(S' \cup \sigma[T]) \qquad \text{for all } T \subseteq D,\ |T| \le h. \tag{14}$$

For finite $D = [m]$ this is precisely the statement that the response tensors $V_h(F, S)$ and $V_h(G, S')$ lie in the same $\mathrm{Sym}(m)$-orbit; for countable D we call (14) *orbit equivalence under* $\mathrm{Sym}(D)$.

Lemma 4. Translation. *Fix $h \geq 0$.*

(a) For every formula ψ of the intervention language and every finite sequence of variables $\pi = (y_1, \ldots, y_\ell)$ with $\ell + \mathrm{dd}(\psi) \leq h$ there is a first-order σ_h-formula ψ^π with $\mathrm{qr}(\psi^\pi) = \mathrm{qr}(\psi)$ and free variables among those of ψ and π, such that for every state (F, S) and every assignment b,

$$(F,\ S \cup b[\pi]),\ b \models \psi \iff \mathfrak{M}_h(F, S),\ b \models \psi^\pi, \qquad b[\pi] := \{b(y_1), \ldots, b(y_\ell)\}. \tag{15}$$

(b) For every first-order σ_h-formula θ there is $\theta^\natural \in \mathcal{L}_h^{\mathrm{anon}}$ with $\mathrm{qr}(\theta^\natural) = \mathrm{qr}(\theta)$ and the same free variables, such that $(F, S), b \models \theta^\natural$ iff $\mathfrak{M}_h(F, S), b \models \theta$.

Proof sketch. Translate Y under a frozen do-variable list $\pi = (y_1, \ldots, y_\ell)$ to $R_\ell(\pi)$, commute the translation with Boolean connectives and quantifiers, and append the do-variable at each intervention. If a later quantifier re-binds a variable already frozen in π, alpha-rename that bound occurrence to the first fresh variable in a fixed enumeration before applying the induction. This is the only non-naive case; it preserves both intervention depth and quantifier rank. The reverse translation replaces each R_j atom by a length-j do-word. Full details are in Supplementary Section S6.12.

Corollary 5. Reduction to first-order logic over response structures. *For every $q \in \mathbb{N} \cup \{\infty\}$,*

$$(F, S) \equiv_h^q (G, S') \iff \mathfrak{M}_h(F, S) \equiv_q \mathfrak{M}_h(G, S'),$$

where $\equiv_q$ is agreement on first-order σ_h-sentences of rank at most q. In particular, anonymous intervention-depth-h equivalence is elementary equivalence of response structures, and orbit equivalence is their isomorphism.

Proof. Apply Lemma 4(a) with $\pi = ()$ to sentences of $\mathcal{L}_h^{\mathrm{anon}}$, and Lemma 4(b) to σ_h-sentences; both maps preserve rank and truth.

Intervention depth and quantifier rank thus become two independent syntactic parameters of a single first-order problem: depth fixes the signature σ_h, while rank is ordinary quantifier rank over that signature. The results below are consequences of this reduction together with standard model theory.

Finite domains: orbit completeness and the price of bounded rank

Theorem 6. Anonymous completeness up to permutation. *For two states over a pure m-element control domain,*

$$(F, S) \equiv_h (G, S') \iff \exists \pi \in \mathrm{Sym}(m) \text{ such that } F(S \cup T) = G(S' \cup \pi(T)) \quad \text{for every } T \subseteq [m],\ |T| \leq h.$$

Moreover, rank m suffices: $(F, S) \equiv_h^m (G, S')$ already implies the right-hand side.

Proof sketch. By the translation lemma it suffices to work with the finite response structures. A rank-m characteristic sentence existentially enumerates the m domain elements and records the truth of every symmetric response relation on every subset of size at most h. A witness in another m-element structure is therefore an isomorphism. Full details are in Supplementary Section S6.13.

The characteristic-sentence construction is the first-order image of the direct tensor-orbit argument.

Proposition 7. Exact rank characterization at depth one. *Let D be finite or countably infinite, $q \geq 1$, and put $P_{F,S} := \{x \in D : F(S \cup \{x\}) = 1\}$. Then $(F, S) \equiv_1^q (G, S')$ iff*

$$\begin{aligned} F(S) &= G(S'), \\ \min\{|P_{F,S}|, q\} &= \min\{|P_{G,S'}|, q\}, \\ \min\{|D \setminus P_{F,S}|, q\} &= \min\{|D \setminus P_{G,S'}|, q\}, \end{aligned}$$

with cardinalities in $\mathbb{N} \cup \{\aleph_0\}$. For $q = \infty$ the condition is $F(S) = G(S')$ together with equality of both cardinalities, which is orbit equivalence at $h = 1$.

Proof sketch. At depth one the response structure is a pure set with one unary predicate $P_{F,S}$ plus the output bit. The q-round Ehrenfeucht–Fraïssé game on unary structures is determined exactly by the two colour-class sizes truncated at q. Corollary 5 transfers this criterion back to the intervention language. Full details are in Supplementary Section S6.14.

Corollary 8. Sharp rank threshold for disjunctions. *Let $|D| = m$ and consider the states $(\mathrm{OR}_A, \emptyset)$, $A \subseteq D$, where $\mathrm{OR}_A(U) = 1$ iff $U \cap A \neq \emptyset$. On this family, rank-q depth-one equivalence coincides with orbit equivalence iff $m \leq 2q$. For every $m \geq 2q + 1$, sets with $|A| = q$ and $|A'| = q + 1$ give states that agree on every rank-$\leq q$ sentence of $\mathcal{L}_1^{\mathrm{anon}}$ but lie in different $\mathrm{Sym}(m)$-orbits.*

Proof. Here $F(\emptyset) = 0$ and $P = A$, so orbit equivalence is $|A| = |A'|$, while by Proposition 7 rank-q equivalence is equality of $\tau(a) := (\min\{a, q\}, \min\{m - a, q\})$ with $a = |A|$. If $m \leq 2q$ and $\tau(a) = \tau(a')$ with $a < a'$, then $a \geq q$ and $m - a' \geq q$, so $a' \leq m - q \leq q \leq a$, a contradiction. If $m \geq 2q + 1$, then $\tau(q) = \tau(q + 1) = (q, q)$.

This identifies the exact rank threshold. Intervention depth and quantifier rank are therefore independent resources with an exact trade-off already at depth one. For threshold mechanisms, by contrast, Theorem 12(i) shows that rank h suffices.

Corollary 9. Anonymous symmetric conjunction. *For $F(X) = \bigwedge_{i=1}^m X_i$, two anonymous states are depth-h equivalent exactly when their missing counts $d = m - |S|$ satisfy*

$$\min\{d, h + 1\} = \min\{d', h + 1\}. \tag{16}$$

Hence the anonymous quotient has

$$N_h^{\mathrm{anon}}(m) = 1 + \min\{m, h + 1\} \tag{17}$$

classes, compared with $1 + \sum_{j=0}^h \binom{m}{j}$ in the named case when $h < m$.

Proof. Permutation symmetry reduces the tensor orbit to the number of missing controls. If $d \le h$, the least number of distinct positive interventions required to make Y true is exactly d, so $d = 0, \dots, h$ are distinct. If $d > h$, every depth-h do-word remains false. Apply Theorem 6.

Theorem 10. Lifted recurrent closure for anonymous threshold mechanisms. *Fix $k \ge 1$ and, for every finite $m \ge k$, let*

$$F_{k,m}(S) = \mathbf{1}\{|S| \ge k\}. \tag{18}$$

On permutation orbits, let the anonymous action α mean "activate one inactive control"; it maps the orbit of a set of size $s < m$ to the orbit of size $s + 1$ and fixes the full orbit. Put $d(S) = \max\{k - |S|, 0\}$. For finite domains, Theorem 6 identifies the anonymous query partition with tensor orbits. Specialized to thresholds, the depth-h orbit query state is exactly

$$\min\{d(S), h + 1\}, \tag{19}$$

so it has $1 + \min\{k, h + 1\}$ classes. This quotient is recurrently closed under α iff $h \ge k - 1$, in which case it has exactly $k + 1$ recurrent states, independent of m.

Under the named action alphabet $[m]$, by contrast, the full Nerode index is

$$1 + \sum_{j=0}^{k-1} \binom{m}{j}. \tag{20}$$

For fixed k this grows as $\Theta(m^{k-1})$, while at the conjunction endpoint $k = m$ it equals 2^m; the anonymous orbit machine has only $m + 1$ states there.

Proof. After j applications of α, the output is one exactly when $j \ge d(S)$, with $d = 0$ already true. Hence all deficits $0, 1, \dots, h$ are distinguished and all larger deficits collapse, proving (19). Equivalently, Theorem 6 says that the depth-h response structure is determined up to isomorphism by the same truncated deficit, so the tensor-orbit statement and the one-letter Moore calculation describe the same partition. If $h < k - 1$, the far class contains deficits $h + 1$ and $h + 2$; one application of α sends the first to the distinguished deficit h and the second back to the far class, so closure fails. If $h \ge k - 1$, all deficits $0, \dots, k$ are already distinct, and α deterministically maps d to $\max\{d - 1, 0\}$.

For named actions, the threshold mechanism has minimal triggers equal to all k-subsets. At a state S with $|S| < k$, its full residual clutter is the family of all $(k - |S|)$-subsets of $[m] \setminus S$, which uniquely determines S. Hence distinct sub-threshold sets have distinct future acceptance languages, while every set of size at least k is an absorbing accepting state. The Nerode classes are therefore the individual sub-threshold subsets plus one accepting class, giving (20).

Theorem 10 is uniform over all finite $m \ge k$: for fixed threshold k, the lifted recurrent state and its closure horizon do not grow with domain size. The trade-off is action

semantics: α activates a fresh inactive control rather than naming a particular individual. The next theorem shows that this orbit machine persists on an infinite domain rather than being an artifact of finite enumeration.

Remark 11. Threshold mechanisms are all symmetric monotone mechanisms. If F is invariant under all permutations of a pure domain, then $F(S)$ depends only on $|S|$. If F is also monotone, then $F(S) = \mathbf{1}\{|S| \geq k\}$ for some $k \in \mathbb{N} \cup \{\infty\}$. Apart from the two trivial constant cases $k = 0$ and $k = \infty$, Theorem 10 and Theorem 12 therefore cover every monotone mechanism whose symmetry group is the full symmetric group, not merely an example family. Symmetry is also what makes the orbit action α deterministic: all fresh choices are conjugate by automorphisms.

Countably infinite domains

Let D be countably infinite. States have finite $S \subset D$; for threshold mechanisms $F_k(S) = \mathbf{1}\{|S| \geq k\}$ and $d(S) = \max\{k - |S|, 0\}$.

Theorem 12. Lifted threshold closure over a countable domain. *In this setting:*

(i) $(F_k, S) \equiv_h (F_k, S')$ *iff* $\min\{d(S), h+1\} = \min\{d(S'), h+1\}$. *Separating sentences have quantifier rank at most h, so the rank-q fragment induces the same partition for every $q \geq h$.*

(ii) The Moore relation $\equiv_h^\alpha$ *over the one-letter alphabet* $\{\alpha\}$ *induces the same partition.*

(iii) This partition is a right congruence for α iff $h \geq k - 1$. In that case the recurrent machine has exactly $k+1$ states and α acts by $d \mapsto \max\{d-1, 0\}$, for finite domains and countable D alike.

(iv) If instead every $u \in D$ is named and has its own positive intervention action, the Nerode index is infinite for every $k \geq 2$: distinct finite sub-threshold sets are pairwise Nerode-inequivalent. For $k = 1$ the index is 2.

Proof sketch. The depth-h response structure of a threshold state depends only on the truncated deficit $\min\{d, h+1\}$: deficits above h make all response relations empty, deficit zero makes them full, and equal intermediate deficits are carried to one another by a permutation of the countable domain. Rank-j separating sentences recover each $d \leq h$. The one-letter action α induces the same truncated-deficit Moore relation, yielding the closure threshold $h = k - 1$. Under named actions, distinct finite sub-threshold sets are separated by suitable fresh activations. Full details are in Supplementary Section S6.15.

The threshold response structures are homogeneous: when $1 \leq d \leq h$ the truncated residual clutter of S is the *infinite* family of all d-subsets of $D \setminus S$, yet finitary sentences suffice because the structure has a large automorphism group. The next three results show that this is one of two finitary mechanisms, and that without either of them finitary orbit completeness genuinely fails.

Finite residual cores. For a monotone F on finite subsets of D, let $\mathcal{H}_F$ be its clutter of minimal true sets and use the residual notation of Eqs. (11)–(12).

Lemma 13. *Let F be monotone, S finite, and $\mathcal{K} = \mathcal{R}_{\mathcal{H}_F,h}(S)$.*

(a) For finite A with $|A| \le h$: $F(S \cup A) = 1$ iff $K \subseteq A$ for some $K \in \mathcal{K}$.

(b) $\mathcal{K}$ is exactly the family of $\subseteq$-minimal sets A with $|A| \le h$ and $F(S \cup A) = 1$.

Proof. (a) $F(S \cup A) = 1$ iff $E \setminus S \subseteq A$ for some $E \in \mathcal{H}_F$. Among the residuals contained in the finite set A, choose one, K_0, of least cardinality. Every residual properly contained in K_0 would also lie in A, so K_0 is $\subseteq$-minimal among all residuals, and $|K_0| \le |A| \le h$; hence $K_0 \in \mathcal{K}$. The converse is immediate. For (b), part (a) says that every true finite set A of size at most h contains some $K \in \mathcal{K}$. Hence an inclusion-minimal such A must equal a member of $\mathcal{K}$; conversely every $K \in \mathcal{K}$ is true and no proper subset is true, because $\mathcal{K}$ is an antichain.

For $1 \le j \le h$ and distinct variables $\bar{x} = (x_1, \dots, x_j)$ let

$$\mathrm{Min}_j(\bar{x}) := R_j(\bar{x}) \wedge \bigwedge_{\bar{y} \subsetneq \bar{x}} \neg R_{|\bar{y}|}(\bar{y}),$$

$$\mathrm{Supp}(z) := \bigvee_{j=1}^{h} \exists y_2 \cdots \exists y_j \Big(\mathrm{Dist}(z, \bar{y}) \wedge \mathrm{Min}_j(z, \bar{y}) \Big).$$

where $\bar{y} \subsetneq \bar{x}$ ranges over proper subsequences and Dist asserts pairwise distinctness. By Lemma 13(b), Supp defines in $\mathfrak{M}_h(F, S)$ the support $\bigcup \mathcal{K}$, with rank $h - 1$.

Theorem 14. Orbit completeness for finite residual cores. *Let D be a countable pure domain (finite or infinite) and $h \ge 1$. Let F, G be monotone and S, S' finite, and suppose $\mathcal{K} = \mathcal{R}_{\mathcal{H}_F,h}(S)$ is finite with support U of size n. Put $\mathcal{K}' = \mathcal{R}_{\mathcal{H}_G,h}(S')$. The following are equivalent:*

(i) $(F, S) \equiv_h (G, S')$;

(ii) $(F, S) \equiv_h^{n+h} (G, S')$;

(iii) some permutation σ of D satisfies $\{\sigma[K] : K \in \mathcal{K}\} = \mathcal{K}'$.

In particular, $\mathcal{K}'$ is then finite with a support of size n. Only one of the two clutters needs to be assumed finite.

Proof sketch. A permutation carrying the truncated residual clutter preserves every response relation by Lemma 13. Conversely, the finite support is first-order definable, and a characteristic sentence naming its n elements and all minimal residuals has quantifier rank at most $n + h$; any model satisfying that sentence yields the required support bijection, which extends to a permutation of the countable pure domain. Full details are given in Supplementary Section S6.1.

For finite D this is the anonymous counterpart of Theorem 21: anonymous depth-h equivalence of monotone states is isomorphism of truncated residual clutters. For countable D it covers every mechanism with finitely many minimal triggers, and more generally every state whose truncated residual is finite. This includes all threshold states with $d > h$ ($\mathcal{K} = \emptyset$) or $d = 0$ ($\mathcal{K} = \{\emptyset\}$).

Infinitary completeness. Let $\mathcal{L}^{\mathrm{anon}}_{h,\omega_1\omega}$ extend (13) by countable conjunctions and disjunctions of formulas with finitely many free variables in total. The recursion of Lemma 4 extends verbatim, with countable connectives commuting with $(\cdot)^\pi$ and $(\cdot)^\natural$. A formula has only countably many binding occurrences and finitely many free variables, so alpha-renaming can inject all bound-variable occurrences into the even-indexed variable symbols while reserving infinitely many odd-indexed symbols. Fresh variables for the rebinding clause are therefore always available.

Theorem 15. Orbit completeness in $\mathcal{L}_{\omega_1\omega}$. *Let D be countable and let F, G be arbitrary mechanisms (not necessarily monotone). Then (F, S) and (G, S') agree on all sentences of $\mathcal{L}^{\mathrm{anon}}_{h,\omega_1\omega}$ iff some permutation of D satisfies* (14)*.*

Proof. By the extended translation, agreement is $\mathcal{L}_{\omega_1\omega}$-equivalence of the countable structures $\mathfrak{M}_h(F, S)$ and $\mathfrak{M}_h(G, S')$ in the finite signature σ_h. By Scott's isomorphism theorem (Scott 1965; Hodges 1993), countable structures in a countable signature are $\mathcal{L}_{\omega_1\omega}$-equivalent iff they are isomorphic. Isomorphism is (14).

Finitary orbit completeness fails in general.

Proposition 16. *Let D be countably infinite. There are monotone rank-two mechanisms F_1, F_2 on D such that $(F_1, \emptyset) \equiv_{h'} (F_2, \emptyset)$ for every finite intervention depth $h' \geq 0$, yet no permutation of D satisfies* (14) *already at depth* 2*.*

Proof sketch. Use one bi-infinite path and the disjoint union of two such paths as rank-two trigger graphs. Their response structures are definitional expansions of 2-regular acyclic graphs. The first-order theory of 2-regular acyclic graphs is complete by the Łoś–Vaught test, but the two graphs are not isomorphic because they have different numbers of connected components. Full details are given in Supplementary Section S6.2.

The depth boundary is sharp. At $h = 1$, Proposition 7 with unbounded quantifier rank gives finitary orbit completeness for every mechanism on a finite or countable pure domain. Proposition 16 shows that finitary orbit completeness can already fail at depth 2, even for monotone rank-two mechanisms.

Remark 17. Countable query classification and recurrent frontier. On countable pure domains, anonymous intervention equivalence is always orbit equivalence in the infinitary logic (Theorem 15). In the finitary logic it is orbit equivalence when the truncated residual core is finite, with explicit rank $|U| + h$ (Theorem 14), or when the response structure is sufficiently homogeneous, with rank h for thresholds (Theorem 12). It fails already for rank-two mechanisms with infinite residuals (Proposition 16).

The *recurrent* side for non-symmetric mechanisms is addressed in the section "Lifted recurrent closure under anonymous actions." Anonymous activation is nondeterministic

on query classes, recurrent closure becomes a bisimulation decided by the 1-types of the next response structure, finite residual cores make the update a finite computation, and the finitary countable case exhibits the recurrent analogue of Proposition 16.

When a query quotient is a recurrent state

From this section onward, $\approx_h$ denotes finite-word behavioral equivalence, distinct from the anonymous first-order relation $\equiv_h$ used above. To connect the logical quotient to classical automata, consider a deterministic action-labelled output system $\mathcal{S} = (\Omega, A, f, o)$ with action alphabet A, transitions $f_a : \Omega \to \Omega$, and output $o : \Omega \to \mathcal{Y}$. For a word $v = a_1 \cdots a_k$, write $f_v = f_{a_k} \circ \cdots \circ f_{a_1}$. An abstraction e is *exact for the word* v when $o \circ f_v$ factors through e, i.e. there exists $\widehat{o}_v$ with $o \circ f_v = \widehat{o}_v \circ e$; exactness for every finite word means this holds for all v. Write $\omega \approx_h \omega'$ when $o(f_v\omega) = o(f_v\omega')$ for every word of length at most h, and write $\omega \equiv_\infty \omega'$ when this holds for every finite word.

Lemma 18. Moore stabilization criterion. *For every $h \geq 0$, the following are equivalent:*

(i) $\approx_h$ *is a right congruence for every action;*

(ii) $\approx_h = \approx_{h+1}$*;*

(iii) $\approx_h = \equiv_\infty$*.*

Proof. The relations refine monotonically: $\approx_{h+1} \subseteq \approx_h$. If $\approx_h$ is a congruence and $\omega \approx_h \omega'$, then for any word av of length at most $h+1$, congruence gives $a\omega \approx_h a\omega'$, so the outputs after v agree. Thus $\omega \approx_{h+1} \omega'$ and (i) implies (ii).

Assume (ii). If $\omega \approx_h \omega'$, then for any action a and any word v of length at most h, the word av has length at most $h+1$, so equality of $\approx_h$ and $\approx_{h+1}$ implies that v produces the same output from $a\omega$ and $a\omega'$. Hence $a\omega \approx_h a\omega'$, proving (i). Reapplying the same argument inductively gives $\approx_h = \approx_{h+k}$ for every k, which is exactly (iii). Finally, Nerode equivalence is a right congruence, so (iii) implies (i).

Recurrent closure as exact do-commuting abstraction

Let $Z_h = \Omega/\approx_h$ and let $e_h : \Omega \to Z_h$ be the quotient map.

Definition 19. Exact do-commuting abstraction. Let $\mathcal{Z} = (Z, A, g, \widehat{o})$ be a second deterministic system over the same action alphabet. A surjection $\tau : \Omega \to Z$ is an *exact do-commuting abstraction* of $\mathcal{S}$ onto $\mathcal{Z}$ when

$$\widehat{o} \circ \tau = o, \qquad \tau \circ f_a = g_a \circ \tau \quad \text{for every } a \in A. \tag{21}$$

This is the deterministic action-labeled specialization of the comuting-square requirement used in causal abstraction: abstracting before or after a do-action gives the same abstract state and output.

Proposition 20. Positioning: recurrent closure is exact abstraction. *For every $h \geq 0$, the following are equivalent:*

(i) $\approx_h$ *is a right congruence;*

(ii) there is a system $\mathcal{Z}_h = (Z_h, A, g, \widehat{o})$ *for which* e_h *is an exact do-commuting abstraction;*

(iii) e_h *is exact for every finite action word, equivalently* $\approx_h = \equiv_\infty$.

Moreover:

(iv) Graded exactness always holds. *For every $h \geq 1$ and $a \in A$ there is a unique $g_a^{(h)} : Z_h \to Z_{h-1}$ satisfying*

$$e_{h-1} \circ f_a = g_a^{(h)} \circ e_h. \tag{22}$$

For the named-control system, $g_a^{(h)}$ is exactly the tensor reindexing in (9).

(v) When (ii) holds, g and $\widehat{o}$ are unique.

(vi) More generally, a surjection $\tau : \Omega \to Z$ is an exact do-commuting abstraction onto some deterministic $\mathcal{Z}$ iff $\ker \tau$ *is a right congruence contained in $\approx_0$. Every such kernel refines $\equiv_\infty$, so the Nerode quotient is the coarsest exact abstraction compatible with the output.*

Proof. (i)⇒(ii): define $g_a([\omega]_h) = [f_a\omega]_h$, which is well defined by congruence, and $\widehat{o}([\omega]_h) = o(\omega)$, which is well defined because the empty word is a depth-h query. The two equalities in (21) then hold by construction. The converse follows immediately from the second commuting equality. Equivalence of (i) and (iii) is Lemma 18.

For (iv), if $\omega \approx_h \omega'$ and $|v| \leq h-1$, then av has length at most h, so $f_a\omega \approx_{h-1} f_a\omega'$. Hence $g_a^{(h)}([\omega]_h) = [f_a\omega]_{h-1}$ is well defined; surjectivity of e_h gives uniqueness. Part (v) is the same surjectivity argument. For (vi), exactness implies that $\ker\tau$ is a congruence and lies inside $\approx_0$; conversely those two properties make $g_a([\omega]) = [f_a\omega]$ and $\widehat{o}([\omega]) = o(\omega)$ well defined. Induction on word length gives $\ker\tau \subseteq \approx_k$ for every k, hence $\ker\tau \subseteq \equiv_\infty$.

The proposition separates two notions that are easy to conflate. A depth-h query quotient is always exact *across* horizons, from Z_h to Z_{h-1}. Reusing the same quotient at a fixed horizon requires stabilization. In causal-abstraction language, query-level adequacy does not by itself guarantee that each deployment do-action has a well-defined high-level counterpart.

For the finite named-control intervention system above, commutativity and idempotence mean that words can be reduced to subsets, and $\approx_h$ is exactly the tensor equivalence in Theorem 2. The lemma therefore gives a sharp test: *a depth-h query representation may be reused at depth h after each action exactly when the depth-h partition has already stabilized.*

Monotone Boolean mechanisms: residual clutters and closure

For a family $\mathfrak{H}$ of monotone mechanisms, write

$$\Omega(\mathfrak{H}) = \{(F_H, S) : H \in \mathfrak{H},\ S \subseteq [m]\}. \tag{23}$$

Congruence is always relative to this ambient state space. This matters when the mechanism is unknown: two states with different mechanisms may be query-equivalent even when their successors are not. Specialize the definitions from "Residual clutters on arbitrary control domains" to the finite named domain $D = [m]$. The members of H are the minimal activating sets of F_H, and $\mathcal{R}_{H,h}(S)$ is the information visible to depth-h positive intervention queries.

Theorem 21. Residual characterization. *For monotone Boolean states (F_H, S) and $(F_{H'}, S')$,*

$$(F_H, S) \equiv_{\mathcal{L}_h} (F_{H'}, S') \quad \Longleftrightarrow \quad \mathcal{R}_{H,h}(S) = \mathcal{R}_{H',h}(S'). \tag{24}$$

Moreover, the exact remaining-horizon update is

$$\mathcal{R}_{H,h-1}(S \cup \{i\}) = \{R : |R| \le h-1,\ R \in \min_{\subseteq} \mathcal{Q}_i(S)\}, \tag{25}$$
$$\mathcal{Q}_i(S) = \{Q \setminus \{i\} : Q \in \mathcal{R}_{H,h}(S)\}.$$

Proof sketch. For a monotone mechanism, the inclusion-minimal activating sets after state S are exactly its residual clutter, so truncating to sets of size at most h determines every declared depth-h query and is recoverable from those answers. After action i, residuals map to $Q \setminus \{i\}$ followed by minimization; a residual omitted because its size exceeds h cannot become a proper subset of any candidate relevant to the $(h-1)$ truncation. Full details are in Supplementary Section S6.16.

Exact separation for conjunction

For the conjunction $F(X) = \bigwedge_{i=1}^{m} X_i$, the residual clutter contains a single missing set $M = [m] \setminus S$. All M with $|M| \le h$ are separately identifiable, while all larger M share the all-zero depth-h response tensor. Histories start at $S = \varnothing$, so all 2^m subsets are reachable. A recurrent predictor may receive all outputs revealed on earlier steps, but after each new action it must emit a prediction $\widehat{Y} \in [0,1]$ for the post-action output before that same output is revealed; error means absolute error $|\widehat{Y} - Y|$.

Theorem 22. Conjunction separation and tight stabilization threshold. *For the m-way conjunction, the number of depth-h query classes is*

$$N_h(m) = \begin{cases} 1 + \displaystyle\sum_{j=0}^{h} \binom{m}{j}, & 0 \le h < m, \\ 2^m, & h \ge m. \end{cases} \tag{26}$$

For $0 \le h \le m-2$, $\approx_h$ is not a congruence; for $h \ge m-1$, it equals Nerode equivalence. Consequently, for $m \ge 2$, any deterministic recurrent state obeying the prediction protocol above and achieving uniform error $< 1/2$ after arbitrary positive-intervention histories has at least 2^m reachable states, even if previously revealed outputs are fed back to the updater.

Proof sketch. For conjunction, the residual is the missing-control set. Distinct missing sets of size at most h give distinct query classes, while all larger ones collapse, giving the stated class count. The far class is not stable when $h \le m-2$: missing sets of sizes $h+1$ and $h+2$ are separated after activating a control from the smaller set. At $h = m-1$ every concrete state is distinguished. The 2^m recurrent lower bound follows by continuing any two distinct histories with actions that complete one while preserving a missing control in the other, so identical recurrent states would force a $1/2$ prediction error on one continuation. Full details are in Supplementary Section S6.17.

Remark 23. One-step versus multi-step output feedback. For $h \ge 1$, the output observed immediately after one action i is already the singleton coordinate $\mathsf{V}_h(F,S)_{\{i\}}$, so one-step feedback cannot refine the successor beyond information already present in the pre-action depth-h tensor. After more than h actions, however, output feedback can reveal genuinely new information (for conjunction, observing $Y=1$ certifies that all missing controls have been supplied). The lower bound above does not deny this: its collision witness keeps all earlier continuation outputs identical and needs only the final prediction to precede its observation. The experiment described below deliberately supplies no output feedback so that the query interface is the only recurrent information channel.

The 2^m recurrent lower bound is exactly the Myhill–Nerode size of the conjunction Moore machine; the contribution of Theorem 22 is the sharp comparison with the much smaller finite-horizon quotient $N_h(m)$ and the stabilization threshold $h = m-1$. For $m = 20$ and $h = 2$, (26) gives 212 query classes versus $2^{20} = 1{,}048{,}576$ recurrent states. This is exponential in discrete state count, not in the number of real coordinates; a neural-dimension lower bound requires precision, margin, or robustness assumptions.

Bounded-rank closure and the role of the hypothesis class

Define $\mathrm{rank}(H) = \max_{E \in H} |E|$ when $H \neq \varnothing$, and $\mathrm{rank}(\varnothing) = 0$ (so both constant monotone mechanisms have rank zero),

$$\mathfrak{H}_{\le h} = \{H : \mathrm{rank}(H) \le h\}, \qquad \Omega_{\le h} = \Omega(\mathfrak{H}_{\le h}). \tag{27}$$

Theorem 24. Class-wide bounded-rank recurrent closure. *On the ambient state space $\Omega_{\le h}$, depth-h logical equivalence is a right congruence and equals Nerode equivalence. The exact fixed-horizon update of the full residual state is*

$$\mathcal{R}_H(S \cup \{i\}) = \min_{\subseteq}\{R \setminus \{i\} : R \in \mathcal{R}_H(S)\}. \tag{28}$$

Hence equal depth-h query states in $\Omega_{\le h}$ have equal depth-h successors under every positive intervention.

Proof. If $\mathrm{rank}(H) \le h$, every residual $E \setminus S$ has size at most h, so truncation is lossless: $\mathcal{R}_{H,h}(S) = \mathcal{R}_H(S)$. By Theorem 21, two states in $\Omega_{\le h}$ are depth-h equivalent exactly when they have the same full residual clutter. Equation (28) depends only on that residual clutter, including when the two states come from different mechanisms in $\mathfrak{H}_{\le h}$. Therefore the quotient is a right congruence; Lemma 18 identifies it with Nerode equivalence.

Proposition 25. Closure is class-wide, not true-mechanism-only. *A rank bound on the true mechanism alone does not certify recurrent closure when the ambient hypothesis class contains higher-rank alternatives. For $h = 1$, let*

$$H = \{\{1\}\}, \qquad H' = \{\{1\}, \{2,3\}\}, \qquad S = S' = \varnothing. \tag{29}$$

Then $(F_H, S) \equiv_{\mathcal{L}_1} (F_{H'}, S')$, but after action 2 the two successor states are not $\mathcal{L}_1$-equivalent.

Proof. Initially both truncated residual clutters equal $\{\{1\}\}$. After action 2, the first remains $\{\{1\}\}$, whereas the second becomes $\{\{1\}, \{3\}\}$. Theorem 21 distinguishes the successors.

Thus recurrent closure is a property of the *whole state space represented by the learner*. If mechanism identity is uncertain, a structural certificate must cover the hypothesis class (or provide some other reason that cross-mechanism collisions cannot occur).

Proposition 26. Positive-query closure does not survive resets. *Keep the query language $\mathcal{L}_h$ positive, but extend the* deployment action alphabet *with last-write-wins reset actions $\mathrm{reset}_i : S \mapsto S \setminus \{i\}$; then the positive-intervention residual state need not be recurrent even at rank one. For $H = \{\{1\}, \{2\}\}$, the states $S = \{1\}$ and $T = \{2\}$ have the same positive-query residual $\{\varnothing\}$, but reset_1 sends them to states with outputs 0 and 1, respectively.*

This small extension delineates the action scope of Theorem 24: allowing interventions that undo earlier changes can resurrect information discarded by a monotone residual quotient.

Boolean degree versus trigger rank. For orientation, if a monotone Boolean F_H has real multilinear degree at most r, then every minimal trigger has size at most r: the Möbius coefficient of a minimal activating set E is exactly one, so a degree-$|E|$ monomial is present. The converse fails sharply: OR_m has trigger rank one and multilinear degree m. The one-line Möbius calculation is recorded in Supplementary Section S6.6. Classical degree–junta results show how restrictive low real degree is for Boolean functions (Nisan and Szegedy 1994; O'Donnell 2014; Chiarelli et al. 2020); bounded trigger rank is the broader closure condition used above. Supplementary Section S2 records complementary intervention-resource bounds for identifying a hidden trigger, connecting the bounded-query setting to exact monotone-DNF learning and pooled group-testing designs (Abasi et al. 2014; Aldridge et al. 2019).

Lifted recurrent closure under anonymous actions

Remark 17 reduced the anonymous *query* side to elementary equivalence and identified conditions under which it coincides with orbit equivalence. We now settle the *recurrent* side for monotone mechanisms. Without names, the deployment action "activate some control" is nondeterministic at the level of query classes, so a Nerode right congruence is the wrong target. The right notion is a bisimulation. It is decided by the 1-types of the next response structure, and for finite residual cores it reduces to a finite computation on one hidden layer. Moore's single criterion $\approx_h = \approx_{h+1}$ splits into three inequivalent conditions, and in the finitary countable case even bounded trigger rank no longer guarantees closure.

Setting. Fix a pure domain D, finite or countably infinite, and a class $\mathfrak{F}$ of monotone mechanisms on finite subsets of D. Put $\Omega^{\mathrm{anon}}(\mathfrak{F}) = \{(F, S) : F \in \mathfrak{F},\ S \subseteq D \text{ finite}\}$. For $\omega = (F, S)$ and $a \in D$ write $\omega_a = (F, S \cup \{a\})$; *anonymous activation* is the relation $\omega \to \omega_a$ ($a \in D$), and the observed output is Y. For x not free in ψ, the formula

$$\Diamond\psi := \exists x\, [\mathrm{do}(X(x) = 1)]\psi$$

has intervention depth one and quantifier rank one, and its Kripke relation is exactly $\to$; thus the anonymous action system is the frame of a modality already definable in $\mathcal{L}^{\mathrm{anon}}$. With $\mathcal{H}_F$ the minimal triggers of F, write

$$\mathcal{K}_h(\omega) = \mathcal{R}_{\mathcal{H}_F,h}(S), \qquad \mathcal{L}_{h+1}(\omega) = \{Q \in \mathcal{R}_{\mathcal{H}_F,h+1}(S) : |Q| = h + 1\},$$

for the *visible core* and the *hidden layer*, $U_{h+1}(\omega) = \bigcup(\mathcal{K}_h(\omega) \cup \mathcal{L}_{h+1}(\omega))$, and $\mathrm{lk}_a\, \mathcal{L} = \{Q \setminus \{a\} : a \in Q \in \mathcal{L}\}$. We say ω has a *finite* $(h+1)$*-core* when $\mathcal{R}_{\mathcal{H}_F,h+1}(S)$ is finite. By Lemma 13(b), $\mathcal{K}_h(\omega)$ is the family of members of $\mathcal{R}_{\mathcal{H}_F,h+1}(S)$ of size at most h.

Definition 27. Three forms of anonymous recurrent closure. Let $h \geq 0$ and consider $\equiv_h$ on $\Omega^{\mathrm{anon}}(\mathfrak{F})$.

(i) $\equiv_h$ is *anonymously closed* if it is a bisimulation for $\to$: whenever $\omega \equiv_h \omega'$ and $a \in D$, some $b \in D$ satisfies $\omega_a \equiv_h \omega'_b$.

(ii) $\equiv_h$ is *stabilized* if $\equiv_h = \equiv_{h+1}$ on $\Omega^{\mathrm{anon}}(\mathfrak{F})$.

(iii) $\equiv_h$ is *type-deterministic* if for all ω, ω', every isomorphism $\sigma : \mathfrak{M}_h(\omega) \cong \mathfrak{M}_h(\omega')$ and every $a \in D$, $\omega_a \equiv_h \omega'_{\sigma(a)}$.

The *successor spectrum* is $\Sigma_h(\omega) = \{[\omega_a]_{\equiv_h} : a \in D\}$. Since $\equiv_h$ is an equivalence relation, (i) holds exactly when Σ_h is constant on $\equiv_h$-classes.

When $\equiv_h$ coincides with orbit equivalence on the ambient space—finite D (Theorem 6), finite h-cores (Theorem 14), or thresholds (Theorem 12)—(iii) says that the successor class depends only on the current class and on the $\mathrm{Aut}(\mathfrak{M}_h(\omega))$-orbit of the activated control. On a finite structure these orbits are exactly the complete 1-types.

Type-determinism is therefore the precise sense in which a *lifted* action "activate a control of type p" is deterministic; Remark 11 is the case in which every inactive control has the same type.

Proposition 28. Anonymous closure is exact nondeterministic abstraction. *Let $e_h : \Omega^{\mathrm{anon}}(\mathfrak{F}) \to Z_h := \Omega^{\mathrm{anon}}(\mathfrak{F})/{\equiv_h}$ be the quotient map, and let $[\omega] \Rightarrow_h [\omega']$ hold when $\tilde{\omega} \to \tilde{\omega}'$ for some $\tilde{\omega} \equiv_h \omega$ and $\tilde{\omega}' \equiv_h \omega'$. The following are equivalent:*

(i) $\equiv_h$ is anonymously closed;

(ii) e_h is a bounded morphism onto $(Z_h, \Rightarrow_h, \widehat{Y})$: $\omega \to \omega'$ implies $e_h\omega \Rightarrow_h e_h\omega'$, and $e_h\omega \Rightarrow_h z'$ implies $e_h(\omega_a) = z'$ for some a;

(iii) some output-preserving bounded morphism onto some transition system has kernel $\equiv_h$.

In that case $\Rightarrow_h$ is the unique relation on Z_h making e_h a bounded morphism, $\Rightarrow_h$ relates $[\omega]$ exactly to the members of $\Sigma_h(\omega)$, and $\equiv_h$-equivalent states satisfy the same formulas of the modal μ-calculus over $\Diamond$ and Y, of every modal depth (Blackburn et al. 2001; Bradfield and Stirling 2007).

Proof sketch. This is the bounded-morphism/bisimulation correspondence specialized to the quotient map: the forth clause is induced by concrete activation, the back clause is exactly anonymous closure, and kernels of bounded morphisms are bisimulations. Uniqueness of the abstract relation follows from forth and back. Full details are in Supplementary Section S6.8. Thus the anonymous counterpart of the do-commuting square is a *set-valued* abstract updater. When it exists, it answers every branching-time question about anonymous intervention plans of any length—for instance whether Y is reachable within k activations ($\Diamond^{\le k} Y$) or unavoidable—from the depth-h class alone.

Lemma 29. Successor interpretation. *For every state ω, $a \in D$, $j \le h$ and $\bar{b} \in D^j$,*

$$\mathfrak{M}_h(\omega_a) \models R_j(\bar{b}) \iff \mathfrak{M}_{h+1}(\omega) \models R_{j+1}(a, \bar{b}).$$

Hence every σ_h-sentence θ of rank q has a σ_{h+1}-formula $\theta^{\langle x\rangle}(x)$ of rank q with $\mathfrak{M}_h(\omega_a) \models \theta$ iff $\mathfrak{M}_{h+1}(\omega) \models \theta^{\langle x\rangle}(a)$. Consequently $[\omega_a]_{\equiv_h}$ is determined by the complete 1-type $\mathrm{tp}(a)$ of a in $\mathfrak{M}_{h+1}(\omega)$, through $\pi(p) := \{\theta : \theta^{\langle x\rangle} \in p\}$.

Proof. Both sides say $F(S \cup \{a\} \cup \{\bar{b}\}) = 1$. Obtain $\theta^{\langle x\rangle}$ by replacing each atom $R_j(\bar{t})$ with $R_{j+1}(x, \bar{t})$ after renaming bound occurrences of x. Corollary 5 identifies $[\omega_a]_{\equiv_h}$ with the complete theory of $\mathfrak{M}_h(\omega_a)$, which is $\pi(\mathrm{tp}(a))$.

Theorem 30. Type criterion and stabilization transfer. *Let $S_1(\mathfrak{M})$ denote the set of complete 1-types realized in $\mathfrak{M}$.*

(a) $\equiv_h$ is anonymously closed iff $\pi[S_1(\mathfrak{M}_{h+1}(\omega))] = \pi[S_1(\mathfrak{M}_{h+1}(\omega'))]$ whenever $\omega \equiv_h \omega'$.

(b) Suppose $\equiv_h$ is stabilized and (TR) *any two $\equiv_{h+1}$-equivalent states of the ambient space realize the same complete 1-types in their depth-$(h+1)$ response structures. Then $\equiv_h$ is anonymously closed. (TR) holds whenever $\equiv_{h+1}$-equivalent structures are isomorphic. This is the case (1) for finite D, (2) when every state has a finite $(h+1)$-core, by Theorem 14 at depth $h+1$, and (3) for threshold mechanisms, by the proof of Theorem 12(i). In cases (1) and (2) the matching is uniform: any permutation σ with $\sigma[\mathcal{R}_{\mathcal{H}_F,h+1}(S)] = \mathcal{R}_{\mathcal{H}_G,h+1}(S')$ satisfies $\omega_a \equiv_h \omega'_{\sigma(a)}$ for every a.*

(c) In $\mathcal{L}^{\mathrm{anon}}_{h,\omega_1\omega}$, where equivalence is orbit equivalence (Theorem 15), stabilization alone implies anonymous closure, for arbitrary mechanisms on countable D.

Proof. (a) is Definition 27 rewritten with Lemma 29. (b) If $\omega \equiv_h \omega'$, stabilization gives $\mathfrak{M}_{h+1}(\omega) \equiv \mathfrak{M}_{h+1}(\omega')$. By (TR), $\mathrm{tp}(a)$ is realized by some b in $\mathfrak{M}_{h+1}(\omega')$, and Lemma 29 gives $\omega_a \equiv_h \omega'_b$. For the uniform clause, an isomorphism of the depth-$(h+1)$ response structures carries a to $\sigma(a)$; Lemma 29 then restricts it to an isomorphism $\mathfrak{M}_h(\omega_a) \cong \mathfrak{M}_h(\omega'_{\sigma(a)})$, and Corollary 5 gives $\omega_a \equiv_h \omega'_{\sigma(a)}$. (c) An isomorphism $\sigma : \mathfrak{M}_{h+1}(\omega) \cong \mathfrak{M}_{h+1}(\omega')$ restricts, via Lemma 29, to an isomorphism $\mathfrak{M}_h(\omega_a) \cong \mathfrak{M}_h(\omega'_{\sigma(a)})$.

Theorem 31. Finite residual cores: one hidden layer decides closure. *Let F be monotone, $\omega = (F, S)$, $\mathcal{K} = \mathcal{K}_h(\omega)$ and $\mathcal{L} = \mathcal{L}_{h+1}(\omega)$. For every $a \in D$,*

$$\mathcal{K}_h(\omega_a) = \min_{\subseteq}\{K \setminus \{a\} : K \in \mathcal{K}\} \sqcup \mathrm{lk}_a\,\mathcal{L}, \tag{30}$$

and no member of either family contains a member of the other. In particular $\mathcal{K}_h(\omega_a) = \mathcal{K}$ for $a \notin U_{h+1}(\omega)$, which includes every $a \in S$. If every state of the ambient space has a finite $(h+1)$-core, then

$$\Sigma_h(\omega) = \big\{\big[\min_{\subseteq}\{K \setminus \{a\}\}\big] \sqcup \mathrm{lk}_a\,\mathcal{L}\big] : a \in U_{h+1}(\omega)\big\} \cup \big\{[\mathcal{K}] : U_{h+1}(\omega) \neq D\big\}, \tag{31}$$

where $[\cdot]$ is the $\mathrm{Sym}(D)$-orbit of a finite clutter; by Theorem 14 (and trivially for $h = 0$) this is exactly the $\equiv_h$-class. Anonymous closure is therefore decided by finitely many isomorphism tests between finite hypergraphs computed from depth-$(h+1)$ cores; no deeper intervention is needed. In particular, $|\Sigma_h(\omega)| \le |U_{h+1}(\omega)| + 1$, so once the finite core is known the number of successor types is independent of the ambient domain size. If $\mathfrak{F}$ is finite and each mechanism has finitely many minimal triggers, only finitely many orbit pairs $([\mathcal{K}_h], [\mathcal{K}_{h+1}])$ occur, so closure is decidable.

Proof. The proof of Theorem 21 is purely set-theoretic and applies verbatim on any pure domain; at horizon $h+1$ it gives $\mathcal{K}_h(\omega_a) = \{R : |R| \le h,\ R \in \min_{\subseteq}\{Q \setminus \{a\} : Q \in \mathcal{R}_{\mathcal{H}_F,h+1}(S)\}\}$, and $|Q \setminus \{a\}| \le h$ exactly when $Q \in \mathcal{K}$ or $a \in Q \in \mathcal{L}$. No $K \setminus \{a\}$ is contained in $Q \setminus \{a\}$ with $a \in Q \in \mathcal{L}$: that would give $K \subseteq Q$, hence $K \subsetneq Q$ since $|K| \le h < |Q|$, inside the antichain $\mathcal{R}_{\mathcal{H}_F,h+1}(S)$. Conversely $Q \setminus \{a\}$ cannot lie strictly inside $K \setminus \{a\}$, because $|Q \setminus \{a\}| = h \ge |K \setminus \{a\}|$. Distinct members of $\mathcal{L}$ containing

a have distinct links of equal size. This proves (30). A control outside $U_{h+1}(\omega)$ meets no member, and $S \cap U_{h+1}(\omega) = \varnothing$. If $U_{h+1}(\omega) \neq D$, some control outside it contributes $[\mathcal{K}]$; otherwise every control lies in $U_{h+1}(\omega)$. Finally, $\mathcal{R}_{\mathcal{H}_F,h+1}(S)$ depends only on $S \cap \bigcup \mathcal{H}_F$, which takes finitely many values when $\mathcal{H}_F$ is finite.

Corollary 32. Ambient-domain-independent finite-core compilation. *Suppose a depth-$(h+1)$ residual core is given explicitly and its support has size n. Its complete depth-h successor spectrum can be computed without iterating over the ambient domain: evaluate* (30) *for the at most n support elements and, when $D \setminus U_{h+1} \neq \varnothing$, one additional outside-support action type. Canonicalizing each resulting finite cluter by brute-force permutations of its support gives time*

$$O((n+1)!\, Mh \log(M+1)),$$

where $M = |\mathcal{K}_h| + |\mathcal{L}_{h+1}|$, and polynomial space. Hence compiling a finite list of such abstract states and checking equality of their successor spectra has no dependence on $|D|$ once the core and the single outside-support bit are known. The substantive point is this ambient-domain independence: because an explicitly represented n-support core already has size bounded by a function of n, we do not present the displayed bound as a separate parameterized-complexity result.

Proof. Theorem 31 gives at most $n+1$ candidate successors. Represent their members as subsets of an n-element support and choose the lexicographically least incidence list over all $n!$ support permutations as a canonical form. Updating one candidate scans M sets of size at most $h+1$; sorting its canonical incidence list contributes the displayed polynomial factor. Equality of canonical spectra is then ordinary finite-set equality.

Supplementary Section S4 reports an implementation check on random rank-two cores supported on seven controls: embedding the same cores in domains of size 8 through 256 leaves the successor-orbit spectrum unchanged (mean 4.52 successor orbits), while the named action alphabet grows from 8 to 256 and the lifted compiler examines at most eight candidate action types.

Corollary 33. Bounded trigger rank: a domain-size-uniform lifted updater. *Suppose every $F \in \mathfrak{F}$ has rank at most h and every state has a finite residual clutter (for instance, every F has finitely many minimal triggers). Then $\equiv_h$ is type-deterministic, stabilized and anonymously closed, on finite and countable D alike. The abstract updater is*

$$[\mathcal{K}] \Rightarrow_h \left[\min_{\subseteq}\{K \setminus \{a\} : K \in \mathcal{K}\}\right] \quad (a \in \bigcup \mathcal{K}), \qquad [\mathcal{K}] \Rightarrow_h [\mathcal{K}] \quad \textit{if } \bigcup \mathcal{K} \neq D,$$

and the successor depends only on the $\mathrm{Aut}(\mathcal{K})$-orbit of a. On hypergraph isomorphism types, it is the same machine for every domain with more than $|\bigcup \mathcal{K}|$ controls.

Proof. Rank at most h gives $\mathcal{L}_{h+1} = \varnothing$ and $\mathcal{K}_h = \mathcal{K}_{h+1} = \mathcal{R}_{\mathcal{H}_F}(S)$, so $\equiv_h$ is stabilized. An isomorphism σ of depth-h response structures carries $\mathcal{K}$ onto $\mathcal{K}'$, because $\mathcal{K}$ is first-order definable from $\mathfrak{M}_h$ by Lemma 13(b). By (30) it then carries $\min\{K \setminus \{a\}\}$ onto $\min\{K' \setminus \{\sigma a\}\}$, and Theorem 14 turns this into $\omega_a \equiv_h \omega'_{\sigma(a)}$.

Corollary 33 is the anonymous counterpart of Theorem 24. It extends the domain-size uniformity of Theorem 10 from fully symmetric mechanisms to every bounded-rank, finitely triggered mechanism: the lifted machine lives on isomorphism types of residual hypergraphs, and orbits (equivalently, 1-types) of controls are its deterministic action labels.

Theorem 34. The Moore criterion splits.

(a) Closure does not imply stabilization, even for one known mechanism at $h = 1$. *Fix $r \geq 2$ and partition a finite pure domain as $D = A \sqcup B \sqcup C$ with $|A| = r + 1$ and $|B| = |C| = r$. Let F_r be the rank-two monotone mechanism whose trigger graph is the disjoint union $K_A \sqcup K_{B,C}$. On the single-mechanism state space $\Omega^{\mathrm{anon}}(\{F_r\})$, $\equiv_1$ has exactly four classes: three false-output classes Q_0, Q_r, Q_{2r} indexed by $0, r, 2r$ singleton residuals and the true absorbing class T. Their successor spectra are*

$$Q_0 \Rightarrow_1 \{Q_r\}, \qquad Q_r \Rightarrow_1 \{Q_r, Q_{2r}, T\},$$
$$Q_{2r} \Rightarrow_1 \{Q_{2r}, T\}, \qquad T \Rightarrow_1 \{T\}.$$

So $\equiv_1$ is anonymously closed. Yet activating one vertex of A and one vertex of B from the initial state produces $\equiv_1$-equivalent successors in Q_r whose depth-two residual graphs are respectively $K_{r,r}$ and K_{r+1}, so they are not $\equiv_2$-equivalent. For $r = 2$ this is the seven-control graph $K_3 \sqcup C_4$.

(b) Stabilization does not imply type-determinism. *Let F be the conjunction of a set Q with $|Q| = h + 1$, on any $D \supsetneq Q$. Then $\equiv_h$ is stabilized, but for $q \in Q$ and $d \notin Q$ the transposition $(q\, d)$ is an automorphism of $\mathfrak{M}_h(F, \varnothing)$, while $(F, \{q\}) \not\equiv_h (F, \{d\})$.*

(c) Without finite cores the finitary transfer fails, even for bounded rank. *Let D be countable, let G_2 be the disjoint union of countably many copies of every finite star $K_{1,n}$ ($n \geq 1$), let $G_1 = G_2 \sqcup K_{1,\omega}$, and let F_i have trigger clutter $E(G_i)$. Then $(F_1, \varnothing) \equiv_{h'} (F_2, \varnothing)$ for every h', and $\equiv_h = \equiv_{h+1}$ on all of $\Omega^{\mathrm{anon}}(\{F_1, F_2\})$ for every $h \geq 2$, although both mechanisms have rank two. Nevertheless $\equiv_h$ is not anonymously closed for any $h \geq 1$.*

Proof sketch. In (a), depth one sees only the current output and the number of singleton residuals. The displayed four-class successor spectra are obtained by considering whether an action hits the same graph component, the other component, or a completing neighbor. The two Q_r witnesses differ one layer deeper because activating one clique vertex leaves $K_{r,r}$ hidden, whereas activating one bipartite-side vertex leaves K_{r+1} hidden. In (b), every depth-h response relation is empty at the initial state, so a trigger/non-trigger transposition is an automorphism, but the two successors have different residual deficits. In (c), the two star forests are elementarily equivalent, yet activating the infinite-star centre produces infinitely many singleton residuals, a successor type omitted by the other mechanism. Full proofs are in Supplementary Sections S6.7 and S6.9.

Remark 35. Interpretation, minimality, and scope. Parts (a)–(c) separate three questions that coincide for named actions. Part (a) lies entirely inside one known mechanism: mechanism uncertainty is not needed for closure without stabilization. A smaller mixed-rank example exists on six controls. With trigger clutter

$$\mathcal{H} = \{\{0,2\},\{0,3\},\{0,4\},\{1,3\},\{3,5\},\{1,2,5\},\{1,4,5\}\},$$

the states $S = \{1\}$ and $S' = \{2\}$ each have one singleton residual and hence are $\equiv_1$-equivalent; after removing that singleton, their depth-two residual graphs are respectively a four-cycle and a triangle, so they are not $\equiv_2$-equivalent. Direct enumeration gives four depth-one classes and verifies anonymous closure. The exhaustive archive rules out every single-mechanism example on one through five controls for every positive horizon $h \geq 1$, so six controls is minimal at positive horizon; the rank-two family in part (a) is cleaner and begins at seven controls. This qualifier is essential: at $h = 0$, the three-control mechanism $F = X_3 \vee (X_1 \wedge X_2)$ is anonymously closed but not stabilized. An anonymously closed interface can therefore be exact for lifted one-step planning while still omitting information exposed by one deeper query.

Part (c) is a type-omission phenomenon: the centre of $K_{1,\omega}$ realizes $\{\exists^{\geq n} y\, R_2(x,y) : n \in \mathbb{N}\}$, which $\mathfrak{M}_2(F_2, \varnothing)$ omits, so hypothesis (TR) of Theorem 30 fails. It mirrors Proposition 16 on the recurrent side and disappears in $\mathcal{L}_{\omega_1\omega}$, where the two initial states are already inequivalent. For thresholds all three notions coincide and hold iff $h \geq k - 1$, recovering Theorems 10 and 12(iii).

Definition 27 is set-valued because ordinary $\Diamond$ records possible successor classes, not their multiplicities. A stronger *graded* closure condition requires the successor multiplicity profile to be constant on each $\equiv_h$-class. The part-(a) witnesses fail it: in $K_3 \sqcup C_4$ the two Q_r witnesses have sorted successor-count multiset $(1,2,4)$ and $(2,2,3)$ (the six-control profiles are given in Supplementary Section S6.9). On a finite domain, if anonymous activation chooses a control uniformly at random, this graded condition is exactly strong lumpability of the induced Markov chain: successor-block probabilities are the corresponding counts divided by $|D|$ (Kemeny and Snell 1976). In modal terms it is the natural closure notion seen by graded, or counting, modalities (Fine 1972; de Rijke 2000). Thus Theorem 34(a) separates stabilization from ordinary modal closure, not from counting-sensitive closure. Our exhaustive scan finds no graded-closure-without-stabilization example through five controls; whether graded anonymous closure implies stabilization on every finite domain remains open.

The archive checks (30) on all $6{,}107{,}808$ four-/five-control cases, scans every single mechanism through five controls, record the three-control $h = 0$ exception, and verifies both positive-horizon witnesses. No graded-closure-without-stabilization case occurs through five controls, including all $1{,}050$ mechanism–horizon cases at $m = 5$. Whether graded closure or type-determinism implies stabilisation remains open.

Secondary quantitative extensions

The recurrent-closure question also has a quantitative side that is useful but secondary to the lifted logic results above. Supplementary Section S6.3 states and proves the

real-valued degree-closure and minimax theorem: for multilinear degree at most r, live Möbius coefficients form an exactly closed recurrent state, and extrapolating from all order-$\leq r$ measurements to a target adding $t > r$ controls has exact minimax sup-norm radius

$$\varepsilon\Lambda_{t,r}, \qquad \Lambda_{t,r} = \sum_{j=0}^{r} \binom{t}{j}\binom{t-j-1}{r-j}.$$

The Möbius estimator attains this radius, and the same lower bound holds inside the globally $[0,1]$-valued class when $m' = |R_S| > r$ and $\varepsilon\Lambda_{m',r} \leq \frac{1}{2}$. This operator norm is analogous to a Lebesgue constant in interpolation (Ibrahimoglu 2016); approximate inclusion–exclusion addresses a related but different low-order-information problem (Linial and Nisan 1990; Kahn et al. 1996). In particular $\Lambda_{t,1} = 2t - 1$ and $\Lambda_{t,2} = 2t^2 - 4t + 1$; for $r = 2$, $t = 20$, and $\varepsilon = 10^{-3}$ the exact radius is 0.721.

A fresh order-r panel after physically executing part of the target changes the radius according to the remaining target size. The fresh-panel corollary in Supplementary Section S6.3 gives the corresponding radii 0.391, 0.161, 0.031, and 0.007 after 5, 10, 15, and 17 of 20 target actions; Supplementary Section S3 tabulates representative degree-two conditioning values. Supplementary Section S6.4 separately states and proves the Waserstein–Lipschitz semantic propagation bound. These results support the distinction among exact structural closure, numerical conditioning, and semantic adequacy; we make no additional lower-floor claim for non-congruent abstractions here.

A finite-sample closure certificate

The Moore/Nerode criterion is exact but, for a learned or simulator-backed interface, builders need a test that can be run from transitions. Let $e : \Omega \to Z$ be any finite abstract interface for a deterministic system with action set A. For each class z, choose a sampling distribution μ_z supported on the fiber $e^{-1}(z)$. Under action a, define the successor-class distribution

$$P_{z,a}(z') = \Pr_{\omega\sim\mu_z}\{e(f_a(\omega)) = z'\} \tag{32}$$

and its impurity $\eta_{z,a} = 1 - \max_{z'} P_{z,a}(z')$. Exact congruence is equivalent to $\eta_{z,a} = 0$ for every pair when each μ_z has full support.

Theorem 36. One-sided finite-sample closure certificate. *Let* $Z_{\mathrm{ns}} = \{z \in Z : |e^{-1}(z)| > 1\}$ *and* $K_* = |Z_{\mathrm{ns}}||A|$. *Singleton fibers are exactly closed and need not be sampled. For each tested pair* (z, a), *draw a fresh i.i.d. sample of* n *concrete states from* μ_z, *and reject closure if any pair produces two different abstract successor classes. The same union bound also applies if one i.i.d. sample from* μ_z *is shared across actions; independence is needed within each tested pair, not between pairs. Fix* $\beta, \rho \in (0, 1)$. *Then*

$$\Pr(\textit{test passes and } \exists(z,a) : \eta_{z,a} \geq \beta) \leq K_*(1-\beta)^{n-1}. \tag{33}$$

Consequently, if $K_*(1-\beta)^{n-1} \leq \rho$, *the probability of falsely passing an interface that contains any* β*-heavy violation is at most* ρ.

If, additionally, every concrete state in every non-singleton fiber has sampling mass at least $\alpha > 0$, then any exact non-congruence has $\eta_{z,a} \geq \alpha$. Setting $\beta = \alpha$ therefore bounds by ρ the probability that the test passes despite an exact non-congruence. Under uniform sampling from non-singleton fibers of size at most $M_{\max}$, one may take $\alpha = 1/M_{\max}$.

Proof sketch. For a tested pair (z, a) with successor probabilities p_j, the probability that n sampled successors all agree is $\sum_j p_j^n \leq (\max_j p_j)^{n-1} = (1 - \eta_{z,a})^{n-1}$. Singleton fibers are automatically closed, so a union bound over only K_* nontrivial class–action pairs gives the stated result; the α-mass clause turns exact non-congruence into a lower bound on impurity. Full details are given in Supplementary Section S6.5.

The refined factor K_* presumes that non-singleton fibers are identifiable under the declared fiber-sampling access; for a learned encoder, singleton status must therefore be verified from the same concrete-state access used to construct the sampler.

A practical horizon selector applies the test to $h = 0, 1, \ldots, H$ with false-pass budget $\rho/(H + 1)$ and returns the first level that passes. The guarantee is then simultaneous over all examined horizons. Without a lower bound on fiber mass this certifies absence of β-*heavy* violations, not exact congruence; rare collisions can remain unseen. This distinction is important for learned state spaces. The test certifies *closure of the supplied partition*, not query adequacy: a constant encoder and an injective encoder are both trivially closed. In practice it must therefore be paired with an adequacy check against the intended query quotient (or direct query loss). The experiment below reports both partition agreement and closure.

Neurosymbolic recurrent experiments

The theory distinguishes four questions that learned systems can conflate: does the interface answer the supervised queries (adequacy), does its quotient survive deployment actions (closure), does perception preserve the required state, and can a recurrent optimizer exploit the retained information? We test these separately on an eight-control lever–door task. The two mechanisms are the eight-way conjunction and the rank-two clutter

$$(X_1 \wedge X_2) \vee (X_3 \wedge X_4) \vee X_5 \vee (X_6 \wedge X_7),$$

with X_8 irrelevant. Exact depth-two query states have 38 and 28 classes, respectively. Unless stated otherwise, neural results use five seeds, 8,000 sampled training rollouts, 35 updater epochs, and 12,000 fresh evaluation rollouts per seed, with no output feedback. Because $m = 8$ has only 256 concrete states, "fresh" here means newly sampled action sequences (and, for Digits, newly sampled held-out glyphs), not generalisation to unseen states. Whenever we exhaustively form a learned partition, the encoder is evaluated on canonical noiseless scalar renders; noisy perceived panels are analyzed separately in the mechanism-learning study.

Query-only supervision and retained encoder detail. The scalar substrate exposes eight sensors, each directly encoding one control bit, so we treat the decision probe as a diagnostic of *non-factoring*, not as evidence that query supervision creates extra state information. The 219 conjunction far-class states are split by state into 153 training and 66 held-out states per seed; held-out states are excluded from encoder training, and pairs are balanced within action. After training-split standardization, the query-trained encoder reaches 82.11% balanced accuracy ($[77.41, 86.82]$; AUROC 0.906), while the same randomly initialized but never-trained architecture reaches 78.96% ($[72.26, 85.67]$; AUROC 0.925). Per-seed differences are not uniformly signed, so query supervision neither clearly creates nor measurably removes the action-relevant detail carried by this architecture. Within the depth-two far class, the probe target is exactly the one-step distinction exposed by depth three: the depth-two quotient has analytic balanced accuracy 50%, the depth-three quotient 100%, and a raw-sensor probe reaches 94.11%. These diagnostics show that a soft neural implementation need not factor through the exact quotient; they do not show that the quotient itself is inadequate. Supplementary Section S1 reports pre-sigmoid and unstandardized controls and the per-seed values used for these diagnostics.

Recurrent compression with scalar and image perception. Table 2 makes the structural comparison on conjunction. Its eight-step scalar soft model is trained separately from the three-step optimized-soft run reported in Suplementary Table S1; the two curves are distinct fitted models. The exact depth-two Bayes curve is computed over the uniform prior on all 256 concrete states and i.i.d. uniform named actions. An exact-V_2 GRU follows that curve closely at early steps and then falls below it from optimization error.* The scalar soft GRU exceeds the quotient curve at later steps, consistent with the probe evidence above, but its input architecture makes such retention easy rather than surprising.

We repeat the recurrent pipeline on the public scikit-learn Digits substrate (Pedregosa et al. 2011; Alpaydin and Kaynak 1998): each control is a held-out 8×8 glyph whose parity encodes the bit, a shared glyph MLP is trained jointly with the depth-two query head, and frozen query probabilities initialize the 64-unit GRU. Exact query-tensor accuracy is 91.76% [89.35,94.17] for conjunction and 89.97% [88.60,91.35] for rank two. To separate state information from updater class, we feed the *same encoder's* eight sigmoid bit-head probabilities to the same 64-unit GRU/action updater. This bit-head GRU reaches $92.51/90.86/89.49\%$ at $k = 1/2/3$ and 90.81% at $k = 8$, compared with $86.33/77.75/68.08\%$ and 39.44% for the depth-two query-head GRU. Feeding the same bit head to an exact symbolic automaton gives $92.31/90.63/89.29/90.90\%$, and propagating Bernoulli bit beliefs is nearly identical ($92.45/90.72/89.34/90.90\%$). Thus the large long-horizon gap is primarily a *state-information* effect: once recurrence receives the richer bit state, neural and symbolic transition mechanisms perform similarly on this task. Complete curves and seed-level intervals are in Supplementary Section S1.

*At $k = 1$ the finite 12,000-rollout estimate is 91.87%, 0.07 percentage points above the analytic 91.80% Bayes rate. This is ordinary evaluation-sample fluctuation, not a true improvement over the Bayes optimum.

Table 2. Eight-step conjunction experiment: five-seed mean class accuracy. The two rightmost controls use the same jointly trained Digits panel encoder and identical evaluation rollouts; the bit-head GRU and exact automaton both receive the encoder's eight-bit state estimate. Intervals and the probabilistic automaton are in S1.

k	Exact-V_2 Bayes	Exact V_2 GRU	Scalar soft GRU	Digits query-GRU	Digits bit-GRU	Digits exact aut.
1	91.80	91.87	90.55	86.33	92.51	92.31
2	81.54	81.72	82.77	77.75	90.86	90.63
3	70.01	70.10	76.11	68.08	89.49	89.29
4	58.31	58.50	69.16	58.27	88.80	88.59
5	47.56	47.50	61.96	48.73	88.75	88.54
6	40.37	39.96	56.98	42.33	89.28	89.11
7	37.21	36.43	54.53	39.50	90.03	90.01
8	37.25	35.67	54.15	39.44	90.81	90.90

Learning the mechanism class: adequacy versus closure. For a rank-≤ 2 hypothesis class, the complete order-two intervention panel has $1 + 8 + \binom{8}{2} = 37$ probes. Across 2,500 randomly generated clutters, exact panel observations recover the mechanism in all cases and the learned residual updater makes zero errors in 250,000 subsequent transition checks. Enlarging the ambient class to permit hidden rank-three triggers creates observationally indistinguishable alternatives for 99.4% of mechanisms, with 19.88 mean alternatives per mechanism; every sampled hidden alternative yields the cross-mechanism collision predicted by Proposition 25. Under noisy perceived panels, recovery degrades while closure does not: with the manuscript's Digits parity grounder, one, three and five independent readings give 51.36%, 94.72% and 99.68% exact clutter recovery, while depth-two successor-query error falls from 7.39% to 0.76% to 0.05%. The learned residual partition has zero measured closure violations in every condition, exactly as Theorem 24 predicts. We treat those zeros explicitly as a theorem-predicted implementation sanity check, not as empirical evidence for closure; the informative quantity is adequacy error. Full five-seed tables are in Supplementary Section S5.

Sampling closure rather than enumerating it. For $m = 8$ exhaustive enumeration is cheap, so it supplies ground truth. We also apply Theorem 36 with a non-oracle uniform-within-fiber sampler. With $n = 512$ samples per non-singleton fiber/action pair, all eight exact conjunction violations are detected; the maximum empirical impurity is 0.1211. This is a detection experiment, not a 5% pass-level guarantee: for the exact far fiber the theorem gives $8(218/219)^{511} = 0.772$, and $n = 1110$ is required to reduce that bound to at most 0.05. Repeating the same $n = 512$ procedure on a thresholded learned partition detects all eight of its exact violating pairs with maximum impurity 0.1055. At $m = 20$ an explicitly oracle boundary sampler detects all 20 action violations with mean impurity 0.1520; it is included only as a power demonstration. Supplementary Section S4 reports the scaling details.

What the experiments establish. These runs diagnose distinct failure sources rather than prove the structural theorems. The GNN check realizes the predicted receptive-field boundary; the state-disjoint probe shows architectural non-factoring; the matched Digits

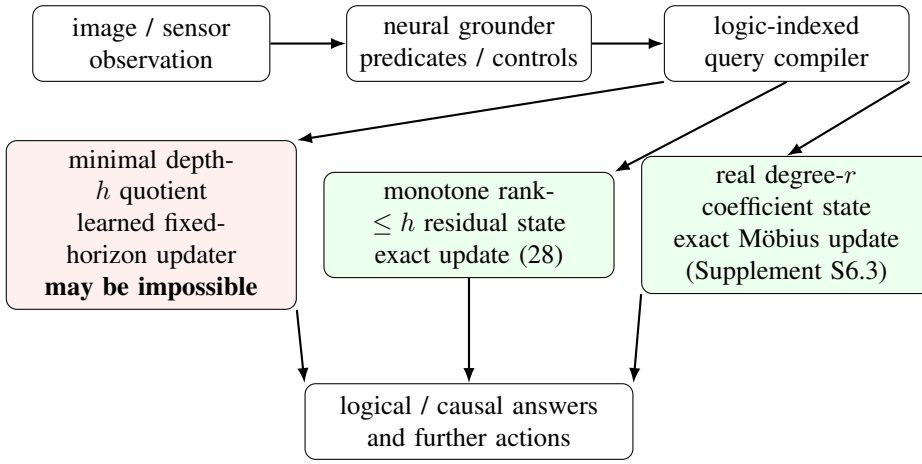


Figure 1. Logic-indexed intervention-interface architecture. Query sufficiency alone does not justify recurrent reuse; either consume the horizon/re-encode, or use a structural state whose quotient is closed under actions.

controls show that replacing the depth-two query state by the richer bit state repairs long-horizon recurrence, while a bit-head GRU and exact automaton perform similarly; and noisy mechanism learning separates adequacy error from theorem-guaranteed closure. Uniform-within-fiber sampling shows that the certificate can find violations without an oracle boundary sampler.

Architecture implied by the theory

Figure 1 summarizes the recommended division of labor. A neural perceptual/query encoder extracts control and query information from unstructured input. A symbolic layer then chooses a state representation whose algebra matches the deployment regime rather than feeding an arbitrary compressed query vector back into a recurrent network.

Three interface rules follow. First, the language must declare what distinctions the representation is intended to preserve. Second, the state module must satisfy a congruence condition for the actions that deployment will apply. Third, numerical approximation must be separated from structural closure: even an exactly closed degree-r state can be numerically fragile (Supplementary Sections S3 and S6.3), while a low-error neural updater cannot recover information erased by a non-congruent quotient.

Discussion

Terminology and neighboring predictive-state traditions

We intentionally avoid calling the quotient a *causal state*. In computational mechanics that term denotes the predictive-equivalence classes of an ϵ-machine (Crutchfield and

Young 1989; Shalizi and Crutchfield 2001). Our object is instead a logic-indexed intervention interface: its equivalence is chosen by a query language and may or may not already coincide with a stable predictive equivalence. This distinction is exactly what Lemma 18 tests.

The response tensor is also a finite deterministic Hankel block. Weighted-automata and spectral-learning work treats Hankel rank as the route to recursively sufficient state (Balle et al. 2014); PSRs encode state by action-conditional future predictions (Littman et al. 2001; Boots et al. 2011). MDP/SMDP homomorphisms impose analogous transition-preservation requirements on abstractions (Ravindran and Barto 2003). The present results specialize these ideas to intervention queries and expose mechanism-level sufficient conditions (bounded monotone trigger rank or real-valued polynomial degree) and measurement robustness.

What a neurosymbolic builder should do differently

Before reusing a query-trained embedding recurrently, first check that it represents the intended query partition and then test whether that partition is closed under deployment actions. Theorem 36 gives a finite-sample one-sided closure test when representative states can be sampled from abstract fibers; Supplementary Sections S1, S4, and S5 illustrate why adequacy and closure must be checked separately and show that a continuous encoder need not factor through the exact quotient; the untrained-encoder control in S1 shows that this can be architectural rather than caused by query supervision. If closure fails, three repairs have different costs: consume horizon and supervise deeper queries; re-encode from a richer external state; or replace the interface by a structural recurrent state such as the residual clutter. Numerical approximation can then be analyzed separately with the quantitative extensions in Supplementary Sections S3, S6.3, and S6.4.

Limitations

The named-control theorems assume persistent positive interventions; last-write-wins resets can invalidate a positive-query residual state (Proposition 26). The anonymous threshold theorem does not solve arbitrary lifted relational model checking, and the finite-sample certificate requires sampling inside abstract fibers; continuous embeddings therefore need a clustering/discretization rule or another test. The lever–door dynamics remain synthetic. The state-disjoint probe is reported only from the archived PyTorch 2.10 reference environment, so its intervals measure seed variation rather than cross-platform uncertainty. The Digits controls use matched perception but compare diferent recurrent state inputs; the bit-head GRU shows that most of the gain comes from the richer state information rather than from symbolic transition mechanics. Neural training and evaluation can vary modestly across CPU/PyTorch builds, so the archived five-seed intervals quantify variation in the reference environment rather than cross-platform uncertainty. These controls are not common-benchmark comparison with NeSyA, DeepDFA, Scallop, or DeepProbLog (Manginas et al. 2025; Umili and Capobianco 2024; Li et al. 2023; Manhaeve et al. 2018); that remains the principal empirical limitation. The supplementary Wasserstein bound assumes uniform primitive error and kernel regularity,

and the discrete state lower bounds are not neural-dimension lower bounds without a precision or robustness model.

Conclusion

A query-trained representation is not automatically a recurrent representation. Named actions require deterministic congruence; anonymous lifted actions require bisimulation. The response-structure translation makes the anonymous query problem first-order, while successor 1-types and finite residual cores connect that logical classification to recurrent reuse. The one-hidden-layer update and its ambient-domain-independent finite-core compilation make this connection constructive rather than merely semantic. Residual clutters provide the corresponding named-action repair, and the finite-sample certificate makes non-closure testable. The experiments reinforce the same separation on local message passing, scalar query interfaces, and held-out handwritten-digit panels: query adequacy, transition closure, perception, and optimization are distinct. For the special-issue theme, the central expressiveness–computation trade-off is therefore not only which formulas can be stated, but which logically sufficient states remain executable after the world changes.

Verification and reproducibility

The exact supplementary and reproducibility package for this version is archived on Zenodo at DOI 10.5281/zenodo.23023084. It contains the supplement, verification scripts, and machine-readable results used for this submission. The verification suite includes 3,000 randomized finite semantics tests of the response-structure translation (including variable rebinding), 477 exhaustive small Ehrenfeucht–Fraïssé game comparisons for the sharp depth-one rank criterion, and independent recomputation of the eight-step exact-V_2 Bayes curve.

Ethical considerations

The lever-door and mechanism-learning data are synthetic. The external perception control uses the public scikit-learn Digits benchmark (Pedregosa et al. 2011; Alpaydin and Kaynak 1998), derived from the UCI Optical Recognition of Handwritten Digits dataset and consisting of previously collected digitized handwriting. This study recruits no participants, collects no new human-subject data, and uses no animals or biological materials.

Supplemental material

The supplementary package contains theorem-verification code, the sampled lever-door experiment, exact-interface controls, learned-interface closure diagnostics, result JSON files, and reproduction instructions.

References

Abasi H, Bshouty NH and Mazzawi H (2014) On exact learning monotone DNF from membership queries. In: *Algorithmic Learning Theory*, LNCS 8776, pp. 111–124.

Aldridge M, Johnson O and Scarlett J (2019) Group testing: an information theory perspective. *Foundations and Trends in Communications and Information Theory* 15(3–4): 196–392. DOI: 10.1561/0100000099.

Alpaydin E and Kaynak C (1998) Optical Recognition of Handwritten Digits [dataset]. UCI Machine Learning Repository. DOI: 10.24432/C50P49.

Angluin D (1987) Learning regular sets from queries and counterexamples. *Information and Computation* 75(2): 87–106. DOI: 10.1016/0890-5401(87)90052-6.

Badreddine S, d'Avila Garcez A, Serafini L and Spranger M (2022) Logic Tensor Networks. *Artificial Intelligence* 303: 103649. DOI: 10.1016/j.artint.2021.103649.

Balle B, Carreras X, Luque FM and Quattoni A (2014) Spectral learning of weighted automata: a forward-backward perspective. *Machine Learning* 96(1–2): 33–63. DOI: 10.1007/s10994-013-5416-x.

Banihashemi B, De Giacomo G and Lespérance Y (2017) Abstraction in situation calculus action theories. *Proceedings of the AAAI Conference on Artificial Intelligence* 31(1). DOI: 10.1609/aaai.v31i1.10693.

Barbero F and Sandu G (2021) Team semantics for interventionist counterfactuals: observations vs. interventions. *Journal of Philosophical Logic* 50(3): 471–521. DOI: 10.1007/s10992-020-09573-6.

Barbero F and Sandu G (2024) Multiteam semantics for interventionist counterfactuals: probabilities and causation. *Journal of Philosophical Logic* 53: 1537–1577. DOI: 10.1007/s10992-024-09777-0.

Barbero F, Schulz K, Smets S, Velázquez-Quesada FR and Xie K (2020) Thinking about causation: a causal language with epistemic operators. In: Martins M and Sedlár I (eds) *Dynamic Logic. New Trends and Applications*, LNCS 12569. Springer, pp. 17–32. DOI: 10.1007/978-3-030-65840-3_2.

Barceló P, Kostylev EV, Monet M, Pérez J, Reutter JL and Silva JP (2020) The logical expressiveness of graph neural networks. In: *International Conference on Learning Representations*.

Bareinboim E, Correa JD, Ibeling D and Icard T (2022) On Pearl's hierarchy and the foundations of causal inference. In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. ACM, pp. 507–556. DOI: 10.1145/3501714.3501743.

Beckers S and Halpern JY (2019) Abstracting causal models. In: *AAAI Conference on Artificial Intelligence*, 33, pp. 2678–2685.

Blackburn P, de Rijke M and Venema Y (2001) *Modal Logic*. Cambridge: Cambridge University Press.

Boots B, Siddiqi SM and Gordon GJ (2011) Closing the learning-planning loop with predictive state representations. *International Journal of Robotics Research* 30(7): 954–966.

Bradfield J and Stirling C (2007) Modal mu-calculi. In: Blackburn P, van Benthem J and Wolter F (eds) *Handbook of Modal Logic*. Amsterdam: Elsevier, pp. 721–756.

Briggs R (2012) Interventionist counterfactuals. *Philosophical Studies* 160(1): 139–166. DOI: 10.1007/s11098-012-9908-5.

Chiarelli J, Hatami P and Saks M (2020) An asymptotically tight bound on the number of relevant variables in a bounded degree Boolean function. *Combinatorica* 40(2): 237–244. DOI: 10.1007/s00493-019-4136-7.

Crutchfield JP and Young K (1989) Inferring statistical complexity. *Physical Review Letters* 63(2): 105–108. DOI: 10.1103/PhysRevLett.63.105.

de Rijke M (2000) A note on graded modal logic. *Studia Logica* 64(2): 271–283. DOI: 10.1023/A:1005245900406.

Ferns N, Panangaden P and Precup D (2004) Metrics for finite Markov decision processes. In: *Uncertainty in Artificial Intelligence*, pp. 162–169.

Fine K (1972) In so many possible worlds. *Notre Dame Journal of Formal Logic* 13(4): 516–520. DOI: 10.1305/ndjfl/1093890715.

Galles D and Pearl J (1998) An axiomatic characterization of causal counterfactuals. *Foundations of Science* 3(1): 151–182. DOI: 10.1023/A:1009602825894.

Geiger A, Ibeling D, Zur A, Chaudhary M, Chauhan S, Huang J, Arora A, Wu Z, Goodman N, Potts C and Icard T (2025) Causal abstraction: a theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research* 26(83): 1–64.

Givan R, Dean T and Greig M (2003) Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence* 147(1–2): 163–223.

Grohe M (2021) The logic of graph neural networks. In: *36th Annual ACM/IEEE Symposium on Logic in Computer Science*, pp. 1–17. DOI: 10.1109/LICS52264.2021.9470677.

Halpern JY and Peters S (2022) Reasoning about causal models with infinitely many variables. *Proceedings of the AAAI Conference on Artificial Intelligence* 36(5): 5668–5675. DOI: 10.1609/aaai.v36i5.20508.

Halpern JY (2000) Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research* 12: 317–337.

Hodges W (1993) *Model Theory*. Cambridge: Cambridge University Press.

Hofmann T and Belle V (2023) Abstracting noisy robot programs. In: *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, pp. 534–542.

Ibeling D and Icard T (2020) On open-universe causal reasoning. In: *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, PMLR 115, pp. 1233–1243.

Ibrahimoglu BA (2016) Lebesgue functions and Lebesgue constants in polynomial interpolation. *Journal of Inequalities and Applications* 2016: 93. DOI: 10.1186/s13660-016-1030-3.

Kahn J, Linial N and Samorodnitsky A (1996) Inclusion-exclusion: exact and approximate. *Combinatorica* 16(4): 465–477. DOI: 10.1007/BF01271266.

Kemeny JG and Snell JL (1976) *Finite Markov Chains*. New York: Springer.

Li L, Walsh TJ and Littman ML (2006) Towards a unified theory of state abstraction for MDPs. In: *International Symposium on Artificial Intelligence and Mathematics*.

Li Z, Huang J and Naik M (2023) Scallop: a language for neurosymbolic programming. *Proceedings of the ACM on Programming Languages* 7(PLDI): 1463–1487. DOI: 10.1145/3591280.

Libkin L (2004) *Elements of Finite Model Theory*. Berlin: Springer.

Linial N and Nisan N (1990) Approximate inclusion-exclusion. *Combinatorica* 10(4): 349–365. DOI: 10.1007/BF02128670.

Littman ML, Sutton RS and Singh S (2001) Predictive representations of state. In: *Advances in Neural Information Processing Systems 14*, pp. 1555–1561.

Lorenz R and Tull S (2026) Causal and compositional abstraction. arXiv:2602.16612. DOI: 10.48550/arXiv.2602.16612.

Loukas A (2020) What graph neural networks cannot learn: depth vs width. In: *International Conference on Learning Representations*.

Manginas N, Paliouras G and De Raedt L (2025) NeSyA: Neurosymbolic Automata. In: *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 5950–5958. DOI: 10.24963/ijcai.2025/662.

Manhaeve R, Dumančić S, Kimmig A, Demeester T and De Raedt L (2018) DeepProbLog: neural probabilistic logic programming. In: *Advances in Neural Information Processing Systems*.

Moore EF (1956) Gedanken-experiments on sequential machines. In: Shannon CE and McCarthy J (eds) *Automata Studies*. Princeton, NJ: Princeton University Press, pp. 129–153.

Mossé M, Ibeling D and Icard T (2024) Is causal reasoning harder than probabilistic reasoning? *Review of Symbolic Logic* 17(1): 106–131. DOI: 10.1017/S1755020322000211.

Nerode A (1958) Linear automaton transformations. *Proceedings of the American Mathematical Society* 9(4): 541–544.

Niepert M and Van den Broeck G (2014) Tractability through exchangeability: a new perspective on efficient probabilistic inference. In: *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pp. 2467–2475. DOI: 10.1609/AAAI.V28I1.9073.

Nisan N and Szegedy M (1994) On the degree of Boolean functions as real polynomials. *Computational Complexity* 4(4): 301–313.

O'Donnell R (2014) *Analysis of Boolean Functions*. Cambridge: Cambridge University Press.

Pearl J (2009) *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge: Cambridge University Press.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay E (2011) Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Ravindran B and Barto AG (2003) SMDP homomorphisms: an algebraic approach to abstraction in semi-Markov decision processes. In: *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pp. 1011–1018.

Rischel EF and Weichwald S (2021) Compositional abstraction error and a category of causal models. In: *Uncertainty in Artificial Intelligence*, PMLR 161, pp. 1013–1023.

Rubenstein PK, Weichwald S, Bongers S, Mooij JM, Janzing D, Grosse-Wentrup M and Schölkopf B (2017) Causal consistency of structural equation models. In: *Uncertainty in Artificial Intelligence*.

Scott D (1965) Logic with denumerably long formulas and finite strings of quantifiers. In: Addison JW, Henkin L and Tarski A (eds) *The Theory of Models*. Amsterdam: North-Holland, pp. 329–341.

Shalizi CR and Crutchfield JP (2001) Computational mechanics: pattern and prediction, structure and simplicity. *Journal of Statistical Physics* 104: 817–879. DOI: 10.1023/A:1010388907793.

Toro Icarte R, Klassen TQ, Valenzano R and McIlraith SA (2022) Reward machines: exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research* 73: 173–208. DOI: 10.1613/jair.1.12440.

Umili E and Capobianco R (2024) DeepDFA: automata learning through neural probabilistic relaxations. In: *ECAI 2024*, Frontiers in Artificial Intelligence and Applications 392, pp. 1051–1058. DOI: 10.3233/FAIA240596.

Umili E, Argenziano F and Capobianco R (2026) DeepDFA: injecting temporal logic in deep learning for sequential subsymbolic applications. arXiv:2602.03486. DOI: 10.48550/arXiv.2602.03486.

van Bremen T and Kuželka O (2023) Lifted inference with tree axioms. *Artificial Intelligence* 324: 103997. DOI: 10.1016/j.artint.2023.103997.

Wang Y, Pu J, Wang Y and Kuželka O (2024) Lifted algorithms for symmetric weighted first-order model sampling. *Artificial Intelligence* 331: 104114. DOI: 10.1016/j.artint.2024.104114.

Xia K and Bareinboim E (2024) Neural causal abstractions. In: *AAAI Conference on Artificial Intelligence* 38(18): 20585–20595. DOI: 10.1609/aaai.v38i18.30044.

Xia K, Lee K-Z, Bengio Y and Bareinboim E (2021) The causal-neural connection: expressiveness, learnability, and inference. In: *Advances in Neural Information Processing Systems*.

Yang Z, Ishay A and Lee J (2020) NeurASP: embracing neural networks into answer set programming. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 1755–1762.

Zennaro FM, Drávucz M, Apachitei G, Widanage WD and Damoulas T (2023) Jointly learning consistent causal abstractions over multiple interventional distributions. In: *Conference on Causal Learning and Reasoning*, PMLR 213, pp. 88–121.

Zhang J (2013) A Lewisian logic of causal counterfactuals. *Minds and Machines* 23(1): 77–93. DOI: 10.1007/s11023-011-9261-z.