Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Birth, Life, and Death of Internal Organization

Yuxuan Deng

dengyx0211@163.com

Independent Researcher https://orcid.org/0009-0000-1531-2152

Physical Sciences - Article

Keywords: consequence organization, internal organization, causal abstraction, causal intervention, learning dynamics, mechanistic interpretability, adaptive systems, cross-realization identity

Posted Date: October 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11356934/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

The Birth, Life, and Death of Internal Organization

Yuxuan Deng Independent Researcher dengyx0211@163.com ORCID: 0009-0000-1531-2152

Version 2. This revision expands the reporting of already executed evidence, including archived cross-architecture lifecycle confirmation, internal-coordinate robustness, real-sensor boundary results and the recovered complete V19 cohort. No new core experiment was run for this revision.

Abstract

Learning systems change internally all the time, yet it remains unclear when such change constitutes an organization: when one comes into being, whether two differently built systems carry the same one, and what remains when it dies. Internal structure is usually compared by representational similarity, by what a probe can decode, or by alignment to a hypothesized algorithm [1, 2, 3, 4]. Here we identify internal organization by what admissible internal interventions do to an independently named future consequence; the resulting *Consequence Organization*, a construction closely related to causal abstraction, is *endogenous* when the system's own consequence history forms it. We show that, across naturalistic and controlled systems, such organizations formed under stable but not shuffled consequence, persisted and reorganized, could be erased, restored and transplanted, and could lose their causal role while carrier-associated readouts or residual carrier content remained. Their identity transferred across independently trained seeds and, in most directions, across architectures, but failed across an intentionally incompatible cross-consequence family even where internal operators remained nearly identical. In the tested systems, internal organization thus has a measurable lifecycle and a consequence-relative identity that neither decodability nor internal-operator similarity alone tracks.

When change becomes organization

Adaptive systems alter parameters, representations and behavior as they learn. Classical state-space models, hidden-state models and system identification compress dynamics into state variables chosen for filtering, prediction or dynamical sufficiency [5, 6, 7]; predictive-state and computational-mechanics approaches instead define state through equivalence of future distributions [8, 9, 10, 11]. For learned internal structure, sameness is usually judged by representational similarity or stitchability between networks [12, 1, 2, 13], by what a probe can decode [3], or by alignment to a hypothesized high-level algorithm [14, 4]. A feature may nevertheless remain decodable after its former role has disappeared, and the same causal role may be instantiated by different coordinates. The question here is therefore not whether a system changed internally, but whether previously interchangeable *internal interventions* became different with respect to an independently fixed consequence, and whether that difference has a lifecycle and an identity of its own.

Following the interventionist distinction between observation and causal manipulation [15, 16], fix a realization r at time t, a prospectively frozen family of admissible internal interventions $\mathfrak{I}_t^{(r)}$, and a named consequence space $\mathcal{Q}$. The real future consequence operator is

$$\mathfrak{C}_t^{(r)} : \mathfrak{I}_t^{(r)} \longrightarrow \mathcal{Q}. \tag{1}$$

Two interventions are organizationally equivalent when they produce the same audited consequence,

$$i \sim_{Q,t}^{(r)} j \quad \Longleftrightarrow \quad \mathfrak{C}_t^{(r)}(i) = \mathfrak{C}_t^{(r)}(j), \tag{2}$$

giving the local realized quotient

$$E_t^{(r)} = \left(\mathfrak{I}_t^{(r)} / \sim_{Q,t}^{(r)}, \overline{\mathfrak{C}}_t^{(r)} \right). \tag{3}$$

The quotient is not yet the fundamental object, because its intervention coordinates belong to one realization. We therefore identify two realizations when there is a bijection between their quotient classes that preserves the anchored consequence map. Two realizations are *compatible* when they are evaluated with the same named consequence space $\mathcal{Q}$ and an admissible intervention family of the same type, so that their consequence maps can be compared class by class; in the experiments below they differ in seed or architecture while task, outcome and action semantics are shared, and the cross-consequence test lies outside this definition by design. Writing this consequence-preserving isomorphism as $\cong_Q$, the object studied here is

$$\boxed{\mathscr{O}_Q(t) = \left[\mathfrak{I}_t^{(r)} \to E_t^{(r)} \to \mathcal{Q} \right]_{\cong_Q}}. \tag{4}$$

We call $\mathcal{O}_Q(t)$ a *Consequence Organization*. If the organization was absent before adaptation and is subsequently created by the system's own consequence history, we call it an *Endogenous Consequence Organization*. A measured subspace, layer stack, low-rank coupling or complete parameter vector may realize or encode $\mathcal{O}_Q$ without being identical to it. Conversely, the definition does not assert metaphysical independence from the complete physical state: any deterministic realization can be represented in a sufficiently rich state description. The empirical claim is instead that the same consequence organization can be identified and manipulated across distinct realizations while coordinate identity is neither necessary nor sufficient.

Formally, the local quotient is a causal coarsening of the intervention family [17, 18], akin to causal states and to bisimulation in Markov decision processes [8, 19], and identification across realizations is an exact causal abstraction of each realization to a common high-level model whose variables are the consequence classes [20, 21, 14, 4]. The contribution is therefore not a new notion of abstraction, nor any one of these formal operations. Here they are combined in a time-indexed empirical construction in which admissible internal interventions are partitioned by a named future consequence, and the resulting organization is followed through formation, persistence, reorganization, death and transport.

The finite closure signature σ_Q over the frozen probe family is only an operational certificate of consequence anchoring at the tested resolution. We do not assume $\sigma_Q \equiv \overline{\mathfrak{C}}$ unless the probe family is separately shown to be separating. The object is also explicitly Q-relative: realization can change while the consequence semantics remain fixed, but no invariance across arbitrary consequence laws is assumed.

Birth is a newly resolved consequence distinction; persistence is preservation of that distinction; reorganization changes its anchored incidence; and death is the merger of a previously active distinction under the named consequence readout. Identity is then tested by intervention: erase should remove the consequence mapping, restore should recover it, and transport should carry the mapping into a compatible realization. Figures 1–4 keep the naturalistic recommendation data and controlled neural-model experiments in separate evidence fields while following these shared causal questions. The cross-architecture and cross-consequence direction rows provide distinct tests of implementation change and consequence-relative identity.

Birth under stable consequence

The flagship experiment used the random-policy component of KuaiRand, a public recommendation dataset containing platform-randomized video replacements inserted into standard recommendation feeds [22]. The released random-policy subset contains 1,186,059 interactions from 27,285 re-indexed users. Our analysis partitioned users into 16,373 development, 5,456 locator and 5,456 frozen-confirmation users. A generic 18,891-parameter MLP received observations and scalar consequence but no organization label, phase identifier, router, expert bank or dedicated memory.

The naturalistic evidence comes from distinct archived experiments, retaining pre-absence, formation, persistence, reorganization, polarity reversal and Null readouts. The later unpaired cross-consequence archive supplies the two task/model families in Figs. 1b and 2a; it is separate from the original long-view manipulation and cross-seed archives. Matched shuffled-consequence rows preserve the relevant marginal data and training budget while breaking action–outcome correspondence. When the always-present release mechanism was removed, unsupported C remained causally active in the retained control rows. These rows separate recruitment from generic training and loss of support from automatic death. A separate controlled-language-model archive supplies executed formation rows and phase-resolved histories. Figure 1 combines the consequence schematic and naturalistic run field with the controlled formation and lifecycle panels, so the birth claim is shown from design, unit-level evidence and phase-resolved histories rather than from a single trajectory. The controlled rows are not pooled statistically with the KuaiRand lifecycle.

Figure 1c keeps each controlled history as one paired unit and displays ordinary/shuffled history–world contrasts $I = x(M_\alpha) - x(M_\beta)$, where $x(M) = \mathrm{Acc}_\alpha(M) - \mathrm{Acc}_\beta(M)$. Figure 1d shows x at four recorded phases for three 3B histories. Persistence and reorganization passed in all three, but ordinary release passed in 2/3 and stack release in 1/3; the partial phase series is retained without retrospective rescue. Complete gate-profile rows are provided in Source Data.

Architecture breadth and its reachability boundary

The neural evidence extends beyond the histories shown in Fig. 1. An archived five-architecture experiment applied a common intervention-defined causal-role audit to an MLP, CNN, GRU, small Transformer and pooled GNN. Its broad matrix crossed five architectures with three residual-task families: formation passed in 9/15 conditions and the complete lifecycle in 8/15. These failures delimit reachability under the tested training schedules; they are retained rather than combined with later successes.

A subsequent fresh confirmation retained the role thresholds and used a 1200-step terminal horizon with two new seeds per architecture in the sine family. Formation passed in 10/10 history conditions, reorganization in 9/10, and both terminal role death and paired continued-support persistence in 10/10. The complete lifecycle passed in 9/10: MLP 2/2, CNN 2/2, GRU 1/2, Transformer 2/2 and GNN 2/2. GRU seed 4020 failed reorganization. Median terminal old-role ratios were 0.01261 under Null and 0.95786 under continued support

(archived one-sided paired Wilcoxon $p = 9.77 \times 10^{-4}$). The stable 10/10 comparison with ten archived controls from the preceding broad matrix (five shuffled and five unstable; formation 0/10) gave the archived one-sided Fisher test $p = 5.41 \times 10^{-6}$. The ten conditions are model histories, not ten architectures, and the two branches of a condition remain paired. The limited migration assay examined two candidate representation sites; it does not exclude migration anywhere in the network. Supplementary Table S1 retains each history, its failure status and both measured terminal ratios.

A real-sensor test and a selective-replacement boundary

An independent real-sensor experiment used released HAPT smartphone accelerometer and gyroscope recordings: 30 participants, six channels at 50 Hz, and 10,929 processed windows of shape 6×128. Each run used 6,324 development windows and 2,381 subject-held-out confirmation windows. A 29,210-parameter one-dimensional convolutional policy received scalar consequences for externally defined activity distinctions, without internal organization labels. Five fixed ordinary seeds were compared with five shuffled-reward seeds; locator and evaluation data were separated by cross-fitting.

All five ordinary histories passed A formation, A persistence, B reorganization, B polarity reversal and C formation, whereas shuffled controls passed none of those formation-dependent stages. Removing C support produced C role death in 5/5 ordinary histories. The harder non-null replacement of B by C passed in only 2/5. The full lifecycle therefore failed its locked requirement of at least three complete ordinary histories; the strong early-stage effects do not overrule that verdict.

This failure distinguishes zero expressed behavior from causal elimination. After C, mean B reward was 0.00023 and B remained decodable, but the intervention-defined residual-B/peak-B ratios were 0.667, 0.051, 0.054, 0.266 and 0.656 for seeds 6101–6105. Only seeds 6102 and 6103 passed the frozen 0.25 ceiling. Three histories therefore retained opposing or compensating causal components despite near-zero net B preference. This experiment supports specific lifecycle stages in held-out real sensory data and an explicit selective-replacement boundary. It does not establish autonomous value formation or online physical interaction.

Role death with carrier persistence

Figure 2a separates rank-four C-channel deletion effects from trace-probe AUC across twelve archived runs. C and Null are different checkpoints of the same run. The probe uses hidden features plus action indicators, and shuffled runs that never formed C reach the same AUC range; its AUC therefore neither isolates hidden-state memory nor indexes the C organization.

In a separate small-network experiment, three learner/world cases each supplied four continuations cloned from their B endpoint and optimizer state (Fig. 2b). Three Null regimes remove B support but differ in trainable layers; continued B retains it. The 60 records track old-mapping role power separately from residual carrier energy at every recorded step. Role power collapsed within 50 steps in every Null branch, whereas carrier content was case-dependent: it persisted or grew in the quadratic case, decayed with the role in the sine case and was near zero throughout in the interaction case. Loss of the audited causal role therefore need not eliminate energy in the former carrier, but it does not always preserve it.

The complete main V19 archive contains a larger cohort than the three mechanistic examples in Fig. 2b: 30 history-condition units, comprising 12 development and 18 fresh confirmation histories, with 480 repeated branch–checkpoint records. Twelve of the 18 fresh histories met the archived formation-and-reorganization eligibility rule before release; the six source-ineligible histories remain in the exported tables. All 12 eligible Null branches passed the original terminal death gate and all 12 paired continued-B branches passed persistence. At step 3200, median history-normalized Null role power was 0.00368236 of its formed-B baseline while median activation energy in the B-frozen carrier coordinates was 0.386598; paired continued-B medians were 0.957478 and 1.04120. These fresh-only medians must not be mixed with the historical 0.006431/0.383898 medians, which use 20 source-eligible development-plus-confirmation histories.

Carrier-energy retention was heterogeneous, including strong retention, increases and near-complete loss. It measures remaining activity in the former carrier coordinates, not independently certified history-specific memory. Branches shared model and optimizer states at cloning; global random state was not cloned, so subsequent training samples are not claimed to be byte-identical. The complete recovered cohort, every excluded history and the existing mechanism and stress examples are supplied alongside the preprint. A separately labelled retrospective first-sampled-death analysis does not replace the fixed terminal endpoint. Supplementary Table S3 and the accompanying three-page V19 evidence supplement expose the history-level heterogeneity.

Erase, restore and transplant

Reversible interventions were frozen in the original long-view recommendation ecology. Rank-four hidden-to-action couplings were measured on locator users, with restore estimated on a disjoint fold; confirmation used 5,456 held-out users and 243,254 exposures. Figure 3b retains ordinary/shuffled seed profiles from a separate gated lifecycle archive. Figure 3a shows 50 fresh controlled worlds, all of which passed the 13 frozen

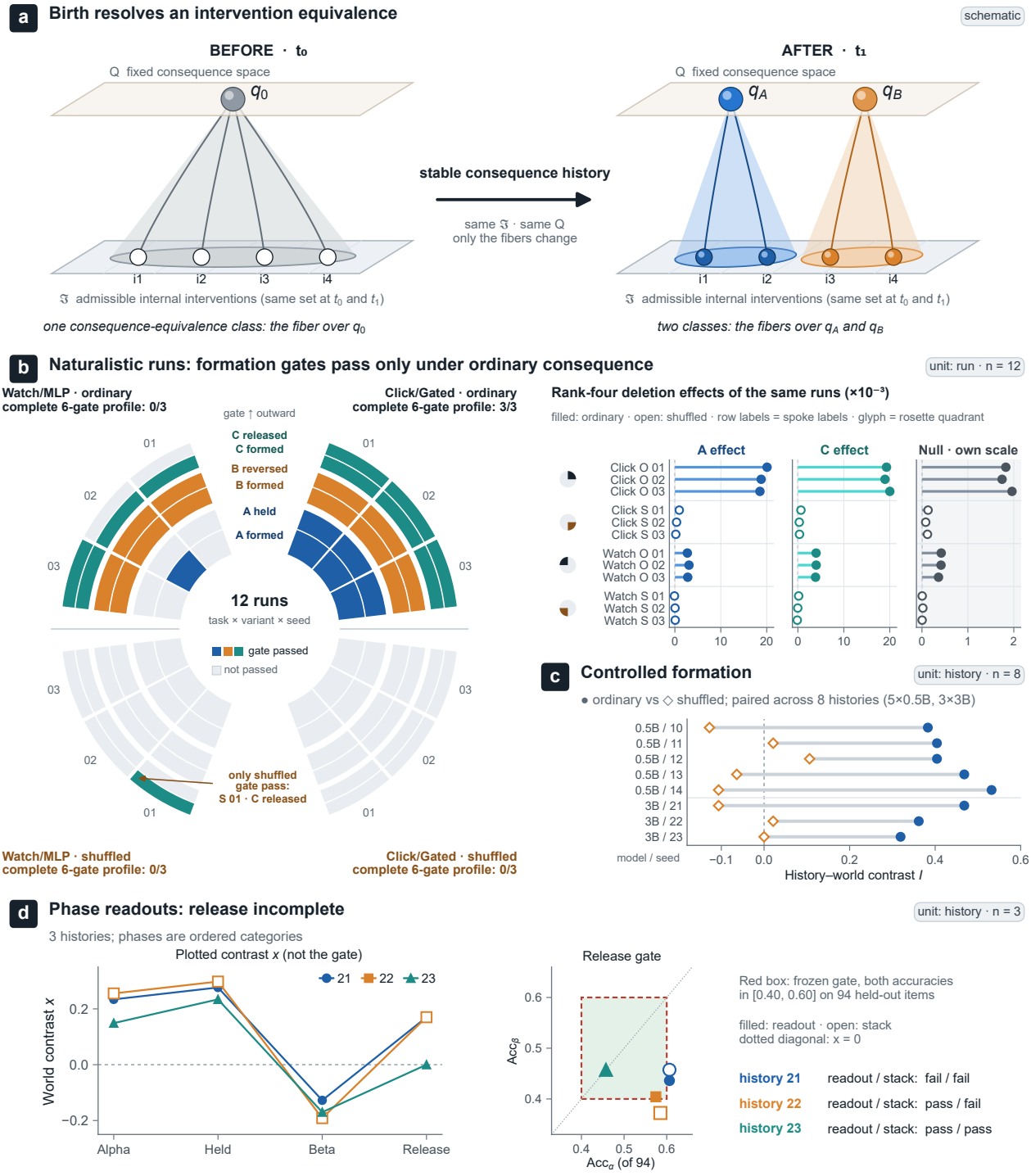


Figure 1. Formation contrasts and phase histories retain their experimental boundaries. a, Schematic fiber split: admissible internal changes $\mathfrak{I}$ begin consequence-equivalent (one fiber over q_0) and separate under fixed Q (fibers over q_A and q_B); positions are not measurements. **b,** Twelve naturalistic runs (two task/model families, ordinary/shuffled variants, three seeds). Task is left/right (Watch/MLP, Click/Gated), variant top/bottom; each sector is one run and six outward equal-area lifecycle gates (coloured, passed; grey, not). The only shuffled pass is Watch S 01 at C release. Adjacent points show the same runs' raw rank-four deletion effects (A/C share an axis; three slightly negative shuffled values are retained). Complete lifecycle gates passed in ordinary Click/Gated (3/3) and none in Watch/MLP (0/3). **c,** Within eight controlled histories (five 0.5B, three 3B), conditions are paired and connectors show raw history–world contrasts I. **d,** Three 3B histories show $x = \mathrm{Acc}_\alpha - \mathrm{Acc}_\beta$ at four phases (left) and release accuracies (right). Release requires both accuracies in $[0.40, 0.60]$ on 94 held-out items, not a threshold on x; filled/open points are readout/stack. Histories 21 and 22 share $x = 16/94$: 21 fails at $\mathrm{Acc}_\alpha = 57/94$ (one item above band), whereas 22 passes at $\mathrm{Acc}_\beta = 38/94$ (lowest passing count); its stack fails at $35/94$. Readout release passed $2/3$ histories and stack release $1/3$; complete release is not established. Experiments are not pooled.

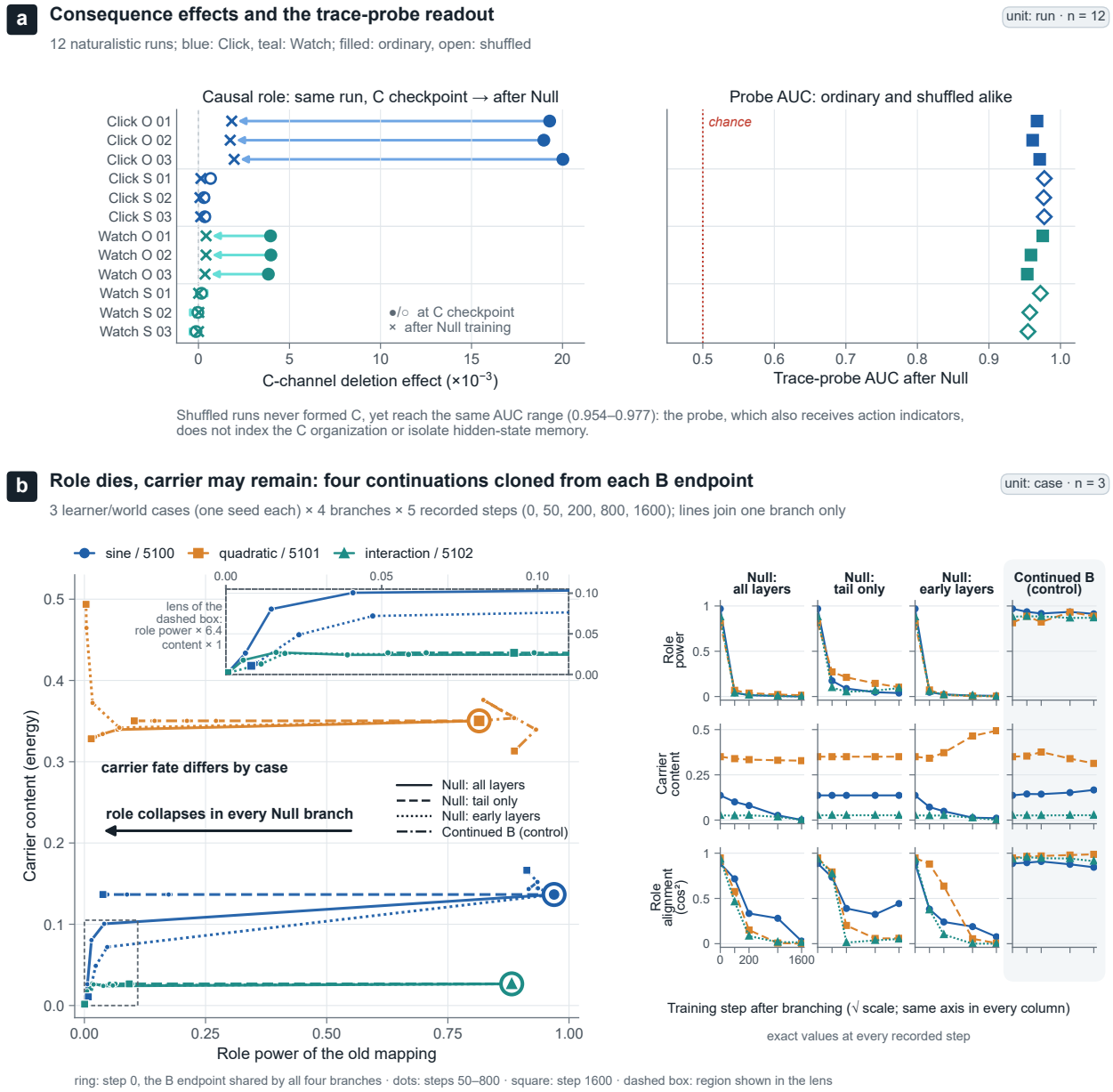


Figure 2. Consequence effects and probe information are separate readouts. a, Twelve naturalistic runs show C-channel deletion effect at checkpoint and after Null in each run (left; ordinary Click, $\sim 0.019 \rightarrow \sim 0.002$; Watch, $\sim 0.004 \rightarrow \sim 0.0004$) and trace-probe AUC after Null (right, 0.954–0.977; chance 0.5). Shuffled runs that never formed C reach the same AUC; the probe includes action indicators and does not isolate hidden-state memory. Colour denotes task; filled/open denote ordinary/shuffled. **b,** Three learner/world cases (sine, quadratic, interaction; one seed each) branch from a shared B endpoint into three Null regimes and continued B. Left, every checkpoint plots role power against carrier content; the inset expands only the near-zero role-power axis (×6.4). Role power collapses in every Null branch, while carrier content persists or grows (quadratic), decays with role (sine, all-layer/early-layer Null), or stays near zero (interaction). Right, each readout is followed across five steps (0, 50, 200, 800, 1600) on a square-root step axis; lines join checkpoints within the same case/branch (60 records). Tail-only Null freezes early layers, so carrier content is constant by design; early-only Null freezes the tail. Branches/checkpoints are repeated records, not independent cases.

gates; each world contributes one point to every operation, so the operations are repeated measurements within world, not additional worlds, and are not pooled with naturalistic seeds. The views pair the gated frozen-locator fraction with the ungated full-signature residual of the same world: every world lies inside the frozen locator interval, while full-signature residuals remain 0.68–1.83 of the native gap. Residual ratios are descriptively larger in worlds with larger locator overshoot and smaller native gap; because that gap is the ratio denominator, neither co-variation is interpreted as an independent association.

Across five ordinary confirmation seeds, the naturalistic manipulation archive gives mean effects for C in its native state, after erase, and after restore of 0.008290, 0.000454, and 0.008285, respectively. Transplanting C into Null raises the mean effect from 0.000263 to 0.007940; restoring Null returns it to 0.000263. On a common $+B$ axis, exchanging the positive-B module into a negative-B recipient gives $+0.006151$, whereas exchanging the negative-B module into a positive-B recipient gives -0.010840; restoring recipient modules recovers the original signs. The accompanying Source Data retains all ten archived run-level rows and their twelve frozen gate outcomes per run; the B-exchange signs are expressed on the common $+B$ axis.

Identity across realizations

Cross-seed transport tested 20 directions among five MLP realizations using locator consequence signatures; confirmation users were excluded from alignment (Fig. 4a). The architecture test retains all 30 directions and its two failures (Fig. 4b). The 18-direction unpaired cross-task test jointly changes user domains, outcomes, action/context maps and architecture (Fig. 4c); near-one internal operator cosines do not rescue the failed full identity gate. These fields separate compatible-realizer transport from the cross-consequence boundary.

Function-preserving changes of internal coordinates

Architecture change and coordinate change are distinct tests. A further archived assay replaced the Euclidean locator by a covariance-whitened variation–sensitivity operator and subjected each trained model to a non-orthogonal invertible linear transformation of its internal coordinates, with compensating downstream weights. The tested transformations had condition number approximately 8.6. Across one stable model of each of five architectures, formation passed both before and after transformation (5/5 paired comparisons); five shuffled models failed formation (0/5). The maximum sampled functional discrepancy in the saved machine-readable table was $5.602836608886719 \times 10^{-6}$; the historical prose report rounded this differently as 5.61×10^{-6}. Function comparison used 256 sampled inputs, not an exhaustive equality proof.

Minimum A-stage and B-stage role-output correlations were 0.7816 and 0.8798. Only 3/5 stable cases passed the full lifecycle under the stricter whitened audit. Thus the formation distinction was robust to these tested coordinate transformations, while complete lifecycle performance remained conditional. Original and transformed readings come from the same trained models and are paired observations, not ten independent histories. This supports bounded linear-coordinate robustness and does not imply invariance under arbitrary nonlinear transformations. Supplementary Table S2 gives every architecture, both errors and its lifecycle result.

Exploratory lifecycle dynamics and its predictive boundary

An archived early-trajectory fit reanalysed 22 formation-and-reorganization-qualified histories from overlapping cohorts rather than adding independent confirmations. It recovered the terminal Null-versus-continued ordering in 22/22 pairs and the ratio-only Null death classification in 21/22; the latter counts correct classifications, not 21 observed deaths. Continuous endpoint prediction did not outperform the archived early-hold baseline: mean absolute errors were 0.0189744 versus 0.0170958 under Null and 0.333066 versus 0.203940 under continuation. The paired-update decomposition, constitutive interpretation and full limitation are retained as exploratory analysis in Supplementary Note S5 and Table S4. These results do not establish a universal kinetic law and are not required for the formation, manipulation or consequence-relative identity conclusions.

Discussion

The experiments identify an adaptive object by causal organization rather than by location. Stable consequence can split a previously coarse intervention equivalence; continued support can preserve that split; changed consequence can reorganize or reverse it; release can merge a former distinction; and certified realizers can be erased, restored, transplanted and exchanged. The same carrier can therefore realize different consequence organizations, while distinct carriers can realize the same $\mathcal{O}_Q$ under a fixed consequence semantics.

This object sits among established notions rather than outside them. State-space and system-identification variables summarize dynamical sufficiency [5, 6, 7]; causal states and predictive-state representations group histories by future prediction [8, 9, 10, 11]; causal representation learning seeks variables with causal semantics [23]; probes, unit analyses and activation patching reveal encoded, localized or mediating structure [3, 24, 25, 26, 27]; and causal abstraction tests whether components realize a high-level model [14, 4]. Consequence Organization uses an intervention-preserving identity criterion closely related to causal abstraction, and its erase, restore and transplant operations are patching interventions. The contribution is not a new general notion of abstraction or a new intervention tool, but an empirical account of how a consequence-anchored organization forms, persists, dies, is restored and is transported, and of where its identity fails across

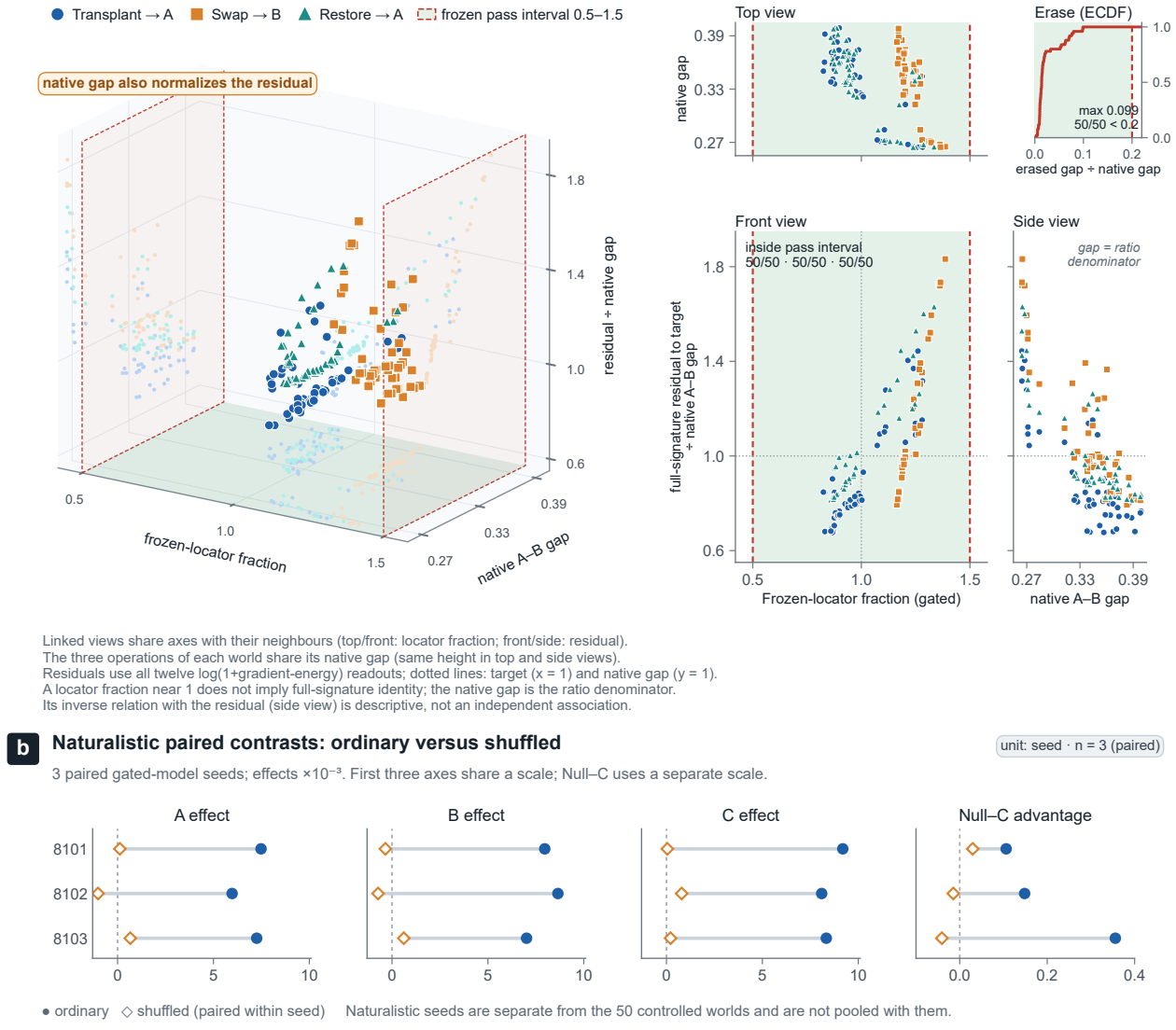


Figure 3. Intervention readouts retain their world and seed units. a, Each point is one of 50 controlled worlds under one operation; the same worlds recur in every operation, so the panel holds 50 units, not 150 or 200. Three measured variables are shown for transplant, swap and restore: the gated frozen-locator fraction (green: frozen pass interval 0.5–1.5; 1 = target reached along the locator), the ungated full-signature residual distance to the A, B and A target, respectively, divided by the native A–B gap, and that native gap. Left, an orthographic three-dimensional view draws the frozen interval as two criterion planes and every observation also as its projection on floor and walls; right, linked top, front and side views share their axes and give exact readings; the three operations of a world share its native gap. The native gap is the ratio denominator; its inverse relation with the residual is therefore descriptive and is not interpreted as an independent association. Signatures contain log(1+gradient energy) readouts for six frozen future tasks. The erase inset is the right-continuous empirical distribution of the erased-state gap divided by the native gap (frozen criterion < 0.2). All 50 worlds passed all 13 frozen gates; a locator fraction near 1 is not full-signature identity. **b,** A, B and C deletion effects and Null–C advantage relative to uniform policy are paired across ordinary/shuffled training within three naturalistic gated-model seeds. The naturalistic experiment is separate from panel a.

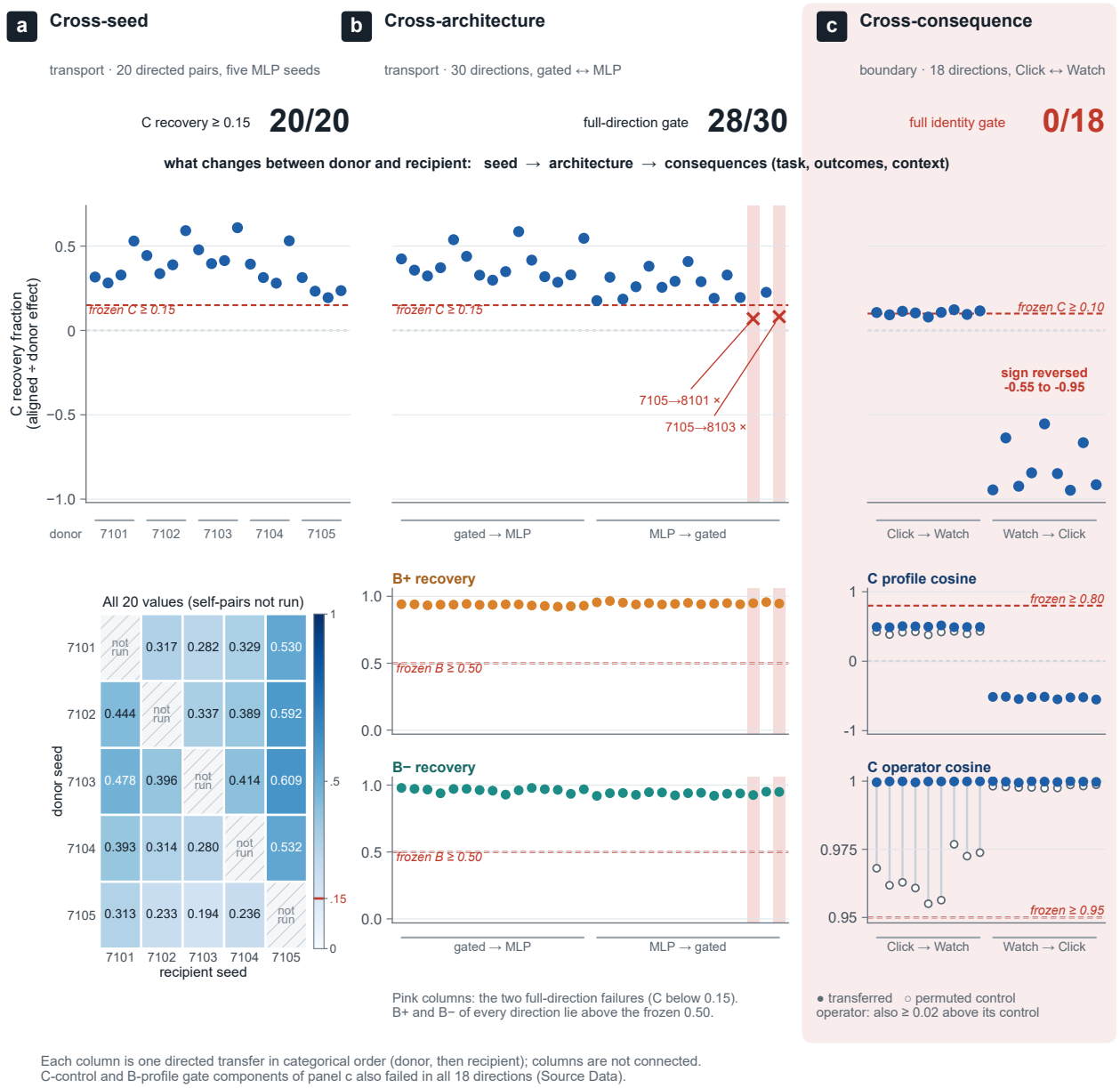


Figure 4. Cross-realization recovery and the cross-consequence boundary remain separate tests. Top, C-recovery fraction (aligned/donor effect) for all 68 directed transfers, with each frozen criterion; categorical columns are not connected, and nested readouts align beneath each transfer. **a,** All 20 cross-seed directions among five MLP realizations; no self-pairs. The 5×5 matrix gives exact off-diagonal values; criterion 0.15 appears on axis and colour scale. **b,** All 30 architecture directions with nested C, B^+ and B^- recovery readouts; criteria are C 0.15 and B 0.50 (red dashes). Two full-direction gates fail below the C criterion; several passes are near it. Directions are not independent seeds. **c,** All 18 cross-task directions are assessed against C recovery (criterion 0.10), C-profile cosine, C-operator cosine and permuted controls; failed C-control and B-profile components are in Source Data. Click→Watch recovers 0.08–0.12 of the donor effect; Watch→Click reverses sign (-0.55 to -0.95). Profile cosines remain below 0.80 and Watch→Click profiles match permuted controls. Operator cosines are near 1, but so are controls, failing the required 0.02 margin. No direction passed; near-one operator similarity does not establish consequence-profile identity.

consequence-incompatible transfers. A state model may encode this object; encoding does not supply that account.

Two empirical separations show that common decodability and internal-similarity criteria are insufficient to track this identity in the tested systems. The first concerns decodability: after Null training the C organization's causal role collapsed, yet a trace-probe readout stayed in the same range as in shuffled runs that never formed C, so that readout does not index the organization (Fig. 2a); in a separate assay, residual carrier content persisted, and in one learner grew, after the role had died (Fig. 2b). Carrier-associated readouts can therefore outlast organizational death, extending to whole organizations a dissociation that amnesic probing has shown for individual features [28]. The second concerns similarity: transport between independently trained seeds and architectures has an analogue in model-stitching tests of cross-model compatibility [12, 2], but in the cross-consequence test internal operator signatures remained nearly identical while consequence identity failed in all 18 directions and Watch→Click recovery reversed sign (Fig. 4c). Together with cross-seed and cross-architecture transport, these results support bounded realization invariance within specified consequence semantics, not a universal or law-independent ontology.

The same separations state how the proposal could fail. If every identification across realizations were already predicted by realization-local coordinate equality, so that consequence-preserving equivalence added no invariance beyond matched state coordinates, the object would have no independent empirical content; the cross-seed and cross-architecture transfers instead succeed without coordinate identity. If, under fixed consequence semantics, transport between compatible realizations failed systematically despite correct identification of each local quotient, the cross-realization identity criterion would be refuted; those fields retain every direction, including the two architecture failures. If loss of the audited role always coincided with loss of carrier content, the distinction between organization and carrier would be unnecessary; Fig. 2 records carrier persistence as well as carrier decay after the same role collapse. Finally, the cross-consequence experiment tests a joint change of outcomes, users, action/context maps and architecture. Its failure shows that internal operator similarity did not preserve the full tested identity contract in that family. Consequence relativity does not require every change of consequence law to fail: any proposed transport would require an independently justified causal correspondence of the consequence semantics.

If this account holds, organization need no longer be identified with a location, a coordinate system or a fixed mechanism; it can be identified experimentally by a consequence-preserving causal identity across realizations. The questions then shift from where a representation resides to which realizations instantiate the same organization, when an organization is born and when it dies, what survives its death, and whether it can be constructed in a different system.

The intervention evidence acts on certified physical realizers of $\mathscr{O}_Q$ at the stated audit resolution. The cross-realization result therefore concerns preservation of consequence-class incidence and anchoring under compatible realization changes; it does not require coordinate identity, microscopic uniqueness or invariance across different consequence laws.

The empirical conclusions concern formation, lifecycle and interventionally tested consequence-relative identity. The recruitment–release decomposition remains exploratory bookkeeping rather than a validated constitutive law for the neural systems: the archived continuous endpoint forecasts did not beat the early-hold baseline. Determining when recruitment, contraction and leakage can be measured before an unseen phase remains a separate predictive problem (Supplementary Note S5). Thus the reported evidence supports a measurable lifecycle and compatible cross-realization identity under fixed consequence semantics, not universal kinetics or a complete theory of adaptive development.

Methods

Naturalistic random-policy data and consequence definition

KuaiRand documents a random-exposure collection stage in which each recommended video had equal probability of being replaced by a random video sampled from an item pool, with approximately 0.37% of interactions replaced in the released logs [22]. The random-policy file contains all interactions resulting from the random intervention. User and item identifiers are re-indexed. The SE analysis used 1,186,059 random-policy interactions from 27,285 users and partitioned users into development, locator and frozen-confirmation subsets before confirmation. The principal consequence was the released `long_view` field, defined by KuaiRand as complete viewing for videos of at most 18 seconds or at least 18 seconds of play for longer videos. This study is a secondary analysis of a dataset containing platform-randomized item replacements; it is not described as a trial randomized by the present author. The two task/model families in Figs. 1b and 2a use long-view outcomes on even-ID users (Watch/MLP) and click outcomes on odd-ID users (Click/Gated), with different opaque action assignments and context weights. A, B and C denote frozen action-group reward channels, not supplied organization labels; Null is audited against the old C channel. Each plotted deletion effect is $s[V(\pi) - V(\pi_{\text{deleted}})]$ for a located rank-four subspace, whereas advantage is $s[V(\pi) - V(\pi_{\text{uniform}})]$. Here $V = \sum_i w_i r_i / \sum_i w_i$ and $w_i = \pi(a_i \mid o_i)/p_i$ use the archived propensity field, and s is the prescribed polarity.

The trace-probe AUC uses both hidden features and observed-action indicators; its input contract does not isolate a history-specific hidden trace.

Repeated exposures from the same user are not treated as independent confirmatory replicates. The primary SE confirmation unit for the causal chains is the frozen model realization (seed). Held-out user exposures estimate the named consequence effects within each realization; protocol success is adjudicated at the realization level. The archive does not contain a complete platform propensity/filtering specification beyond the public random-replacement description, so claims are limited to the released random-policy subset and named consequence.

Formation, lifecycle and death audits

Formation required pre-absence followed by a selective post-consequence distinction and matched-control separation. Lifecycle gates tested persistence under continued support, reorganization or sign reversal under changed consequence and dissolution under release. In the phase-resolved 3B histories, persistence and reorganization were signed criteria on x, whereas readout and stack release each required $0.40 \leq \mathrm{Acc}_\alpha \leq 0.60$ and $0.40 \leq \mathrm{Acc}_\beta \leq 0.60$ on the same 94 held-out items; equal x can therefore receive different release verdicts. The same-checkpoint death experiment branched Null and continued-support futures from the same checkpoint and optimizer state and audited old-mapping consequence separately from residual carrier content. Figure 2b is the small-network mechanism archive, separate from the language-model phase series. Its sine/ReLU, quadratic/Tanh and interaction/residual-ReLU cases each retain one seed. All four continuations share each case's B endpoint and RMSprop state; the code does not clone global random state for subsequent samples. Null removes the B-differential training target while retaining the shared base target; continued B retains both. Tail-only Null freezes early layers, whereas early-only Null freezes the tail. With h_d the condition-centered hidden response, V the B-frozen intermediate basis, y_d the B-target differential and δp_d the intact-minus-deleted output differential, carrier content is $\langle (h_d V)^2 \rangle / \langle y_d^2 \rangle$, role power is $\langle (\delta p_d)^2 \rangle / \langle y_d^2 \rangle$, and alignment is $\cos^2(\delta p_d, y_d)$. Means include all displayed latent coordinates where applicable; squared alignment is not a signed-polarity readout.

Erase, restore, transplant and signed swap

A rank-four hidden-to-action coupling was estimated on locator data. Erase removed that measured coupling. Independent restore used a disjoint locator fold. C transplantation inserted the active C realization into a released Null recipient, followed by recipient recovery. Signed exchange transplanted positive B into a negative-B host and vice versa. Equal-rank random subspaces, action-row permutations, wrong-organization modules and shuffled-consequence modules served as separation controls. These interventions identify a reproducible realizer at the stated resolution; they do not identify a unique or minimal physical carrier.

Cross-realization identity

Cross-seed transport has 20 directed pairs among five MLP realizations; cross-architecture transport has 30 directed pairs between MLP and gated-residual families. Both retain C, positive-B and negative-B readouts; frozen per-direction criteria included donor recovery ≥ 0.15 for C and ≥ 0.50 for signed B, plus aligned-effect and control-margin criteria. The 18-direction unpaired boundary jointly changes outcome semantics, user domain, action/context maps and architecture. All full-direction gates fail. Its frozen per-direction criteria were C recovery ≥ 0.10, held-out profile cosine ≥ 0.80, operator-signature cosine ≥ 0.95 and at least 0.02 above its permuted-locator control, together with control-margin and restore criteria. The profile and score-operator readouts test that joint transport contract; they do not isolate one causal factor.

Exploratory dynamics and admission boundary

Lifecycle gates were frozen independently of the dynamics. The paired-update decomposition and its constitutive interpretation are detailed in Supplementary Note S5. They do not supply an independently validated neural constitutive law. The retrospective fit is evaluated against the archived early-hold baseline on the same 22 selected histories, retaining both branch-specific continuous errors. No universal timescale or prospective deep-neural closure is claimed.

Controlled model experiments and unit definitions

The controlled formation archive includes five 0.5B and three 3B language-model histories; the lifecycle archive includes three phase-resolved 3B histories. The displayed small-network branch experiment contains three learner/world configurations, one seed per configuration, four cloned branches per case and five repeated steps per branch (60 branch–checkpoint records). The cases differ in task function and network settings; they are not three identical-configuration seed replicates. Figure 3b contains six lifecycle rows (three ordinary/shuffled seed pairs), separate from the ten-run naturalistic manipulation archive reported above. Figure 3a contains fifty controlled-world units, which are not pooled with either naturalistic experiment. For the 50-world experiment, the functional signature S contains twelve readouts: $\log(1 + \|\nabla_\theta L_{Q_j}\|^2)$ and $\log(1 + \|\nabla_{\theta_h} L_{Q_j}\|^2)$ for each of six frozen future tasks Q_j. The distance ratios use the native $\|S_A - S_B\|$ as denominator: transplant and restore retain residual distance to S_A, swap to S_B, and erase the residual gap between the two erased states. Locator

fractions instead measure displacement along a frozen development-time locator. The frozen pass intervals were $(0.5, 1.5)$ for each locator fraction and < 0.2 for the erase ratio, alongside eleven further frozen gates; all 50 confirmation worlds passed all thirteen. A locator fraction near 1 is not equality of the complete signature. Cross-seed transport contains twenty directed contrasts among five MLP realizations; cross-architecture transport contains thirty directed rows; and the cross-consequence boundary contains eighteen directed rows. Direction and checkpoint observations remain nested within their realization, branch or transfer design rather than being counted as independent model histories.

Restored architecture and coordinate assays

The five-architecture fresh-horizon table contains ten history conditions, two seeds per architecture in the common sine family, and a 1200-step terminal branch horizon. Stable and matched unstable/shuffled runs retain the original thresholds. The earlier fifteen-condition broad matrix is reported separately. The coordinate assay applies covariance whitening on audit-train, then tests paired original/transformed internal intervention responses under invertible linear changes with compensating downstream weights. The sampled function-comparison bank contains 256 inputs. Neither paired transformed readings nor the control conditions are extra independent architectures.

Complete V19 cohort and denominator-specific reporting

The main archive contains twelve development and eighteen fresh confirmation history conditions. Six configurations crossed with three optimizers provide one fresh seed per confirmation cell; there is no within-cell seed replication. Eligibility requires the original formation criteria and B's old-A ARI below 0.55. Terminal death uses the archived conjunction of absolute frozen-basis differential role below 0.025, role/B below 0.10, and current differential power below 0.025. Continued support requires role/B above 0.35 and current differential power above 0.20. Energy is measured in the B-frozen coordinates and each displayed history fraction uses that history's B baseline. Newly refitted checkpoint clustering and carrier energy do not provide a PRE-controlled reader of history-specific content. Source comments describe gates fixed before confirmation, but no external preregistration is asserted.

Real-sensor control and negative lifecycle adjudication

The HAPT confirmation compares five fixed ordinary seeds and five shuffled-reward seeds. Biological participants supply nested sensory samples, not thirty independently trained models. Each policy is the lifecycle unit; subject-held-out observations estimate its effects. The full lifecycle requires at least three of five complete ordinary histories and zero of five shuffled A formations. Only two ordinary histories completed it; the formal result remains not established. The supplied tables retain all seeds, stage-specific gates and selective-death ratios.

Retrospective fit evaluation

The normal-form rows re-use 22 formation-and-reorganization-qualified histories from V20, V20b and V21. Forty-four branch rows are repeated Null/continued observations of those histories, not forty-four new histories. The early fit and early-hold baseline are compared using the archived continuous endpoint errors, not only successful binary classifications.

Statistical units and reporting

Confirmatory success is adjudicated at the unit defined by each design: model history for formation, independent world for the controlled confirmation, and the frozen seed or branch for the corresponding manipulation and lifecycle experiments. Users, exposures, directions and checkpoints are nested measurements where the design specifies repeated observations; they are not substituted for independent histories. The twenty cross-seed directions arise from five realizations, the thirty architecture-transfer rows from directed cross-family pairs, and the eighteen cross-consequence rows from the frozen transfer family. Continuous readouts and gate outcomes are retained separately in the accompanying Source Data.

Ethics and governance status

This study performs secondary analysis of the publicly released KuaiRand dataset. The author had no interaction with the original users, performed no new participant recruitment, intervention or data collection, and analysed only the released research records, in which user and item identifiers were re-indexed by the dataset providers. Public documentation describes the platform-randomized exposure procedure used to generate the relevant interactions. We were unable to verify an authoritative public statement specifying the original collection's ethics-review, exemption or consent status, and attempts to obtain clarification from the dataset authors did not yield such a statement before submission. We therefore make no claim regarding an original approval, exemption or consent status that is not documented in the available public record.

Use of generative AI

Generative AI tools from OpenAI, including ChatGPT and Codex, were used during study planning and research discussions, to assist with code drafting and modification and the development and review of analysis

workflows, and in preparing the manuscript. The author reviewed and verified AI-assisted code, analyses and text, made the scientific judgments, and accepts full responsibility for the final content.

Data availability

The Version 2 source package retains the eleven existing Paper I derived source tables, including the recovered ten-run naturalistic manipulation table, and adds the archived architecture, coordinate and normal-form tables plus the complete recovered V19 and HAPT result tables. Original source bytes, archive-member paths and SHA-256 digests are recorded in the accompanying source manifest. These derived tables expose model-history, branch, direction and checkpoint units separately. They do not include model weights or original user-level KuaiRand records. The original KuaiRand random-policy data are available from the providers (DOI: 10.5281/zenodo.10439422). The separate experimental consolidation record at 10.5281/zenodo.22535187 was described as restricted with no downloadable files in the baseline availability statement; its current access status is not asserted here. The Version 2 source package accompanies this manuscript, without claiming a new DOI deposit. Historical outputs, later source recoveries and separately executed replications are not interchangeable provenance.

Code availability

The accompanying baseline figure-generation software reproduces the four unchanged main figures from their derived source tables. The Version 2 source package additionally includes the archived V20b/V21/V22 analysis sources, recovered tables and a deterministic assembly/verification script. It reproduces the displayed tables and numerical checks. It does not re-execute the original neural training or replace missing upstream inputs, checkpoints or full-curve files. Only Paper I materials are included; project-wide knowledge registers and separate mature-model cognition and future-control studies are not part of this package.

Acknowledgements

The author thanks Professor Yubiao Sun for helpful discussions and comments on this work, and the author's family for personal financial support that enabled the computational research.

Author contributions

Yuxuan Deng conceived the study, designed the experiments, developed and validated the code, analyzed the data, and wrote the manuscript.

References

[1] S. Kornblith, M. Norouzi, H. Lee and G. Hinton. Similarity of neural network representations revisited. *Proc. ICML*, PMLR **97**, 3519–3529 (2019).

[2] Y. Bansal, P. Nakkiran and B. Barak. Revisiting model stitching to compare neural representations. *Adv. Neural Inf. Process. Syst.* **34** (2021).

[3] J. Hewitt and P. Liang. Designing and interpreting probes with control tasks. *Proc. EMNLP-IJCNLP*, 2733–2743 (2019).

[4] A. Geiger *et al.* Causal abstraction: a theoretical foundation for mechanistic interpretability. *J. Mach. Learn. Res.* **26**(83), 1–64 (2025).

[5] R. E. Kalman. A new approach to linear filtering and prediction problems. *J. Basic Eng.* **82**, 35–45 (1960).

[6] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286 (1989).

[7] L. Ljung. *System Identification: Theory for the User*, 2nd ed. Prentice Hall (1999).

[8] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Lett.* **63**, 105–108 (1989).

[9] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: pattern and prediction, structure and simplicity. *J. Stat. Phys.* **104**, 817–879 (2001).

[10] M. L. Littman, R. S. Sutton and S. Singh. Predictive representations of state. *Adv. Neural Inf. Process. Syst.* **14** (2001).

[11] S. Singh, M. R. James and M. R. Rudary. Predictive state representations: a new theory for modeling dynamical systems. *Proc. UAI* (2004).

[12] K. Lenc and A. Vedaldi. Understanding image representations by measuring their equivariance and equivalence. *Proc. CVPR*, 991–999 (2015).

[13] S. K. Ainsworth, J. Hayase and S. Srinivasa. Git Re-Basin: merging models modulo permutation symmetries. *Proc. ICLR* (2023).

[14] A. Geiger, H. Lu, T. Icard and C. Potts. Causal abstractions of neural networks. *Adv. Neural Inf. Process. Syst.* **34** (2021).

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

[16] J. Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press (2003).

[17] K. Chalupka, P. Perona and F. Eberhardt. Visual causal feature learning. *Proc. UAI*, 181–190 (2015).

[18] K. Chalupka, F. Eberhardt and P. Perona. Multi-level cause-effect systems. *Proc. AISTATS*, PMLR **51**, 361–369 (2016).

[19] R. Givan, T. Dean and M. Greig. Equivalence notions and model minimization in Markov decision processes. *Artif. Intell.* **147**, 163–223 (2003).

[20] P. K. Rubenstein *et al.* Causal consistency of structural equation models. *Proc. UAI* (2017).

[21] S. Beckers and J. Y. Halpern. Abstracting causal models. *Proc. AAAI* **33**, 2678–2685 (2019).

[22] C. Gao *et al.* KuaiRand: an unbiased sequential recommendation dataset with randomly exposed videos. *Proc. CIKM*, 3953–3957 (2022). doi:10.1145/3511808.3557624.

[23] B. Schölkopf *et al.* Toward causal representation learning. *Proc. IEEE* **109**, 612–634 (2021).

[24] D. Bau *et al.* Understanding the role of individual units in a deep neural network. *Proc. Natl Acad. Sci. USA* **117**, 30071–30078 (2020).

[25] J. Vig *et al.* Investigating gender bias in language models using causal mediation analysis. *Adv. Neural Inf. Process. Syst.* **33** (2020).

[26] K. Wang, A. Variengien, A. Conmy, B. Shlegeris and J. Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. *Proc. ICLR* (2023).

[27] K. Meng, D. Bau, A. Andonian and Y. Belinkov. Locating and editing factual associations in GPT. *Adv. Neural Inf. Process. Syst.* **35** (2022).

[28] Y. Elazar, S. Ravfogel, A. Jacovi and Y. Goldberg. Amnesic probing: behavioral explanation with amnesic counterfactuals. *Trans. Assoc. Comput. Linguist.* **9**, 160–175 (2021).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- PaperIResearchSquarev2Supplement.pdf