CAUSAL SEPARATION: A THEORY OF REPRESENTATION AND RECURSIVE SELF-IMPROVEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Recursive self-improvement aims to let machines improve how they learn from experience. A system can continue to act on a familiar relation after the conditions supporting it have changed. The resulting failure exposes a dependence that sustained its activity without being represented. We characterize this problem by tracing the closed causal paths through which activity continues. An initial separation between represented and unrepresented relations leaves participating support on one side and its symbolic projection on the other. We show how this causal separation has corresponding environmental and representational forms. When the same represented state admits different consequences, we prove that their received separation cannot be generated by the old internal relations alone. A cycle of representational reproduction incorporates these symbolic effects into reusable relations and changes subsequent interaction. We derive its coupled response and conditions for learning without labels for hidden causes. In finite settings, the cycle recovers after support changes and removes error floors that further fitting within the old representation cannot overcome. For a two-layer linear family, the original gradient update rule executes the self-modifications selected from experience. With consistent estimation, selection, and execution, the system asymptotically achieves the lowest attainable expected loss after a fixed training period.

1 INTRODUCTION

Recursive self-improvement (RSI) aims to let machines improve how they learn from experience. Successful activity does not require an internal account of every condition that makes it work. When one of those conditions changes, the system can encounter a consequence its existing representations do not explain.

Consider a delivery robot that opens a service door by pressing a wall button. After maintenance interrupts the motor supply, the button still clicks but the door stays closed. The program executes correctly, yet the delivery fails. The rule still generates the same command, but the supply no longer makes that command effective. Detecting the failed operation does not require identifying the supply fault.

At a more fundamental level, the program and the physical conditions of its operation belong to one causal process. The represented rule helps generate the command; the runtime, circuit, motor, and supply determine what follows. Structural causal models describe these mechanisms at a chosen causal resolution (Bongers et al., 2021; 2022). The model includes all elements of the specified interaction, including computation, sensing, actuation, and physical support. An unrepresented mechanism can still determine the system's next state. The robot's representation is part of that process because it organizes an actual operation. It expresses a relation between pressing and opening without expressing every condition realizing that relation.

The same dependence appears in a mother–child interaction. A child's call elicits a response that helps organize the child's next activity, and that activity enters the interaction again. Consequences sustain further activity through this return, the property studied as organizational closure (Montévil & Mossio, 2015).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Tracing the return crosses a boundary in what is represented. The child's represented relation organizes the call, but the mother's activity completes its causal return. Neither side contains the complete return. This separation is already present in the causal structure; we term it *causal separation*. Its *initial separation* is between represented relations and their unrepresented causal environment. A *subject* is a representation-using organization sustained and reorganized through such returns.

A neural network's computation can be expressed as a causal model (Geiger et al., 2021). To study training, the same model includes data sampling, error calculations, optimizer state, and parameter updates (Appendix B.6). Sensing and actuation likewise appear as causal paths. A mechanism belongs to the unrepresented causal part when its determining relations are absent from the system's current representations. An optimizer can therefore change weights without its update rule being represented by the network. These causal mechanisms already specify how an operation changes the environment and how its consequences return. The separation follows causal roles, not hardware boundaries.

The mother's response is part of the child's return, although her other interactions continue beyond it (De Jaegher & Di Paolo, 2007). Participation in this closure separates the unrepresented support from those other causal relations. This is *environmental separation*. The received response occupies the other end of the same path. It enters the child's activity without an internal account of the relations producing it. This is the corresponding *representational separation*. When the mother turns to answer the father, an unchanged expectation meets a changed response. The event makes the separation consequential before its cause has been understood. Infants' sensitivity to interrupted responsiveness motivates this example (Tronick et al., 1978).

Binding research connects systematic generalization to three linked problems: forming representations from experience, maintaining their separation, and composing them into new inferences (Greff et al., 2020). Separation allows knowledge to be reused without losing which entity a property or relation concerns. We follow these requirements through a causal process, from causal separation to the relations subsequently built from its received effects.

The father's influence is already present in the child's changed experience, although the relation producing it has not been represented. It has a *projection* within the child's represented causal structure. A shadow likewise carries an effect without containing the process that produces it. *Symbolic representations* are these projections of unrepresented causal influences within the system. *Relational representations* express the causal dependencies the system can follow and revise internally. The projection brings an external effect into the represented process without bringing its complete determining relations with it. Figure 1 follows an unrepresented cause into an exposed effect and its symbolic contact. This symbol opens a new possibility and is subsequently connected into the existing closure.

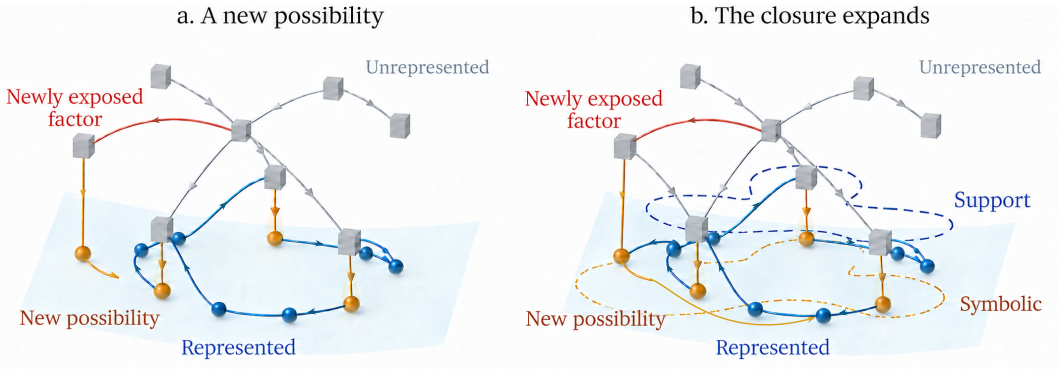


Figure 1: An exposed effect carries an unrepresented causal dependence into the system. Left: hidden relations produce the exposed effect, whose contact with the represented surface is symbolic and opens a new possibility. Right: this symbol is connected back into the ongoing closure, making the new consequence available for comparison and renewed use. The two dashed contours mark participating support and symbolic contributions on the two sides of the initial separation. The surface displays representational scope, not a physical enclosure.

Representational reproduction develops this coupling over successive encounters. An externally determined effect changes the associations of an existing representative. Comparison relates these effects across encounters and separates their conditions. Reconstruction makes the resulting dependencies usable in a new prediction or operation. The symbolic effect becomes part of the relational organization of further activity. This changed activity again acts on the unrepresented processes through which it takes effect. The same cycle can incorporate the conditions and consequences of learning operations, allowing an acquired relation to reorganize how the system learns next.

Contributions.

- **Causal separation (C1).** We characterize causal separation in actual causal returns and derive its environmental and representational forms. We prove when the old internal relations cannot determine the received effects (Section 2).
- **Representational reproduction (C2).** We derive the causal order and coupled response of the cycle. Our finite constructions form new relations without hidden-cause labels, recover usable relations after support changes, and remove error floors imposed by the old representation (Section 3).
- **Recursive self-improvement (C3).** For the studied two-layer linear class, we prove that self-modification can be executed through the original gradient update rule using internally generated targets. With consistent estimation, selection, and execution, expected loss after a fixed training period converges to the attainable minimum (Section 4).

2 Causal Separation and Representation

2.1 The initial separation

The robot's rule, the computation issuing its command, and the circuit opening the door all help determine what happens. We place them in one causal domain, together with processes that can change their operation. At a declared temporal resolution, a dynamic structural model writes

$$X^i_{t+1} = F_i(X^{\mathrm{pa}(i)}_t, U^i_{t+1}), \qquad X_0 \sim \mu_0. \tag{1}$$

The parent relations describe direct dependencies, including computational and physical mechanisms (Bongers et al., 2021; 2022). Execution, sensing, actuation, data selection, and parameter updates are all mechanisms in Equation 1. Their paths determine how activity changes the environment and how consequences return to the system. This is a model of the actual process, not knowledge supplied to the system.

Within that process, an activity can help sustain the conditions of its continuation. A child's call, a caregiver's response, and the child's ensuing activity can form such a return. Organizational closure concerns the mutual maintenance realized through these returns (Montévil & Mossio, 2015). Closure identifies an organization within the common causal domain, even when its return passes through processes it does not represent.

Tracing these returns reveals *causal separation* in the underlying causal structure. The separation depends on which relations participate in the closure and where their return is completed. Appendix B.3 expresses this property through the topology of the traced paths. For representational activity, the *initial separation* is between represented relations and their unrepresented causal environment. The actual return crosses this boundary. A representation is itself causal activity used to express conditions and consequences of a process; it can belong to that process and need not be simpler than it. The separation concerns causal roles, rather than the physical location of their carriers. An unrepresented runtime can therefore lie inside the machine.

A *subject* is a representation-using organization sustained through such returns. Its consequences help enable and organize further activity, including reconstruction after a disruption. Appendix B.2 specifies this continuation through actual mechanisms, resources, initial states, and positive-time returns. A causal path back to an earlier variable is insufficient if the next operation is no longer enabled. Appendix B.4 specifies the actual source and use of a representation.

Neural computation can be represented directly as a causal model (Geiger et al., 2021). Predictions enter error calculations, and updates change the state used for later predictions. These operations can

sustain a training process under its input and execution conditions (Appendix B.6). An optimizer can therefore act within the machine without its rule being represented in the prediction it changes.

2.2 Environmental and representational separation

Environmental separation is relative to the subject's closure. Some unrepresented relations participate in its return; other unrepresented relations lie outside that return. The participating support cannot complete this particular return within the unrepresented side alone. Its continuation passes through the subject's activity. Maternal responsiveness has this role in the child's closure, even though the mother's activity also participates in other closures.

Representational separation is the corresponding separation on the represented side. Relational representations express dependencies that can be followed internally. Symbolic representations carry effects whose determining paths pass through unrepresented relations. The symbol belongs to represented activity, but its causal return is not contained within represented relations alone. The same cross-boundary closure accounts for both forms of separation (Figure 1).

At the resolution of recurring causal roles, each side contains only part of a cross-boundary return. Its completion requires the other side (Appendix B.3). The mother's response and its symbolic projection occupy the two sides of the same return. When she turns to answer the father, a changed response makes the separation consequential for the child's expectation. This event exposes the separation; the closure already supplies its relative structure.

In a deterministic realization, let r be the operative represented state and e the remaining causal conditions. For an operation a, with other inputs fixed or included, write

$$r^+ = F_a(r, e), \qquad e^+ = G_a(e, r). \tag{2}$$

At fixed r, effects through F_a determine which external conditions must remain separated. At fixed e, outward effects through G_a determine which internal states must remain separated. A representation preserving either response must preserve the separation of conditions that produce different values of that response (Appendix B.7). Changing the operation or its support can therefore change the response separation required.

2.3 Symbolic projections and relational structure

A door remaining closed can interrupt the robot's activity before the supply condition is understood. The unrepresented cause has an effect within the represented process. This is a causal projection: the effect enters the system's activity, but the relations determining it have not been expressed there. The projected part is *symbolic*; the causal dependencies that can be expressed and revised internally form the *relational* part. The separation is already realized by the causal interaction.

The receiving mechanism makes this projection precise. Let $r = \pi(x)$ be the operative representation of the full causal state x, $z = E_a(x)$ the received signal, and T_a the executed transition. With every other receiver parent fixed or included, the actual equation is

$$\pi(T_a x) = U_a(\pi(x), E_a(x)), \qquad \Gamma_a(r) = \{\pi(T_a x) : \pi(x) = r\}. \tag{3}$$

If $\Gamma_a(r)$ contains different next states, the same old r admits different received effects. A function of r alone cannot determine which one occurs. Two signals have the same effect at this resolution precisely when $U_a(r, z)$ is the same. Their response classes correspond one-to-one to the next states in $\Gamma_a(r)$ (Appendix B.9). This gives the received separation an operative identity in the process that already runs.

The response classes describe the symbolic part of this separation at the receiving operation. They record the separation of projected effects without requiring an internal reconstruction of their causes. Binding research studies how representations can remain separately identifiable as their relations change (Greff et al., 2020; Hummel & Holyoak, 1997). Here these identities are present in the effects already received. Relational reconstruction subsequently connects those effects to conditions the subject can use.

The received separation can constrain further relations. If two available histories grouped by the same representation have different consequence laws, a common prediction cannot fit both. At least

one prediction error in total variation is half that distance or greater (Appendix A.7). The subsequent reconstruction must make the relevant difference usable.

An *environmental remainder* is the causal influence left unexpressed by the current relations. Its received effect is already part of the subject's activity. An encounter connects an adopted reference, an operation, available conditions, and that effect. The reference can remain identifiable as comparison changes its associated relations. The next section follows this effect into a reconstructed relation and then back into interaction with the processes supporting activity.

3 A Cycle of Representational Reproduction

3.1 From an unexplained effect to a usable relation

An adopted representative organizes activity through unrepresented support. The activity can change that support, and changed support can alter what the same operation produces. The cycle follows how one such effect becomes a relation organizing another encounter. A command once associated only with success can also become associated with interruption or delay. The environmental remainder acts through this difference. Reception links the adopted reference c, the executed operation a, available conditions h, and the received representative σ. The reference acquires associations with new consequences, and the consequence acquires associations with the conditions under which it occurred.

Comparison uses these associations in the reverse direction. The reference that organized an encounter now becomes an object compared through its consequences. The same operation can be compared across conditions, and different operations can be compared through a common outcome. Shared associations can be expressed together; different responses can be distinguished by their conditions. Binding makes such comparison possible without losing the identities being related (Greff et al., 2020; Hummel & Holyoak, 1997).

Handwritten digit recognition presents a familiar demand: different handwritings should support the same digit decision, although different digits can share local strokes. Accumulating examples under a category and organizing the relations needed to distinguish it are different operations.

Reconstruction makes the compared dependencies available for renewed use. The result is more than a record that two encounters differed: a condition can now determine which prediction or operation is used. The received effect has acquired a place in a conditional relation. That relation can itself be adopted as a representative in another encounter. The reconstructed relation directs renewed interaction with the processes supporting activity. The two-sided separation in Figure 1 thus develops into a cycle of *representational reproduction*. The effect becomes usable without requiring a complete internal model of its causes.

The four positions follow the information dependencies of the same encounter. Adoption uses a representative to organize it (a); reception adds its actual consequence (b). Comparison works on these available relations (c); reconstruction makes the resulting relation usable again (d). Use precedes its consequence, and a reconstruction based on that consequence follows reception and comparison. Appendix B.15 proves this order for the acquisition and deployment of information from one encounter. Earlier encounters supply the material for later adoptions (Figure 2). The roles recur as the organization changes; the complete state need not return to its starting value. Unrepresented causes can affect every position and the transitions between them.

3.2 A new separation changes earlier comparisons

Predictive-state reconstruction groups histories by their future consequences (Shalizi & Shalizi, 2004), and model minimization compares states through outcomes and transitions to equivalence classes (Givan et al., 2003). We use these ideas to give an explicit realization in which an operating system records, reconstructs, and installs the resulting representational separation. The available conditions H and executable operations A are finite. Each operation has a stable conditional outcome mean $g(h, a)$ and transition law $P_a(h, h')$ under the allowed policies. Relevant implementation changes are included in the recorded configuration.

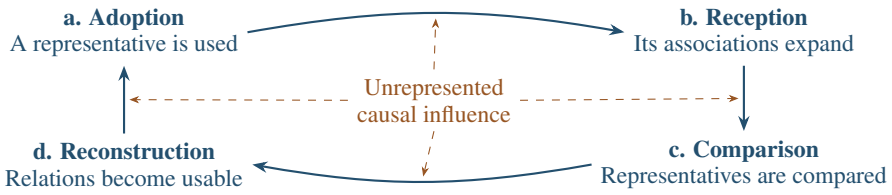


Figure 2: Unrepresented causes can affect every phase of representational reproduction. A reference organizes an encounter; its consequences expand associations, comparison relates encounters, and reconstruction makes a conditional relation usable again. Blue arrows track the ordered roles of one encounter. Dashed arrows mark possible influences throughout the cycle, including transitions, rather than separate input channels. Different cycles can overlap.

In the finite device realization (Appendix B.11), previously merged sensor readings acquire different predicted consequences, and the installed relation changes the next operation.

The general comparison must include what an operation leads to, not just its immediate outcome. For a current grouping Π, its response profile is

$$\chi_\Pi(h) = \big(g(h,a),\, P_a(h,B) : a \in A,\; B \in \Pi\big). \tag{4}$$

Suppose two routes initially reach the same *service area* category. Once powered and unpowered destinations are distinguished, the two routes can have different probabilities of reaching them. Equation 4 then distinguishes the routes as well. A new representational separation changes the comparisons through which further representations form. For multiple received codes, the profile also includes their individual response probabilities. Separation of one consequence thus constrains the earlier conditions that lead to it. Appendix A.5 proves that this refinement completes the smallest function algebra closed under the declared outcomes and transitions. The unresolved dependence determines what must be added; already usable relations are retained.

Actual returns update statistics for the conditions and operations that produced them. The system estimates the profiles, reconstructs a partition from its initial grouping Π_0, and installs the resulting prediction and decision variables. Later evidence starts another reconstruction, allowing early separations caused by noise to be withdrawn. With bounded results, persistent visits to the relevant condition–operation pairs, and continuing recording, reconstruction, and deployment, we prove almost-sure eventual recovery of the coarsest stable refinement of Π_0 (Appendix A). Neither that partition nor its separation gaps are supplied. Within this controlled interface, the system learns the representational separation required by its consequences and makes it available for use. This same rule can repair a relation after its support changes. If the controlled laws stabilize after an unannounced finite change, the accumulated records and unchanged reconstruction rule eventually recover the new stable partition (Corollary 4). The new returns correct the grouping through the ongoing cycle; no change notification, record reset, or replacement representation is supplied.

3.3 A RELATION CAN BE AVAILABLE IN EXPERIENCE BUT ABSENT FROM THE UPDATE

Grouping conditions is one way to reconstruct relations. A second limitation arises when the records remain distinct, but the active prediction functions cannot express their joint effect. Two observed signs can, for example, determine a response through their product even though an additive prediction uses each sign separately. More accurate coefficients within the additive class do not create the missing interaction. The comparison must change what the system can express.

Let $E \subseteq L^2(\mu)$ be the finite-dimensional space of currently usable prediction functions, and let P_E be its orthogonal projection. For a square-integrable scalar return Y and a finite available record H with positive probabilities μ, define

$$g(h) = \mathbb{E}[Y \mid H = h], \qquad u = (I - P_E)g, \qquad E^+ = E \oplus \operatorname{span}\{u\}. \tag{5}$$

Here u is the predictable part of experience that the current functions cannot express. We prove that their minimum excess half squared risk is $\frac{1}{2}\|u\|_\mu^2$, and that E^+ is the smallest linear extension removing this error floor. Conditional averages from independent paired episodes with full support

consistently estimate g, so the new function can be formed without labels for the hidden factor (Appendix D.3). The extension retains the old functions, preserving shared structure.

Installing the new function turns an object of comparison into a condition of subsequent learning. For $u \neq 0$, normalize the new component as $\psi = u/\|u\|_\mu$. In orthonormal readout coordinates, the update kernel becomes

$$K_{\text{new}}(h, h') = K_{\text{old}}(h, h') + \psi(h)\psi(h'). \tag{6}$$

An error at one record can now change another prediction through the acquired relation. Thus reconstruction changes the way experience acts on the system. At the old optimum $f_0 = P_E g$, the old projected gradient has stopped even though $u \neq 0$. After installing the reconstructed component, the same readout update gives $\dot{f}(t) = e^{-t}u$ and converges to g (Corollary 18). All remaining correction in this realization follows the formerly unavailable direction. The gap becomes an effective cause of change through the return that incorporates it. Appendix D.3 also gives finite-sample guarantees for an estimated component. The new relation can subsequently enter an adopted operation, exposing consequences that the earlier operation family did not distinguish (Appendix A.9).

3.4 Reconstruction and generalization

Reconstruction gives experience a use beyond the encounter that produced it. Comparison separates the conditions and relations available in past encounters; reconstruction makes them jointly determine a prediction for a new one. A symbolic argument can stand for different concrete realizations, so the subject can estimate a consequence without reconstructing every mechanism that will produce it. The mother's responses, for example, can become partly predictable without her activity becoming fully transparent to the child.

The validity of this reuse depends on the causal differences the representation leaves unresolved. Across the declared contexts, a common prediction for each represented state is valid exactly when contexts represented alike have the same consequence law under every allowed operation (Appendix F.2). It connects the compositional reuse emphasized by binding to the stability of its unrepresented support; causal transportability studies this latter requirement across environments (Pearl & Bareinboim, 2014).

Unknown details can therefore remain unknown when their differences preserve the required consequence law. When they cease to be equivalent, the same representation no longer supports a common correct prediction. If both contexts must still be handled, their difference becomes a new requirement for separation and reconstruction. The return thus gives rise both to a range of generalization and to evidence about where that range fails. The finite constructions estimate relations on their sampling domain. For a new context sharing an already learned representation, Appendix F.2 separates source estimation error from a difference in consequence laws. It also relates reuse to tasks not specified during formation.

3.5 One coupled response, successive changes of role

The same return also has a dynamical description. Consider a smooth implementation with symbolic coordinates s, relational coordinates ℓ, and environmental coordinates e. Further unrepresented conditions w affect their joint dynamics $\dot{x} = F(x, w)$, where $x = (s, \ell, e)$. For a fixed initial state and a differentiable adoption $s^+ = \mathcal{A}(\ell(T))$, variation of the dynamics gives

$$\delta s^+ = D\mathcal{A}(\ell(T)) \int_0^T \mathcal{U}_{\ell \leftarrow w}(T, t)\, \delta w(t)\, dt. \tag{7}$$

The propagator $\mathcal{U}_{\ell \leftarrow w}$ follows how an external change reaches the evolving relations, including intermediate interactions through the environment. The adoption derivative then converts that relational change into a change in the next representative. Equation 7 expresses how the return carries an unrepresented influence into renewed representational use.

Expanding the joint propagator as time-ordered products of local Jacobians resolves the successive interactions along this movement. With symbolic and relational role projections P_s, P_ℓ, its internal response consists of $P_i U P_j$ for $i, j \in \{s, \ell\}$. These four contributions describe effects within and between the two roles; each can contain many intermediate returns. The four positions above track when an experience becomes available for a new use, and this expansion computes how influence

propagates through the interacting representations. Appendix D derives the expansion and the corresponding adoption-event formulas. The finite refinement and response-completion constructions specify how the relations carried by this movement become usable in prediction and learning.

The same dynamics express the continued participation of unrepresented support. Eliminating environmental coordinates leaves a history-dependent memory term and an unresolved driving term in the linearized response dynamics (Appendix D); such effects are standard in projection methods (Chorin et al., 2000). Reconstruction can make their response regularities usable without replacing the mechanisms that realize them. The subject's acquired ability is therefore jointly realized through represented relations and their continuing causal support.

4 Learning Mechanisms and Recursive Self-Improvement

4.1 When a learned relation changes how learning proceeds

The handwriting example concerned relations that make a new input recognizable. Those relations affect a decision because they are used in producing it. The same requirement applies to learning itself: a relation about the consequences of an update must enter the process choosing or executing an update. The causal model introduced in Section 2 already includes that process. Learning operations have conditions and consequences of their own, so they can become objects of the same comparison and reconstruction.

Backpropagation illustrates how an expressed result already constrains its producer. For a prediction $\widehat{y} = f_\theta(x)$, a target y determines an error signal and hence a parameter change:

$$\widehat{y} = f_\theta(x), \qquad e = \nabla_{\widehat{y}}\ell(\widehat{y}, y), \qquad g = J_\theta f_\theta(x)^\top e, \qquad \theta^+ = \theta - \eta g.$$

The execution context connects this error to the prediction and parameters that produced it (Baydin et al., 2018). A self-generated target can occupy the same position. Stop-gradient fixes its contribution during differentiation even as the target-generating parameters remain changeable across updates (Tian et al., 2021). These mechanisms show how a representation produced within the system can constrain later changes to that system.

The neural realization below makes this requirement testable. A probe setting A is chosen before the first return $Z = Y_1$ arrives. The observed pair $H = (A, Z)$ is then available for predicting Y_2. Its conditional mean contains an interaction βAZ that the original space $\mathrm{span}\{1, A, Z\}$ cannot express. Reconstruction forms this interaction from paired events and adds it to the usable representation. The relation then guides a parameter preparation that changes the response to future training. The received effects have become a condition for how the machine will learn next.

Missing parts of the cycle have specific learning consequences. If all retained observations merge two environments requiring different preparations, the useful choice remains unlearnable. If observations retain the difference but the active functions cannot express it, more accurate fitting within those functions does not remove the error floor. If the relation is formed but never used to change the learning process, it does not improve the successor. The following results connect all three requirements within specified executable processes.

In the finite controlled setting of Section 3.2, consistent estimation and vanishing planning error let the selected rule approach the optimal discounted value within the declared operation family (Appendix F.6). Recursive use preserves recoverable sources, informative interaction, reconstruction, and deployment in each successor. Continued exploration has a separate cost.

4.2 Self-modification through the original neural update

Parameter-symmetry work shows that preserving a current function or loss need not preserve later optimization (Zhao et al., 2024). We characterize the preparations attainable through a fixed gradient update rule and prove that feedback can identify preparations that improve subsequent learning. The analyzed organization includes calibration, relation formation, and execution of the selected preparation; its weights are part of that process.

Take a two-layer linear network at unit scalar input, $f_{W,v}(1) = Wv$, initially $W = D = \mathrm{diag}(1, 2)$ and $v = 0$. For half squared loss with target y, its output evolves as

$$\dot{f} = K(y - f), \qquad K = WW^\top + \|v\|^2 I.$$

The response matrix K determines how strongly each error direction changes the prediction. Initially the prediction is zero and $K = D^2 = \mathrm{diag}(1, 4)$. Internally chosen training targets move the parameters through the unchanged gradient rule. The target changes; the update formula does not. The attainable returned endpoints are exactly $(QD, 0)$ with $Q \in SO(2)$, giving response $QD^2Q^\top$. Thus preparation can rotate the stronger learning response toward a useful direction and restore the prediction to zero. It reallocates sensitivity across directions. Appendix E proves the complete orbit in arbitrary output dimension and constructs its training paths. Both branches can execute these paths; experience determines their selection.

Calibration records connect a probe setting to its two returned measurements. Their joint relation identifies which of two possible lines will contain the future target. The target has an independent random sign, revealed only after preparation; learning the line already tells the system where to direct its stronger response. Removing the probe identity makes the two environments observationally identical and leaves this choice unlearnable. The system estimates the relation from paired records, compares attainable orientations over the fixed adaptation period, and executes the selected preparation through this gradient update rule. Appendix E.7 specifies the observation law and implementation.

At adaptation period 0.08, the complete process and a no-deployment control both reach prediction loss 0.187. Deploying the acquired relation reduces successor loss from 0.688 to 0.525, excluding preparation costs. Removing source correspondence or fixing the original feature class blocks the process earlier. Appendix E.7 reports the controls and numerical results, and Appendix F.7 states the conditions for recurring use.

We prove that this difference persists beyond finite data. Source-erased preparation has positive excess successor loss for sufficiently short fixed adaptation periods, regardless of its data count. With retained sources and increasing estimation, search, and execution accuracy, the selected successor satisfies

$$\mathbb{E}L^{\mathrm{actual}}_{\theta,N}(\tau) \longrightarrow \mathbb{E}\min_{p\in\mathcal{P}} L_\theta(p;\tau). \tag{8}$$

Here θ is the environmental configuration, $\mathcal{P}$ the attainable preparation family, and τ the fixed adaptation period. Experience therefore teaches the system to select and execute the preparation that gives the lowest attainable expected loss after subsequent learning.

5 Discussion and Conclusion

The door failure exposed a condition participating in the robot's causal closure. The initial separation between represented relations and their unrepresented environment gives this closure a relative boundary. Environmental separation concerns the unrepresented support participating in the return. Representational separation concerns its symbolic effects within represented activity. These are paired forms of causal separation, defined by the same closure. Relational reconstruction makes the received effects usable and reorganizes subsequent interaction. The subject continues through this return, still dependent on processes it has not represented in full.

The learning consequence is concrete. A difference merged by the current grouping, or a relation absent from its active functions, can remain beyond further fitting in that representation. Our constructions show when the cycle forms the missing separation, makes it usable, and changes subsequent responses. Applied to learning operations, the acquired relation can guide an actual modification of the learning process. In the studied neural class, beneficial preparations are executed through the original gradient update rule with internally generated targets. As estimation, selection, and execution become accurate, fixed-period successor loss approaches the attainable minimum. Recursive use requires the modified system to continue receiving informative effects and reconstructing and using their relations.

AI USE STATEMENT

AI tools assisted with literature retrieval, research ideation, mathematical derivation and checking, manuscript drafting and revision, and preparation of code, document sources, and the conceptual illustration in Figure 1. Their suggestions were examined against primary sources, explicit derivations, and numerical checks. Responsibility for the submitted claims and references rests with the authors.

ETHICS STATEMENT

This work contains no human-subject study, private dataset, or autonomous deployment. Learning ability alone supplies no safety or normative guarantee.

REPRODUCIBILITY STATEMENT

The appendices specify the models, proofs, assumptions, and numerical protocol. Code implements the empirical response-formation and neural preparation test.

REFERENCES

Awni Altabaa, Taylor Webb, Jonathan D. Cohen, and John Lafferty. Abstractors and relational cross-attention: An inductive bias for explicit relational reasoning in transformers. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6b070264f20d25fbd1488de8ea51575f-Paper-Conference.pdf.

Garrett Andersen and George Konidaris. Active exploration for learning symbolic representations, 2017. URL https://arxiv.org/abs/1709.01490.

Xabier E. Barandiaran, Ezequiel Di Paolo, and Marieke Rohde. Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, 17(5):367–386, 2009. doi: 10.1177/1059712309343819.

Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research*, 18(153):1–43, 2018. URL https://jmlr.org/papers/v18/17-468.html.

Sander Beckers and Joseph Y. Halpern. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. URL https://ojs.aaai.org/index.php/AAAI/article/view/4117.

J. Nicholas Betley, Shengjin Xu, Zhen Fang Huang Cao, Rong Gong, Christopher J. Magnus, Yang Yu, and Scott M. Sternson. Neurons for hunger and thirst transmit a negative-valence teaching signal. *Nature*, 521:180–185, 2015. doi: 10.1038/nature14416.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021. doi: 10.1214/21-AOS2064.

Stephan Bongers, Tineke Blom, and Joris M. Mooij. Causal modeling of dynamical systems, 2022. URL https://arxiv.org/abs/1803.08784.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. Adapting language models to compress contexts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3829–3846, 2023. doi: 10.18653/v1/2023.emnlp-main.232. URL https://aclanthology.org/2023.emnlp-main.232/.

Alexandre J. Chorin, Ole H. Hald, and Raz Kupferman. Optimal prediction and the Mori–Zwanzig representation of irreversible processes. *Proceedings of the National Academy of Sciences*, 97(7):2968–2973, 2000. doi: 10.1073/pnas.97.7.2968.

Hanne De Jaegher and Ezequiel Di Paolo. Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences*, 6:485–507, 2007. doi: 10.1007/s11097-007-9076-9.

Leonidas A. A. Doumas, John E. Hummel, and Catherine M. Sandhofer. A theory of the discovery and predication of relational concepts. *Psychological Review*, 115(1):1–43, 2008. doi: 10.1037/0033-295X.115.1.1.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/fe131d7f5a6b38b23cc967316c13dae2-Abstract.html.

Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sablé-Meyer, Lucas Morales, Luke Hewitt, Luc Cary, Armando Solar-Lezama, and Joshua B. Tenenbaum. DreamCoder: Bootstrapping inductive program synthesis with wake-sleep library learning. In *Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation*, 2021. URL https://www.neurosymbolic.org/papers/EllisWNSMHCST21.pdf.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2106.02997.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003. doi: 10.1016/S0004-3702(02)00376-4.

William M. Goldman. Geometric structures on manifolds, 2021. URL https://math.umd.edu/~wmg/gstom.pdf.

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks, 2020. URL https://arxiv.org/abs/2012.05208.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/1911.05722.

John E. Hummel and Keith J. Holyoak. Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104(3):427–466, 1997. doi: 10.1037/0033-295X.104.3.427. URL https://reasoninglab.psych.ucla.edu/wp-content/uploads/sites/273/2021/04/Hummel-Holyoak.1997.pdf.

Michael Kirchhoff, Thomas Parr, Ensor Palacios, Karl Friston, and Julian Kiverstein. The markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, 15(138):20170792, 2018. doi: 10.1098/rsif.2017.0792.

George Konidaris. Constructing abstraction hierarchies using a skill-symbol loop, 2015. URL https://arxiv.org/abs/1509.07582.

Maël Montévil and Matteo Mossio. Biological organisation as closure of constraints. *Journal of Theoretical Biology*, 372:179–191, 2015. doi: 10.1016/j.jtbi.2015.02.029. URL https://montevil.org/publications/articles/2015-mm-organisation-closure-constraints/.

Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014. doi: 10.1214/14-STS486. URL https://ftp.cs.ucla.edu/pub/stat_ser/r400-reprint.pdf.

Jürgen Schmidhuber. Gödel Machines: Self-Referential Universal Problem Solvers Making Provably Optimal Self-Improvements, 2006. URL https://sferics.idsia.ch/pub/juergen/gm6.pdf. Version dated 17 December 2006.

Cosma Rohilla Shalizi and Kristina Lisa Shalizi. Blind construction of optimal nonlinear recursive predictors for discrete sequences, 2004. URL https://arxiv.org/abs/cs/0406011.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10268–10278, 2021. URL https://proceedings.mlr.press/v139/tian21a.html.

Edward Tronick, Heidelise Als, Lauren Adamson, Susan Wise, and T. Berry Brazelton. The infant's response to entrapment between contradictory messages in face-to-face interaction. *Journal of the American Academy of Child Psychiatry*, 17(1):1–13, 1978. doi: 10.1016/S0002-7138(09)62273-1.

Taylor Webb, Ishan Sinha, and Jonathan D. Cohen. Emergent symbols through binding in external memory. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2012.14601.

Bo Zhao, Robert M. Gower, Robin Walters, and Rose Yu. Improving convergence and generalization using parameter symmetries. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.13404.

Christopher A. Zimmerman, Yen-Chu Lin, David E. Leib, Ling Guo, Erica L. Huey, Gwendolyn E. Daly, Yiming Chen, and Zachary A. Knight. Thirst neurons anticipate the homeostatic consequences of eating and drinking. *Nature*, 537:680–684, 2016. doi: 10.1038/nature18950.

A A FINITE CONTROLLED REALIZATION OF REPRESENTATION SEPARATION

A.1 THE CAUSAL PROCESS AND EXECUTION ASSUMPTIONS

A time-unrolled structural causal model has assignments $X^i_{t+1} = F_i(X^{\mathrm{pa}(i)}_t, U^i_{t+1})$, an initial state $X_0 \sim \mu_0$, and a specified joint law for X_0 and its exogenous processes. The state includes the execution state of programs, parameters, data acquisition, update application, and the relevant environmental support. A change of program is a state change for the total execution mechanism. An intervention replaces a declared assignment, rather than merely conditioning on its output. Dynamic structural causal models provide the general setting for this interpretation (Bongers et al., 2022).

The algorithmic realization below requires complete execution of its operating cycle. The operations needed to acquire a result, update the internal record, form a representation, deploy it, and initiate another interaction are realized by the model. These scheduling requirements give a sufficient implementation of the organizational continuation condition in Appendix B.2; they do not require a periodic state trajectory. A supplied resource can support execution without being represented by the machine.

Fix a finite acyclic graph of within-round execution dependencies. If its inputs arrive, each internal operation terminates, the scheduler eventually executes every enabled operation, and the installed result initiates the next round, then that round and its restart finish in finite time. To prove this, order the within-round dependencies topologically and apply finite termination in that order. Induction proves completion of any specified finite number of rounds whenever the same conditions continue to hold. This execution statement supplies no statistical success by itself.

A Markov blanket imposes a different property (Kirchhoff et al., 2018): for specified variables and a specified distribution, $I \perp E \mid B$. It follows from neither execution completeness nor causal return, and is not itself a learning guarantee. The manuscript uses the causal execution conditions directly.

A.2 INITIAL STATE AND AN EMBODIED REALIZATION

An initial distribution μ_0 specifies the joint state from which execution begins. For an embodied interpretation, it can include environmental state e, bodily maintenance variables b, neural activity n, and a plastic organization θ. These are coordinates at a declared modeling scale: b is not the whole organism with n and θ counted again. A representation can be a functional readout $r = \alpha_\theta(n)$ used in prediction or control. The choice of this readout identifies a representational use; naming a physical variable does not establish that use.

One continuous causal organization compatible with this interpretation is

$$\dot{b} = F_b(b, e, a), \qquad \dot{n} = F_n(n, b, e; \theta),$$
$$\dot{\theta} = G(\theta, n, b, e), \quad a = \pi(n, b, \theta), \qquad \dot{e} = F_e(e, a).$$

Noise and further support processes can be included as specified inputs. The equations express paths from bodily and environmental changes into neural activity, from activity into action, and from action back into bodily conditions. Under suitable regularity they can be given structural dynamical causal semantics (Bongers et al., 2022). Digital updates and discrete resets instead require their own time-unrolled or hybrid description; continuous-process regularity is not assumed automatically for them.

The same return can regulate present activity and reorganize how later activity proceeds. A change in b can restore a maintained quantity through existing relations. When the encounter also leaves a retained change in θ that is used in a comparable situation, the return has altered those relations. Learning is this persistent reorganization within the return, rather than an additional movement after regulation is complete. The biological findings cited in the main text motivate particular connections between need signals, conditioning, and ongoing behavior (Betley et al., 2015; Zimmerman et al., 2016). They do not show that every instance of this causal template learns, or that biological systems implement the partition algorithm below.

Initial states differ in the observations, actions, resources, and plasticity they make available. A learning claim must specify a region in which its required operations remain possible for the analyzed interval. The present automatic-formation theorem expresses this through continued sampling,

information retention, and completed reconstruction and deployment. It does not derive those conditions from a physiological deficit. Doing so would require additional properties of F_b, F_n, G, π and the accessible environment, including evidence that the induced activity is both informative and sustainable.

For the statistical implementation below, one explicit initial state consists of a permitted record h_0, initial partition Π_0, zero counts and result sums, a specified exploratory policy, and an enabled execution state. The initial policy and subsequent exploration must realize the coverage condition. Thus this is an autonomous learning realization with a declared starting organization, not a claim that its observation and update machinery is generated from an initially unorganized system.

A.3 Successful prediction cannot retain a harmful merge

Let H and A be finite sets of available records and operations. Let $g(h, a)$ be the conditional mean of the declared result under operation a. An interventional interpretation requires the corresponding causal sampling condition. Let $\mu(h, a) > 0$ be a fixed evaluation distribution. All information used by a predictor is included in its effective representation: $f_n(h, a) = d_n(q_n(h), a)$. There is no additional decoder input carrying separations omitted by q_n. Define the squared excess prediction risk

$$\mathcal{E}_n = \sum_{h,a} \mu(h, a)\big(f_n(h, a) - g(h, a)\big)^2.$$

The representation and decoder may depend on the training history.

Proposition 1 (Separation forced by consistency)**.** *If $\delta = g(h, a) - g(h', a) \neq 0$, then on the event $q_n(h) = q_n(h')$,*

$$\mathcal{E}_n \geq c_{h,h',a} := \frac{\mu(h, a)\mu(h', a)}{\mu(h, a) + \mu(h', a)}\delta^2 > 0.$$

Consequently, $\mathcal{E}_n \to 0$ almost surely implies that the pair is eventually never merged again, almost surely. If only $\mathbb{E}\mathcal{E}_n \to 0$ is known, the merge probability tends to zero.

Proof. On a merge both predictions equal some z. Writing $u = \mu(h, a)$ and $v = \mu(h', a)$, their contribution is

$$u(z - g)^2 + v(z - g')^2 = (u + v)\left(z - \frac{ug + vg'}{u + v}\right)^2 + \frac{uv}{u + v}(g - g')^2.$$

All remaining terms are nonnegative. Almost-sure convergence eventually puts the risk below the fixed positive lower bound. For convergence in expectation, the merge probability is at most $\mathbb{E}\mathcal{E}_n / c_{h,h',a}$. □

The almost-sure and expectation conclusions must not be interchanged. Independent merges with probabilities $1/n$ can occur infinitely often even though their expected risk vanishes. The proposition concerns operative separation, not two physical modules. It also does not supply the learning rule that achieves consistency.

A.4 A stable controlled process observed during operation

Let $H = \{1, \ldots, m\}$ and let A be a finite common set of executable operations. At interaction t, the system observes h_t, chooses a_t, and receives $Y_t \in [0, 1]$ and h_{t+1}. The filtration $\mathcal{F}_t$ includes the choice of a_t but precedes the new result. For every allowed non-anticipating operation policy or operation intervention, assume the same policy-invariant causal laws

$$\mathbb{E}[Y_t \mid \mathcal{F}_t] = g(h_t, a_t), \qquad \mathbb{P}(h_{t+1} = x \mid \mathcal{F}_t) = P_a(h_t, x).$$

The conditional means and controlled kernel are fixed during the analyzed learning regime. The operation policy may depend on the complete available past and on the changing representation. The invariance across allowed policies is an assumption of a sufficient controlled interface, not inferred from observational conditioning alone. An independently randomized operation with the corresponding sequential causal-identification condition is one realization. The result is an immediate measured consequence, or a return collected under a fixed evaluation protocol satisfying these

conditional laws. A return under an arbitrarily changing continuation policy does not automatically satisfy them.

The record h is available to the system; it is not an unobserved world-state label. It may encode a finite sufficient history or configuration. The theorem learns how to group such records for subsequent operations. It does not generate physical distinguishability from an observation channel that has none.

Let Π_0 be the initial operative partition. A partition Π is stable if, whenever h, h' belong to one block,

$$g(h,a) = g(h',a), \qquad P_a(h,B) = P_a(h',B) \quad (a \in A,\ B \in \Pi).$$

This preserves the specified result means and transitions between represented conditions. Additional observed-result probes can be included if their response separations are required. A constant result and a one-block initial partition need not induce any further separation.

For a partition Π, let $\mathfrak{R}(\Pi)$ split each block by equality of the profile in Equation 4. Starting from Π_0, this iteration stops after at most $m - |\Pi_0|$ strict refinements. Its limit Π_* is the unique coarsest stable refinement. Indeed, any stable refinement Q of Π_0 refines every iterate: each current block is a union of Q-blocks, and equal transitions into Q-blocks give equal transitions into their unions. The limit is stable and coarser than every such Q. This is the standard stable-partition structure of stochastic bisimulation (Givan et al., 2003).

A.5 Refinement as completion of represented dependencies

The stable-partition construction can also be expressed in the function spaces used for response completion in Appendix D.3. For a partition Π of finite H, let $E_\Pi \subseteq \mathbb{R}^H$ be the space of functions constant on each block. This is a unital algebra under pointwise multiplication. Choose any strictly positive probability measure μ on H and let P_E denote orthogonal projection onto $E = E_\Pi$ in $L^2(\mu)$. Write $g_a(h) = g(h,a)$, $\mathcal{G} = \operatorname{span}\{g_a : a \in A\}$, and regard each controlled transition as the operator

$$(P_a f)(h) = \sum_{h' \in H} P_a(h,h') f(h').$$

The functions g_a describe immediate consequences, and $P_a f$ describes the expected next value of a currently expressible function f. Define the missing response space and its completed algebra by

$$\mathcal{N}_E = (I - P_E)\left(\mathcal{G} + \sum_{a \in A} P_a E\right), \qquad \mathcal{C}(E) = \operatorname{Alg}(E + \mathcal{N}_E), \tag{9}$$

where Alg is the smallest unital pointwise algebra containing the indicated functions.

Proposition 2 (Stable refinement and response completion)**.** *For the profile-refinement operator $\mathfrak{R}$ above,*

$$\mathcal{C}(E_\Pi) = E_{\mathfrak{R}(\Pi)}.$$

Moreover, $\mathcal{N}_E = \{0\}$ if and only if Π is stable. Iteration from E_{Π_0} reaches E_{Π_*}, the smallest unital algebra containing E_{Π_0} and $\mathcal{G}$ that is invariant under every P_a. The completed algebra and its limit are independent of the chosen positive measure μ.

Proof. Because P_E takes values in E,

$$E + \mathcal{N}_E = E + \mathcal{G} + \sum_a P_a E.$$

The block indicators $\mathbf{1}_B$, $B \in \Pi$, span E. Consequently, $\mathcal{C}(E)$ is generated by the old block indicators, the functions g_a, and $P_a \mathbf{1}_B(h) = P_a(h,B)$. Their common level sets are precisely the blocks of $\mathfrak{R}(\Pi)$. On a finite set, polynomial interpolation of each generator's finitely many values supplies the indicators of these common level sets. Hence their generated algebra is $E_{\mathfrak{R}(\Pi)}$. The missing space vanishes exactly when $\mathcal{G} \subseteq E$ and $P_a E \subseteq E$ for every a, which are the two stable-partition conditions. Each strict refinement increases the dimension, so the iteration stops after at most $|H| - |\Pi_0|$ strict expansions. Any unital algebra containing E_{Π_0} and $\mathcal{G}$ and invariant under all P_a contains every iterate by induction. The limit is therefore the smallest such algebra. The displayed linear-space equality also removes the dependence on μ from $\mathcal{C}(E)$. □

This identity describes how an unresolved consequence constrains the reconstruction of its predecessors. Suppose a comparison has formed a block $B \in \Pi_n$ and two records h, h' still belong to the same block of Π_n. If, for an operation a,

$$P_a(h, B) \neq P_a(h', B),$$

then the next comparison separates h and h'. They can no longer share a prediction about reaching the newly distinguished condition. Repeating this argument propagates every surviving response contrast through the profile dependencies that detect it. A nonzero path to B alone is insufficient: two paths can have identical response probabilities. The propagation concerns separation required to represent consequences, rather than a claim that every physical influence turns toward one location.

The construction retains the old functions while completing their missing dependencies within each reconstruction pass. A newly expressible function also becomes an object whose future value must be representable, which explains the repeated use of $P_a E$ in Equation 9. At the fixed point, subsequent uses of the declared operations remain expressible in the completed algebra. This is the finite controlled counterpart of the operation-preservation criterion in Appendix B; the unrepresented mechanisms still realize those operations. The statistical rule below forms this closure from actual returns and installs it for further use.

A.6 An internally executed statistical refinement rule

Memory retains counts $N(h, a)$, result sums $S(h, a)$, and next-record counts $C(h, a, x)$. An actual result increments precisely its source row. At a reconstruction epoch, use the empirical means $\widehat{g}$ and probabilities $\widehat{P}$. The scheduler starts reconstruction at unboundedly many interaction times and keeps collecting data after each reconstruction.

If any row is unvisited, retain Π_0 and the initial exploratory rule for that epoch. An unvisited row queried by another operation uses zero result mean and a fixed probability row until data arrive. Otherwise set

$$b(k) = \sqrt{\frac{2\log(k+1)}{k}}, \qquad \varepsilon_n = \max_{h,a} b(N(h, a)), \qquad \tau_n = 4m\varepsilon_n.$$

Rebuild from Π_0 at every epoch. Within each current block, connect h, h' when

$$\widehat{d}_\Pi(h, h') = \max\left\{ \max_a |\widehat{g}(h, a) - \widehat{g}(h', a)|, \max_{a \in A,\, B \in \Pi} |\widehat{P}_a(h, B) - \widehat{P}_a(h', B)| \right\} \leq \tau_n.$$

Replace the block by the connected components of this graph, and repeat until no block changes. Connected components are necessary because threshold proximity is not itself transitive. Each epoch terminates after at most $m - |\Pi_0|$ strict refinements.

Assign each block its least record index as a current identifier and retain versioned source correspondences for past operations. The block's transition row and result estimate supply its relational interpretation. An internal deployment operation installs this partition and its predictive or decision rule. Rebuilding from Π_0 permits correction of early noise-induced over-refinement. A fixed point terminates this reconstruction pass, not future sampling or future passes.

Theorem 3 (Automatic stable separation)**.** *Suppose every source row (h, a) is visited infinitely often almost surely, the controlled laws above hold, and the internal recording, estimation, reconstruction, and deployment operations continue to execute correctly. Then, almost surely, there is a finite reconstruction epoch after which every installed partition equals Π_*. If Π_* strictly refines Π_0, the system actually generates and eventually retains the corresponding new operative distinctions without being supplied the target partition.*

Proof. At successive visits to one source row, the result deviations and next-state indicator deviations are bounded martingale differences. Conditional Hoeffding bounds therefore give, for each empirical coordinate after k visits,

$$\mathbb{P}(|\widehat{z}_k - z| > b(k)) \leq 2e^{-2kb(k)^2} = \frac{2}{(k+1)^4}.$$

There are finitely many rows and coordinates. Borel–Cantelli and infinite visits imply that, eventually, all coordinate errors obey their bounds simultaneously, with $\varepsilon_n \to 0$. In particular,

$$|\widehat{g} - g| \le \varepsilon_n, \qquad \|\widehat{P}_a(h,\cdot) - P_a(h,\cdot)\|_1 \le m\varepsilon_n.$$

For every subset $B \subseteq H$, including a random block constructed from the same data, its transition error is bounded by this row L^1 error. Thus $|\widehat{d}_\Pi - d_\Pi| \le 2m\varepsilon_n$ uniformly over all partitions.

Consider the finite deterministic chain of true refinements from Π_0 to Π_*. Among its unequal within-block profile pairs, let $\Delta > 0$ be the minimum nonzero distance, if any exist. For all sufficiently late epochs, $6m\varepsilon_n < \Delta$. Equal true profiles then have empirical distance at most $2m\varepsilon_n$ and are connected. Unequal profiles have distance exceeding $4m\varepsilon_n$ and are disconnected. Hence each empirical refinement exactly reproduces the corresponding true refinement, by induction along the finite chain. It stops at Π_*. If no true refinement occurs, all within-block profiles at Π_0 are equal and the same argument gives immediate stability. The eventual error bounds hold at all later epochs, proving eventual equality. □

Corollary 4 (Reconstruction after an unannounced support change)**.** *Keep the finite record and operation sets, initial partition Π_0, bounded observations, cumulative statistics, thresholds, and reconstruction rule above. After a fixed finite number T_0 of completed interactions, suppose the actual controlled laws are g^+ and P^+ for all subsequent interactions and all allowed non-anticipating policies. The preceding records can be arbitrary within the same observation ranges. Every source row is still visited infinitely often almost surely, and the original recording, reconstruction, and deployment operations continue to execute. The change time is not supplied to the system: it retains all its counts and applies the same rule, including reconstruction from Π_0 at each epoch. Then almost surely every sufficiently late installed partition equals the coarsest stable refinement Π_*^+ for (g^+, P^+). The empirical outcome means and transition probabilities converge to these new laws.*

Proof. Fix a source row and one outcome or transition-indicator coordinate. Let z^+ be its new conditional mean, and let $K \le T_0$ be the number of its observations before the change. Write $\mathcal{F}_{T_0}$ here for the completed history through the change epoch and condition on it, so K and those observations are fixed. After $k > K$ total visits, its cumulative estimate satisfies

$$\widehat{z}_k - z^+ = \frac{1}{k}\sum_{j=1}^{K}(Y_j - z^+) + \frac{1}{k}\sum_{j=1}^{k-K} D_j,$$

where the post-change D_j are martingale differences along successive visits to this row. The controlled-law and non-anticipation assumptions give the martingale property even when visit times depend on earlier results. Each observation is in $[0,1]$, so the old contribution has absolute value at most K/k and each D_j has conditional range of length at most one. The same radius $b(k) = \sqrt{2\log(k+1)/k}$ obeys

$$\mathbb{P}\left(\left|\frac{1}{k}\sum_{j=1}^{k-K} D_j\right| > \frac{3}{4}b(k) \,\middle|\, \mathcal{F}_{T_0}\right) \le 2\exp\left(-\frac{2(3kb(k)/4)^2}{k-K}\right) \le \frac{2}{(k+1)^{9/4}}.$$

These bounds remain summable after averaging over the change-epoch history. Borel–Cantelli and the finite number of rows and coordinates give the tail bound eventually for every coordinate, almost surely. Also $K/k \le b(k)/4$ eventually, so $|\widehat{z}_k - z^+| \le b(k)$ eventually with the original radius. Infinite visits therefore restore the simultaneous coordinate bounds used in Theorem 3, with the same $\varepsilon_n \to 0$. Applying its finite refinement induction to the nonzero profile gaps of (g^+, P^+) yields eventual installation of Π_*^+. The coordinate bounds also prove convergence of the cumulative estimates. □

The corollary makes reconstruction within the return a form of repair: consequences that no longer fit an old relation change the records, comparison, and installed organization of subsequent activity. The same causal process absorbs the mismatch; it needs neither a separate repair controller nor a teacher announcing when or how the support changed. Old separations can also be withdrawn because each reconstruction starts from Π_0, although the accumulated evidence is retained. The statement assumes that the changed support has a stable controlled law at the available-record level

and that the receiving, sampling, and reconstruction mechanisms remain operative. It proves recovery of usable outcome and transition relations after a finite change, not survival of every physical component or tracking of arbitrary continuing changes. When these relations enter the implemented selector of Appendix A.8, they can change executed operations; a new prediction alone does not establish that further step.

The unknown gap Δ is not an input to the system. It affects the sample requirement. The argument applies to any implementation providing the same eventual error control and refinement/deployment semantics; literal tables are one realization. For returns in $[-M, M]$, first rescale to $[0, 1]$ or enlarge the concentration constants. No statement here identifies every hidden physical mechanism.

Infinite visits can be guaranteed by an actual exploration mechanism. For example, in a finite communicating controlled process, a fixed exploratory policy assigning every action positive probability induces an irreducible finite chain. Every row is visited infinitely often. A reset interface is another possibility when the environment actually provides one. Irreversible traps or inaccessible response separations require different conditions. Exploration coverage is not itself a guarantee of optimal online reward.

A.7 A CHANGE IN SUPPORT CAN REQUIRE REPRESENTATIONAL SEPARATION

The causal model and the system's representation are different objects. Let x be a pre-operation causal state, including history when needed, let $h = \eta(x)$ be an available record, and let $q_0(h)$ be its initial operative representation. The analyst can describe hidden variables in x without making them inputs to the learning rule.

To make the two unrepresented roles concrete, consider support B_t, a further process C_t, and an activation condition Λ_t. A local part of a time-unrolled model can have the form

$$Y_t = f(A_t, B_t, \epsilon_t), \qquad B_{t+1} = \beta(B_t, R_t, A_t, \Lambda_t, C_t, V_t).$$

Here R_t is the internal state acting through the operating loop. Suppose β is independent of C_t when $\Lambda_t = 0$, but can depend on it when $\Lambda_t = 1$. The second regime permits C_t to alter the support of a later operation. The complete mechanisms need not change. The roles concern the active dependence in an operating regime, not whether C_t is absent from the full causal graph. This local form illustrates conditional activation; the result below uses the resulting response contrast, not the particular form of β.

Corollary 5 (From support conditions to representation separation)**.** *Consider a fixed causal realization with reachable available records h_0, h_1 resulting from different support conditions. Suppose $q_0(h_0) = q_0(h_1)$ but $g(h_0, a) \neq g(h_1, a)$ for a common executable operation a. Then:*

1. *Under the prediction interface and positive evaluation weights of Proposition 1, retaining this merge incurs its strictly positive excess-risk bound. Almost-sure consistency implies eventual separation of the pair.*
2. *Under the controlled-interface and execution assumptions of Theorem 3, its internally installed partition eventually separates h_0 and h_1, almost surely.*

The second conclusion also holds if a fixed finite sequence of executable operations yields different expected measured consequences from the two records.

Proof. The first assertion substitutes the two records into Proposition 1. No stable refinement can merge records with unequal result means for the same operation. Hence Π_* separates them, and Theorem 3 proves their eventual installed separation.

For the extension, stable blocks have equal immediate result means and equal transition probabilities into every stable block under each operation. After one operation, the distributions over blocks therefore agree for two starting records in the same block. Inducting over a fixed operation sequence preserves this agreement, and equality of the blockwise means gives equal expected measured consequences at each step. A finite sequence with unequal expectations thus excludes membership in the same stable block. Apply Theorem 3 again. □

An external causal contrast and an observable response contrast must not be conflated. A hidden change may alter particular outcomes without changing any conditional mean or transition law available at H. The corollary requires the contrast specified in its statement; it does not supply a hidden-state label to the machine. Nor does consistency require recovering the entire physical mechanism: a sufficient conditional representation meets the conclusion.

Conversely, fixed mechanisms need not look stationary through an inadequate representation. Under the controlled laws, for a fixed initial grouping q_0 and a positive-probability conditioning event,

$$\bar{g}_t(r,a) = \mathbb{E}[Y_t \mid q_0(h_t) = r, a_t = a] = \sum_{h:q_0(h)=r} \mathbb{P}(h_t = h \mid q_0(h_t) = r, a_t = a)\, g(h,a).$$

Changing support conditions or operation policies can change the mixture weights even though g and P are fixed. This expresses how a familiar represented operation can appear to change its effect. The theorem distinguishes its component conditions by refining q_0.

The fixed laws at the available-record level remain an explicit assumption. If omitted history or hidden support instead makes those laws vary with time, the automatic theorem does not apply until a genuine observation or history extension supplies a sufficient controlled interface. The existence of a fixed, larger causal model alone does not prove that such an extension is available or finite.

Finally, preserving a reusable reference and revising a conditional relation are separate functional requirements. When a result cannot identify its source, the source-preservation argument in Appendix B supplies the additional condition for reuse. Together with refinement, it motivates symbolic-reference and relational-revision roles within one organization. It does not require two physical stores or identify a unique symbolic architecture.

A.8 The learned representation supports executable selection

Let Y now be a bounded per-operation utility, with $|g| \le M$, and fix $0 \le \gamma < 1$. Once the partition is correct, its quotient process preserves the optimal discounted value. To see this, the Bellman operator maps block-constant functions to block-constant functions, because both result means and transitions into blocks are equal. Its unique fixed point is therefore block-constant. An optimal quotient policy lifts to an optimal original-state policy.

Estimate the quotient rows using a canonical record in each block, or consistent pooled rows. Let e_g be the uniform result-mean error and e_P the maximum row L^1 transition error. For any fixed block policy, Bellman contraction gives

$$\|V^\pi - \widehat{V}^\pi\|_\infty \le D_n := \frac{e_g}{1-\gamma} + \frac{\gamma M e_P}{(1-\gamma)^2}.$$

If an internally computed value function is within η_n in supremum norm of the empirical optimal value and π_n is greedy for it, then

$$\|V^* - V^{\pi_n}\|_\infty \le 2D_n + \frac{2\gamma\eta_n}{1-\gamma} \longrightarrow 0$$

almost surely when $\eta_n \to 0$. The first two terms compare each policy between the two models; the last is the standard greedy-policy error bound from Bellman contraction. Value iteration with a vanishing Bellman-residual tolerance supplies such an internal planning procedure. This evaluates each deployed decision rule. A sampling policy that continues to mix exploration has an additional exploration cost.

When a executes a modification to a learning process and h includes a finite observable sufficient configuration, this is a self-modification selection result. A finite menu alone does not imply a finite sufficient configuration. Operation availability must be common, or included in the preserved signatures. The discount is per operation; variable physical execution time requires time/cost states or a semi-Markov formulation. The utility may include the costs specified by the task. Self-modifications must preserve the recording, exploration, estimation, and execution conditions of the theorem. The theorem does not compare against modifications outside the declared family.

A.9 A NEW OPERATION CAN REQUIRE A FINER REPRESENTATION

For deterministic operations T_a and a required initial readout q_0, define

$$x \sim_A x' \quad \Longleftrightarrow \quad q_0(T_w x) = q_0(T_w x') \text{ for every finite word } w \in A^*.$$

The empty word preserves q_0. Appending an operation proves that the quotient has a well-defined transition for every $a \in A$. Every other operation-closed refinement of q_0 refines this relation, by induction over words. Thus it is the coarsest such refinement, with no linearity assumption on the maps.

If $A \subseteq A'$, then $\sim_{A'} \subseteq \sim_A$. Strict separation occurs precisely when an added experiment distinguishes states equivalent under all old experiments. Naming a composition of already closed old operations does not refine this separation of histories. This identifies what must be rechecked when an acquired representation becomes a new operation.

B CAUSAL SEPARATION, RECURSIVE ADEQUACY, AND AN ORDERED RETURN

B.1 AN INITIAL SEPARATION WITHIN ONE CAUSAL PROCESS

Begin with the joint dynamic causal model, its initial law, and the relevant actual and possible interactions. Execution, resources, observation, organisms, machines, and changes in their support belong to this common process. The model is the analyst's setting, not knowledge supplied to a participant.

Causal separation is a topological division of causal relations relative to an actual causal closure. Its topological content concerns participation in the return and completion of paths across its boundary. The structural mechanisms and operating regime establish that the return actually continues. The representation-use relation specifies which of its dependencies are expressed by the organization. These specifications refer to causal roles and can select different aspects of the same physical carrier.

The initial separation is between represented relations and the unrepresented causal environment in this common process. Relative to the same closure, environmental separation singles out participating unrepresented support from other unrepresented relations. Representational separation has the corresponding boundary role: a received symbolic projection participates internally, but its determining path passes through the unrepresented side. The participating support and the symbolic projection are therefore the two sides of one cross-boundary return. Neither side, restricted to its part of that return, supplies the complete return on its own.

A change in support can make this existing separation apparent in unequal received effects. The change is an event within the separated causal structure, not the definition of separation. The topological result below makes the paired structure explicit. The implemented-receiver result then specifies when the old represented state cannot determine the received effect. Response-preserving separation, empirical reconstruction, and successor selection retain their stated conditions.

B.2 CAUSAL RETURN AND ORGANIZATIONAL CLOSURE

At a stated process and temporal resolution, let $\mathcal{O}$ be a collection of actual activity roles and their implementing mechanisms. A return has a compatible positive-time witness in the time-unrolled causal model,

$$p_{t_0} \longrightarrow q^1_{t_1} \longrightarrow \cdots \longrightarrow p_{t_k}, \qquad t_k > t_0.$$

Its edges must coexist under the conditions of one realization. Returns contributing to the continuation of the activities composing $\mathcal{O}$ give the organization its causal closure. The relevant property is stronger than the existence of one path back to a chosen interface.

Let $\mathcal{K}_{\mathcal{O}}$ be the joint states in which these activities remain enabled. This set is specified by their actual resource, mechanism, and scheduling conditions, independently of a proposed representation or its accuracy. For successive return epochs t_n, taken as stopping times of the full-history filtration, closure in the declared operating regime requires

$$\mathbb{P}\big(0 < t_{n+1} - t_n < \infty, \quad X_{t_{n+1}} \in \mathcal{K}_{\mathcal{O}} \,\big|\, \mathcal{F}_{t_n}\big) = 1 \tag{10}$$

on the admissible histories of that regime. Here $\mathcal{F}_{t_n}$ is the history of the causal realization through the current return epoch. The returned consequences must also causally enter the enabling or use of later activity. Membership in an invariant set without this actual return does not establish organizational closure. For a finite analysis, the condition is required for the continuations being analyzed, not for an unlimited lifetime.

At the resolution of recurring activity roles, let $\mathcal{G} = (V, \mathcal{E})$ be the causal dependency graph and let $O \subseteq V$ denote the roles of the identified organization. Its graph boundary is

$$\partial O = \{(i, j) \in \mathcal{E} : \mathbf{1}_O(i) \neq \mathbf{1}_O(j)\}.$$

Membership is specified by the actual organization under analysis; it is not obtained from an arbitrary decoder. A compatible return that leaves O and reaches it again crosses this boundary in both directions. The graph records these dependencies but does not replace the continuation condition. In particular, organizational closure is not closure under all causal ancestors: supporting processes can lie outside O. At a representational standpoint, the initial separation concerns which dependencies are actually expressed, regardless of where their carriers lie. The two notions use the same boundary construction but need not give identical subsets of variables.

This transfers the distinction between organizational maintenance and material openness from closure-of-constraints accounts (Montévil & Mossio, 2015) to the specified activity roles. The structural model must provide the actual continuation conditions; a graph partition cannot supply them. The results begin with such an organization and do not prove that every dynamic SCM develops one.

This is a model of sustained activity, not resource independence or repetition of an identical state. External supplies can remain necessary, and the organization's states and effective mechanisms can change. A graph's strongly connected component describes possible mutual reachability; it alone neither verifies Equation 10 nor selects a unique organization. For example, a path can return to an interface after the resources for its next use have been exhausted.

The finite implementation in Appendix A supplies one sufficient realization. Available inputs, termination of each enabled within-round operation, fair scheduling, and installation that enables the next round yield a positive finite return to an enabled state. Topological induction within a round, followed by induction over the requested rounds, proves Equation 10 while these support conditions persist. The returned data also change the statistics used in subsequent activity, supplying the required causal dependence. No statistical learning guarantee follows from this continuation statement alone.

The subjects studied here are organizations sustained through such actual causal returns. A failure of one represented dependency need not terminate every activity in $\mathcal{O}$. Reception, comparison, and modification can continue and reorganize the failed dependency through the same return. This representation-mediated restoration is the process analyzed in Section 3, not a separate controller appended to closure. Corollary 4 gives a sufficient realization: after an unannounced support change, the already executing reconstruction process recovers usable outcome and transition relations. Continuation alone does not entail those reconstruction mechanisms; the realization specifies how they participate in the organization's actual dynamics.

The organization is identified before its representational standpoint. An implemented interface p and operative readout $\pi_p : X \to \mathcal{R}_p$ then describe a representational standpoint within it. A subject can have several such interfaces; choosing another readout does not by itself produce another subject. Fixed probes and passively received consequences are permitted, since the representation need not select every external operation.

Different continuing organizations can share return mechanisms without sharing their representational standpoints. Maternal responsiveness, for example, can contribute to a child's return while the mother's activity also continues through other interactions. This is the limited intersubjective role used here (De Jaegher & Di Paolo, 2007). It does not provide the child with the mother's unobserved state. Broader organizational accounts of agency address additional individuality and normative questions (Barandiaran et al., 2009).

The initial separation introduced above remains relative to the actual representational use within this organization. Among the unrepresented dependencies, environmental separation is between those participating in the ongoing return and those outside the traced return. Relations outside the traced return can change its support, but their possibility of change does not define the separation. These

are relative causal roles rather than two physically isolated regions. With the subject and standpoint fixed, suppress p in the actual readout π_p, receiver $U_{p,a}$, and successor set $\Gamma_{p,a}$. In applying a local result, its states and transitions are restricted to the declared compatible operating regime. The local reception results below require their original assumptions; organizational continuation does not substitute for their nontrivial-response condition. Likewise, the formation and successor guarantees retain their observation, reconstruction, and execution assumptions.

B.3 Separation relative to the topology of a causal return

Use the recurring-role graph and compatible positive-time witnesses of Appendix B.2. The actual representational use assigns the relevant roles to V_R or V_N, the represented and unrepresented sides. Consider a realized return $w = (v_0, \dots, v_m = v_0)$ at this role resolution. Repeated roles occur at later times in its witness. The sequence records one execution; connections drawn from incompatible executions cannot complete it.

Environmental separation records which unrepresented relation occurrences participate in these returns and which do not. Representational separation records their internally received projections alongside internally expressed relations. The following elementary graph fact explains why a boundary-crossing return has a counterpart on each side.

Lemma 6 (Paired separation of a causal return)**.** *Let w be a directed closed walk with a compatible causal witness and with roles in both V_R and V_N. If $n_{R\to N}(w)$ and $n_{N\to R}(w)$ count its boundary crossings, then*

$$n_{R\to N}(w) = n_{N\to R}(w) \geq 1.$$

Removing the crossing occurrences decomposes the traced return into alternating same-side segments. Each segment has a boundary entry and exit in the original return; recovering that return requires the complementary-side segments and crossings.

Proof. Set $b(v) = \mathbf{1}_{V_R}(v)$. The telescoping sum $\sum_{j=0}^{m-1}(b(v_{j+1}) - b(v_j)) = b(v_m) - b(v_0) = 0$ equates the two crossing counts. Because the closed walk visits both sides, each count is positive. Between consecutive crossings, all vertices lie on one side, and every crossing changes that side. Cutting at these occurrences gives the asserted alternating segments. Rejoining consecutive segments restores precisely the deleted crossings and the original traced return. □

The segments are taken as occurrences along the return; they can themselves contain shorter local cycles. The lemma concerns completion of the chosen cross-boundary return, not absence of every other cycle in an induced subgraph. An unrepresented supporting segment is separated from the rest of its side by its participation in this return. At the represented receiving end, a symbolic projection is determined through that supporting segment. The receiver in Appendix B.9 specifies the functional resolution of this projection. This connects the topological separation to the state-based reception result without identifying the whole next state with a symbol.

The closure condition supplies an actually continuing organization, the side assignment records its representational use, and the crossing structure gives the paired separation. None is replaced by choosing a decoder after observing the data. The formation results supply the additional mechanisms through which this separation can be reconstructed and reused.

B.4 State-based realizations of representational relations

The results below use state-based realizations of the relative causal role described in Section 2. Such a realization specifies a source, an encoding or retrieval, and an actual use. For example, an available event record H supplies a state $R = Q(H)$. An implemented receiving operation uses it as $B = D(R, C)$, with its other parents C specified. Here B is a prediction, an action selection, or an update concerning the source event or condition. The relation to that source identifies what the state represents; the actual use identifies where it has that role. The encoding may include earlier internal activity, so a self-generated representative need not copy a current sensory input.

Changing R at this use, with the other parents held fixed, must be able to change the associated operation on the declared domain. An unused encoding is not an operative representation. Nor does an arbitrary decoder invented by the analyst make a physical variable representational: Q and

D describe implemented source and use relations. The source may be absent when the state is used, but delayed use is not required. These mechanisms specify the representational role in this realization; a physical carrier or an arbitrary correlation does not specify it.

Accuracy is not part of this definition. A state can organize a false prediction or an unsuccessful action and still be the representation actually used in the operation. The adequacy conditions below test which consequence laws can be expressed through it. Symbolic and relational representations remain roles within this use: the symbolic part consists of the received projections described below, and the relational part expresses dependencies available for further operations.

B.5 A CAUSAL REALIZATION AND ITS REPRESENTATION BOUNDARY

For an organization identified by the closure condition above, a discrete-time coordinate realization has representational state R_t, remaining state E_t, operation A_t, and return Y_t:

$$A_t = \pi_t(R_t, \nu_t), \quad Y_t = H_t(E_t, A_t, \xi_t), \quad R_{t+1} = U_t(R_t, A_t, Y_t), \quad E_{t+1} = G_t(E_t, A_t, Y_t, \zeta_t).$$

Its exogenous variables and their joint law are specified as part of the model. The split (R, E) partitions coordinates of this realization, relative to its actual representational use. The process represented by R can include both R and E; it need not coincide with the complementary coordinates E. Representation neither requires a disjoint physical domain nor a reduction in dimension or complexity. It does not assert a unique physical boundary; unrepresented support can be inside the hardware or body. The causal model describes the realization, rather than a complete model available to the system. An operation intervention replaces the equation for A_t. The remaining equations then determine its causal consequences. Programs and their modifications can be included as state variables, so the same description includes a change in an effective update rule.

A received history contains every record and recovery path allowed by the protocol. A limited memory or symbolic signature is a measurable function of that history. The analyst may discuss hidden variables to define an environment, but an implemented estimator receives only the declared history. Variables used to choose a prediction must be available before its outcome. These requirements prevent a latent cause, future target, or unrecorded action from entering a proof as free side information.

B.6 A NEURAL TRAINING PROCESS IN THE COMMON MODEL

A feedforward network can be represented directly by structural equations for its inputs, activations, and output (Geiger et al., 2021). To include its training, we add the mechanisms that generate an error and update the state used by the next computation. For differentiable forward computation and loss, with the sampled inputs and targets held fixed during differentiation, one time step can contain

$$\widehat{y}_t = f_{\theta_t}(x_t), \qquad d_t = \nabla_{\widehat{y}} \ell(\widehat{y}_t, y_t), \qquad g_t = J_\theta f_{\theta_t}(x_t)^\top d_t,$$
$$(\theta_{t+1}, m_{t+1}) = \mathrm{Opt}(\theta_t, m_t, g_t; \lambda_t).$$

Here m_t is optimizer state and λ_t contains specified update settings. The sample (x_t, y_t), settings, randomness, and execution conditions have their own declared parents and initial states. A nonsmooth implementation additionally specifies its derivative convention. The causal model describes these actual computations; backward differentiation reverses the order of derivative calculation, not causal time (Baydin et al., 2018).

Relative to an actual prediction, the mechanisms supplying its data or changing its parameters can be unrepresented causal support. Whether a hidden activation, a prediction, or an update setting has a representational role is determined by the actual source and use described in Appendix B.4. A fixed physical location is not a symbolic or relational type.

Under the continuation conditions of Appendix B.2, each completed step leaves the next computation enabled, and the updated state causally affects its use. The resulting training process can satisfy the paper's operational organization criterion. A stored weight array or an isolated forward call does not by itself supply that continuation. This argument concerns the specified running process and interval; it does not establish automatic recovery of a missing feature or beneficial self-modification. Those conclusions require the formation and execution mechanisms proved in the corresponding appendices.

B.7 Response separation on both sides of a boundary

Consider a deterministic realization on an admissible set $\mathcal{D}_a \subseteq \mathcal{R} \times \mathcal{E}$ for a fixed operation a:

$$r^+ = F_a(r, e), \qquad e^+ = G_a(e, r).$$

All other causal parents are fixed or included. Here r is the actual operative readout and e completes the specified causal conditions; e is not an observation supplied to the system. Fixing exogenous inputs gives such a deterministic comparison. If stochastic responses rather than fixed-input realizations are compared, replace the response values below by their conditional laws.

Set $\mathcal{E}_a(r) = \{e : (r, e) \in \mathcal{D}_a\}$ and $\mathcal{R}_a(e) = \{r : (r, e) \in \mathcal{D}_a\}$. Define

$$e \sim_{a,r} e' \iff F_a(r, e) = F_a(r, e'), \qquad r \approx_{a,e} r' \iff G_a(e, r) = G_a(e, r').$$

These equivalences describe the response separation made by the joint process at the stated response resolution.

Proposition 7 (Two-sided response factorization)**.** *For fixed (a, r), the quotient $\mathcal{E}_a(r)/\!\sim_{a,r}$ is in bijection with the image $F_a(r, \mathcal{E}_a(r))$. A code α_r on $\mathcal{E}_a(r)$ preserves this response through a map $\widehat{F}_{a,r}$,*

$$F_a(r, e) = \widehat{F}_{a,r}(\alpha_r(e)),$$

if and only if equality of codes implies $e \sim_{a,r} e'$. The corresponding statements hold for $\mathcal{R}_a(e)/\!\approx_{a,e}$, the image of $G_a(e, \cdot)$, and any code preserving that outward response. If either image is non-singleton, a constant code on that side cannot preserve its responses.

Proof. The map $[e] \mapsto F_a(r, e)$ is well-defined and injective by the definition of $\sim_{a,r}$, and surjective onto the stated image. Factorization implies that equal codes have equal responses. Conversely, if equal codes have equal responses, define $\widehat{F}_{a,r}$ on an attained code by the response of any representative. The implication makes this independent of the representative. The same argument with G_a proves the outward statement. A constant code has only one image value and hence cannot preserve a non-singleton response image. □

This is a set-level statement; for measurable factorizations use finite spaces or the quotient sigma-algebra convention below. Applied to the implemented receiver, its incoming half is Proposition 10. Neither quotient is a new observation channel or an algorithm installed in the environment.

One side can change without the other acquiring a new representational category. Both nonconstant response families are required to obtain nontrivial response separation in both directions. The result identifies their common structural form, not equal partitions, an environmental learning rule, or four temporal phases. The ordered use of one acquired separation is treated separately in Appendix B.15. When actual suborganizations admit the same mechanism and operating-regime description, the criterion applies at each of their boundaries. Their existence and continuation at arbitrarily many levels do not follow from this local result.

B.8 Adequacy of a representation

Let $\mathcal{H}$ be a set of available histories, $q : \mathcal{H} \to \mathcal{S}$ a description, and $\mathcal{I}$ a declared family of interventions or finite experiments. For $i \in \mathcal{I}$, let $\kappa_i(h)$ be the probability law of an independently specified observed consequence under that experiment. The consequence is defined by the task or evaluation protocol, rather than by a desire to recover a newly introduced label. Call *q adequate for $\mathcal{I}$* if there are kernels $\bar{\kappa}_i$ with $\kappa_i(h) = \bar{\kappa}_i(q(h))$ for every relevant h.

Proposition 8 (Causal adequacy)**.** *The description q is adequate precisely when $q(h) = q(h')$ implies $\kappa_i(h) = \kappa_i(h')$ for every $i \in \mathcal{I}$. If two histories in one description class have consequence laws at total-variation distance $d > 0$, every common predicted law has error at least $d/2$ for at least one of them.*

Proof. Factorization implies equality on every class. Conversely, define $\bar{\kappa}_i(s)$ by choosing any h with $q(h) = s$. Classwise equality makes this definition independent of the choice. For the quantitative statement, the triangle inequality gives $d \le \mathrm{TV}(\kappa_i(h), \nu) + \mathrm{TV}(\nu, \kappa_i(h'))$ for every common prediction ν. □

For finite spaces the factorization is automatically measurable. In general spaces the proposition is stated with the quotient sigma-algebra, or with a measurable factorization assumption. No extra claim about measurable selections is needed.

For a deterministic realization with state space X, let $T_a : X \to X$ be a permitted executed operation and $O_a : X \to Z$ its declared result. Write $\pi : X \to S$ for the retained representation, with $S := \pi(X)$. For a learnable implementation, it must be obtainable from available records, for example $\pi = q \circ \eta$ with $\eta : X \to \mathcal{H}$. It is not an arbitrary hidden-state input to the machine.

Proposition 9 (Operations on a reduced representation)**.** *There exist maps $\bar{T}_a : S \to S$ and $\bar{O}_a : S \to Z$ satisfying*

$$\pi \circ T_a = \bar{T}_a \circ \pi, \qquad O_a = \bar{O}_a \circ \pi$$

for every $x \in X$ if and only if, whenever $\pi(x) = \pi(x')$,

$$\pi(T_a x) = \pi(T_a x'), \qquad O_a(x) = O_a(x').$$

Proof. The factorization implies both equalities. Conversely, define $\bar{T}_a(\pi(x)) = \pi(T_a x)$ and $\bar{O}_a(\pi(x)) = O_a(x)$. The equalities make these definitions independent of the representative x. For general measurable spaces, use the same quotient or measurable-factorization convention as above. □

This is the elementary operation-preservation condition used in causal abstraction (Beckers & Halpern, 2019), here applied to an implemented operation and its usable representation. It explains both the value and the limit of leaving a mechanism implicit. When it holds, the full mechanism need not be reconstructed at each use. When it fails, one retained state cannot support the different required continuations. For a stochastic state kernel K_a, exact one-step closure requires the pushforward $K_a(x, \pi^{-1}(B))$ to be constant on each π-class for every measurable B. Together with preservation of the specified result means, this is the controlled stability condition used in Appendix A. Equality of only one-step means does not imply transition closure. Alternatively, defining $\mathcal{I}$ to include all allowed finite experiments gives a direct future-response criterion.

B.9 From an incomplete projection to actual reception

An unrepresented causal difference can affect the operative representation without its determining relation being represented. The projection below records this actual effect at the implemented readout. The symbolic component consists of the received separation carried by that effect, relative to the current operation and represented state.

Let $T_a : X \to X$ be an executed transition and let $\pi : X \to R = \pi(X)$ be the declared operative representation. Define

$$\Gamma_a(r) = \{\pi(T_a x) : \pi(x) = r\}.$$

Every $\Gamma_a(r)$ is a singleton precisely when $\pi T_a = f_a \pi$ for an internal map f_a (Proposition 9). If a fiber has different successors, any exact completion $\pi T_a(x) = F_a(\pi(x), v_a(x))$ must distinguish them through v_a. Equal pairs $(\pi(x), v_a(x))$ must give equal successors; this condition is also sufficient to define F_a on attained pairs. In particular, v_a cannot factor through π on a nonconstant fiber. This algebraic requirement is now applied to the receiving mechanism that actually runs.

Fix an operation a and a reception regime. The signal $z = E_a(x)$ reaches the receiving mechanism through the declared causal return. Its actual structural equation is

$$r^+ = U_a(r, z), \qquad \pi(T_a x) = U_a(\pi(x), E_a(x)). \tag{11}$$

All other parents of the receiver, including its memory or operating mode, are fixed or included in the declared input. Thus U_a is a mechanism of the modeled system, not a completion fitted after observing the successor. A boundary intervention substitutes the received input while retaining this mechanism and its other inputs. This is a causal comparison in the model; it does not grant the machine control of that input.

For fixed (a, r), set

$$Z_a(r) = \{E_a(x) : \pi(x) = r\}, \qquad z \sim_{a,r} z' \iff U_a(r, z) = U_a(r, z').$$

Let $\Sigma_a(r) = Z_a(r)/\sim_{a,r}$. The quotient $\Sigma_a(r)$ describes the *symbolic component* of the causal projection at this receiving operation. Its classes distinguish how unrepresented causal inputs affect the operative representation. A received signal and its actual effect realize one such class. This is a description of the separation already implemented by U_a, not a further operation that assigns representatives to effects. The complete next state can also carry internally determined structure; it is not identified wholesale with the symbolic component.

Proposition 10 (Implemented symbolic separation)**.** *Under Equation 11, the map*

$$\overline{U}_{a,r} : \Sigma_a(r) \to R, \qquad \overline{U}_{a,r}([z]) = U_a(r, z)$$

is well-defined and injective, with image $\Gamma_a(r)$. If this image is not a singleton, then for every actual x in that fiber there is an admissible x' in the same fiber whose received contribution has a different response class and changes the operative successor. No function of the old r alone can replace the received contribution and preserve all these actual responses.

Proof. Equality of receiver responses defines $\sim_{a,r}$, proving well-definedness and injectivity. The implemented equation gives

$$\overline{U}_{a,r}(\Sigma_a(r)) = \{U_a(r, E_a(x)) : \pi(x) = r\} = \Gamma_a(r).$$

For any x in a fiber with more than one successor, choose x' with a different successor. Then $E_a(x)$ and $E_a(x')$ have different response classes. Substituting the latter input into the same receiver changes its output by the structural equation. A function of r is constant on the fiber and cannot make this distinction. □

At fixed (a, r), the actual successor r^+ codes its response class without hidden-state access or enumeration of all classes. For measurable receivers this concrete encoding also avoids a measurable-selection requirement for abstract quotient labels. A finite implementation can use a common observed outcome code $s = \kappa(z)$, such as OPEN or CLOSED, provided equal codes give the same selected receiver response in each compared context. Such a code may retain more information than the minimal response class. Other parts of the received trace may also enter later relational comparisons.

The theorem gives representational separation in the current implemented response. It does not infer learning from recurrence alone. For a stochastic transition kernel K_a, an exact reduced law instead requires $K_a(x, \pi^{-1}(B))$ to agree on each π-fiber. Different samples from one correctly represented law do not violate that requirement. An individual realized reception can be analyzed by including its exogenous realization in x. The formation results below learn conditional laws or their declared means, rather than removing irreducible sample variability.

B.10 One event, distinct source and return readouts

Let c_t be the adopted source reference and a_t its executed operation. Let z_t be the actual return and σ_t its symbolic response code. An available event record can be written

$$m_t = (h_t, c_t, a_t, z_t, \sigma_t, h_{t+1}), \qquad \mathrm{src}(m_t) = c_t, \quad \mathrm{ret}(m_t) = \sigma_t.$$

Only actually available fields are included; a sufficient statistic may replace the literal tuple. Two events can have the same c_t and different σ_t. The source, return, and learned condition label are therefore different readouts, not one scalar being renamed.

Event correspondence can be preserved by row-wise updates, without storing a new event identifier. For a finite, consistently coded return alphabet Σ, let

$$N_n(h, a, s) = \sum_{t \le n} \mathbf{1}\{h_t = h,\ a_t = a,\ \sigma_t = s\}, \qquad \widehat{p}_n(s \mid h, a) = \frac{N_n(h, a, s)}{\sum_{s'} N_n(h, a, s')}.$$

Use any fixed convention for an unvisited row. These counts associate the same received symbols with the conditions and operations under which they occurred. Their support defines a weighted association relation between Σ and $H \times A$. Following it from a symbol retrieves its contexts; following it in reverse retrieves the returns associated with a context. Complete event records are not used as comparison keys that would separate every observation in advance.

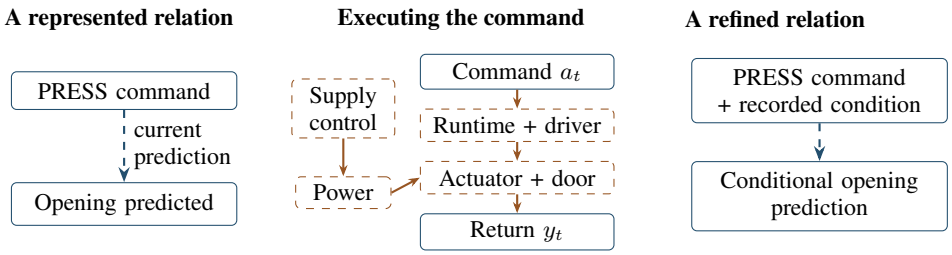


Figure 3: The represented prediction leaves steps from command to outcome implicit. The middle panel unfolds the command's causal path and support conditions; it is the analyst's model, not knowledge supplied to the robot. Accessible readings and returns can refine the prediction without identifying every mechanism. Dashed arrows show predictions; solid arrows show causal paths.

Proposition 11 (Reconstruction of the same symbolic response)**.** *Suppose the finite controlled-interface and continued-execution assumptions of Theorem 3 hold, with stable conditional probabilities $p_s(h,a)$ for each received code $s \in \Sigma$. Precisely, for every allowed policy, $\mathbb{P}(\sigma_t = s \mid \mathcal{F}_t) = p_s(h_t, a_t)$, where $\mathcal{F}_t$ includes the chosen source row and action but precedes this return. Compare the symbol coordinates by $\max_{a,s} |p_s(h,a) - p_s(\widetilde{h},a)|$ and include the transition coordinates as before. The error radius uses the source-row count $N(h,a) = \sum_s N(h,a,s)$; individual symbols need not each occur infinitely often. Then the installed partition eventually equals their coarsest stable refinement almost surely. The empirical relations consistently estimate the conditional probabilities of these same received symbols. The result does not require the condition grouping $q_0(h)$ to encode the output of an earlier event.*

Proof. For each s, the observed indicator $\mathbf{1}\{\sigma_t = s\}$ is a bounded result with conditional mean $p_s(h_t, a_t)$. Apply the coordinate error bound in Theorem 3 to this finite set of indicators and to the transition indicators. A finite union preserves summability, so all bounds hold eventually almost surely. The maximum-norm profile error remains at most $2m\varepsilon_n$ for $m = |H|$. The same finite refinement induction gives the coarsest stable partition. Infinite visits yield convergence of every response frequency. If a reference c carries information needed beyond (h,a), include its observed sufficient part in the source row before applying the result. □

For binary OPEN/CLOSED returns, $p_{\mathrm{OPEN}} = g$ and $p_{\mathrm{CLOSED}} = 1 - g$, so the original scalar-result theorem already supplies this reconstruction. The new block label $b = q_n(h)$ denotes a learned condition. It is neither the adopted c_t nor the received σ_t. Reintegration places the same received code in an estimated conditional response relation. A prediction of that response and its actual occurrence remain distinct events.

B.11 A FINITE DOOR PROCESS CARRYING THE WHOLE RETURN

Each observation window begins with the door closed and ends with a declared spring or reset mechanism returning it to that state. Let the unrepresented supply be $p_t \in \{0,1\}$ and maintenance be $M_t \overset{\text{iid}}{\sim} \text{Bernoulli}(1/4)$:

$$p_{t+1} = p_t \oplus M_t, \qquad h_t = p_t, \qquad Y_t = \mathbf{1}\{a_t = \text{PRESS}\}p_t.$$

Maintenance switches the door motor's supply; the robot's sensor and receiving controller remain operative. The procedure receives a measured reading h_t, its adopted operation, the observed door result Y_t, and the next reading. It is not supplied with the equations connecting the reading, supply, and result. Initially $q_0(0) = q_0(1)$, and the familiar PRESS reference predicts an opening.

At reception, the actual control mechanism sets a passage flag to Y_t. Before the return, its selected operative input consists of the old group, the adopted command, and an unset result field. The two supply states give the same selected input under PRESS but different received results. The receiver sets the field and passage flag accordingly. Thus Proposition 10 applies to this declared reception interval even though the raw sensor reading is separately available for later comparison. CLOSED already changes the current action.

The controlled result means and transitions are

$$g(0, \text{PRESS}) = 0, \quad g(1, \text{PRESS}) = 1, \quad g(h, \text{WAIT}) = 0, \qquad P_a = \begin{pmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{pmatrix}.$$

A fixed positive-probability exploration of both operations gives infinite visits to all four source rows. The stable refinement is $\Pi_* = \{\{0\}, \{1\}\}$. The source counts learn the conditional probabilities of the original OPEN/CLOSED symbols, and the installed predictor uses them directly. No new hidden-mechanism label is needed.

For a uniform evaluation over the four (h, a) rows, every predictor retaining the one-block condition grouping has minimum squared excess risk $1/8$, or $1/16$ for half squared loss. Indeed, its best PRESS prediction is $1/2$ and its WAIT prediction is zero. The refined relation has zero prediction error in this noiseless example. These are analytical values for the stated causal process, not an additional training experiment. For a fixed pre-existing command cost $0 < \kappa < 1$, let the task criterion be $Y - \kappa \mathbf{1}\{a = \text{PRESS}\}$. Once the conditional means are learned, the deployed greedy rule presses at reading 1 and waits at reading 0. A fixed positive exploration probability preserves access to both operations; its cost is separate from the greedy rule's value.

B.12 Reception in the neural response-completion realization

In the paired-return protocol, a probe setting A is fixed before the returns. The first coordinate $Z = Y_1$ arrives before the prediction of the second coordinate Y_2. Thus $H = (A, Z)$ is an available prefix at that prediction, not information available before the entire interaction. The source reference is A; the second received symbol is the observed sign Y_2.

The numerical program forms conditional means from paired array rows:

$$\overline{y}_H = \frac{1}{n_H} \sum_{i : H_i = H} Y_{2i}.$$

For a nonempty row, keep its membership and the other returns fixed. Replacing one received Y_{2j} changes the actually stored mean by

$$\Delta \overline{y}_H = (Y'_{2j} - Y_{2j})/n_H.$$

A nonzero replacement therefore has a nonzero receiving effect. This is the batch receiving operation implemented by the program, of the form in Equation 11.

The response-completion theorem then estimates $g(H) = \mathbb{E}[Y_2 \mid H]$ from those same rows and separates its component outside the old function space. The full prefix H is retained: the missing interaction in $\text{span}\{1, A, Z\}$ is a function-space restriction, not a merger of two observable input records. Installing the new component and fitting its coefficient makes the original second return predictable. A subsequent selector uses that relation to choose a preparation, which is executed through the original gradient flow. The future task is revealed only after that execution. These timing and readout correspondences connect reception, formation, and deployment without identifying their variables.

B.13 Reference preservation during revision

The following statement isolates the role requirement from a physical implementation. Let z be a complete usable internal state, $c(z)$ an adopted reference, and $b(z)$ a response required by an independently specified future consequence. Suppose that $c(z_0) = c(z_1)$ but $b(z_0) \neq b(z_1)$.

Proposition 12 (Reuse and revision)**.** *A system whose complete revision rule factors through c cannot realize b on both states. A system that both reuses the same reference and realizes b must have a reference-preserving distinction available to the revision rule. This need not be a product state space or two physical modules.*

Proof. If $b = \bar{b} \circ c$, equal references imply equal responses, contrary to the assumption. Hence some accessible distinction outside the reference value is necessary. Conversely, a state z with separate readouts $c(z)$ and $b(z)$ can support the two functions even if both readouts share an implementation. □

The proposition proves a difference in operational resolution. It does not prove that every successful system contains an explicit causal graph. In this paper, relational representations are the dependencies available for formation and revision; claims about their causal accuracy require the experiment conditions below.

Now let C identify the operation actually adopted and Y its return. If the required response $b(C, Y)$ differs for two sources compatible with the same received Y, a state retaining only Y cannot realize that response. All accessible recovery paths must be counted. This proves the source-preservation claim in the main text and specializes the information results in Appendix C. A new conceptual partition may change with learning without changing which earlier operation occurred. An implementation can therefore re-encode all its records under a new partition, provided that the original sources or sufficient equivalents remain recoverable.

B.14 Recursion and the propagation of separation

Suppose $q_1 : X_0 \to X_1$ and $q_2 : X_1 \to X_2$ are compatible descriptions. If $q_1 T_a = T_a^{(1)} q_1$ and $q_2 T_a^{(1)} = T_a^{(2)} q_2$, then substitution gives

$$(q_2 q_1) T_a = T_a^{(2)} (q_2 q_1).$$

Induction proves the same result for any specified finite number of levels. This is the elementary composition structure used in causal abstraction (Beckers & Halpern, 2019).

Failure also has an exact propagation condition. If two states merged at one level produce different required outcomes, that level is inadequate. A further projection inherits the difference only if its required outcome distinguishes those continuations. A constant readout, for example, can suppress a response separation. Thus neither causal connection nor nesting alone proves that every difference must propagate. The theory applies the same adequacy criterion at each level and identifies which task-relevant differences survive. It does not infer a physically infinite hierarchy from a finite model.

An exact abstraction of every primitive operation remains exact for every finite composition of those operations. For deterministic maps this follows by repeated substitution. For controlled Markov kernels satisfying lumpability it follows by induction over the history, including policies that use only the abstract history. Consequently, merely naming a sequence of already closed operations cannot create new environmental information. Representation formation can instead change computational access, retain previously discarded evidence, or enable an interaction outside the earlier effective protocol. These cases must be specified separately.

B.15 The minimal ordered return of one event

Fix an interaction event and its allowed internal history. A prior return can first expose an unexplained difference. Let C identify a candidate condition formed from that prior record, and let A denote the actual diagnostic operation. The candidate can guide the choice of operation or organize comparison under a fixed probe protocol. The pair (C, A) is the adopted source for this return. Let Y denote its new actual consequence; let V be a relation estimate whose empirical content depends on Y; and let B^+ be an executed rule that uses V. The causal dependencies include

$$A \longrightarrow Y, \qquad (C, A, Y, \text{available records}) \longrightarrow V \longrightarrow B^+.$$

The first arrow records execution and its consequence; the second records the inputs used to form and deploy the relation. No causal edge from C to A is required when the probe protocol is fixed. The four functional positions are adoption of (C, A), reception of Y, comparison forming V, and reconstruction deploying B^+. The first failed return that motivated C belongs to the preceding cycle; it is not placed after C in this chain. Comparison includes evaluation against existing records or candidate explanations; it need not be a separate batch operation.

Proposition 13 (Order of an informative return)**.** *Under a non-anticipating execution protocol, adoption precedes the return attributed to it. A relation estimate cannot use empirical information from that return before it becomes available. An improvement attributed to the estimate requires a subsequent executed rule to depend on it. If the source is needed to interpret the return and cannot be recovered from it, erasing the source blocks the corresponding inference.*

Proof. Let $\mathcal{F}_-$ be the complete pre-return information. The adopted source is measurable before its return is received. Every estimate executed before reception is measurable with respect to $\mathcal{F}_-$ and independent randomization conditionally uninformative about the unknown mechanism. If two mechanisms give the same pre-return law, such an estimate has the same law in both. It cannot exploit a distinction introduced only by the return. After reception, a statistic of $(\mathcal{F}_-, Y)$ can exploit that distinction. If the deployed rule and all of its inputs have the same law whether or not this statistic is formed, the deployed behaviour also has the same law. Thus a behavioural gain attributed to the statistic requires a causal use of it. The source-erasure claim follows from Proposition 12, or from the reception factorization in Appendix C. □

The proposition gives a partial order of dependency, not a lower bound of four physical time steps. Reception, comparison, and installation can be fused in one implementation. The account is minimal in following one informative return without inserting a further adopted operation between its source and the installation of its consequence. It also does not prohibit learning directly from a sufficiently informative passive observation. What it excludes is obtaining empirical support from a result before receiving it, or attributing behavioural improvement to a relation that is never used.

B.16 Forming a conditional representation without hidden labels

Let pre-action records range over a finite set $\mathcal{H}$ and probes over a finite set $\mathcal{A}$. Formation episodes are independent and identically distributed. They draw $H \sim \mu$ with $\mu(h) > 0$ for every h, assign A using a fixed distribution $\rho(a \mid h) > 0$ for every (h, a), and observe $Y \in [-M, M]$ from a fixed response law. Assume consistency and sequential randomization: conditional on H, probe assignment is independent of the potential returns. Then the observable conditional means identify $m_a(h) = \mathbb{E}[Y \mid H = h, do(A = a)]$. Stationarity is required over these episodes. Without this causal assumption the same argument remains a predictive, rather than interventional, result.

For an initial description q_0, define

$$h \sim_* h' \quad \Longleftrightarrow \quad q_0(h) = q_0(h') \text{ and } \quad m_a(h) = m_a(h') \text{ for every } a.$$

Let q_* denote its class index. It is the coarsest refinement of q_0 that makes all declared response means functions of the new description. For a finite-valued return, using its indicator vector in place of Y gives the corresponding statement for full consequence laws.

Construct q_N from the first N episodes as follows. At each (h, a), estimate $m_a(h)$ by its sample average. For pairs with the same q_0, join h and h' when $\max_a |\widehat{m}_{a,N}(h) - \widehat{m}_{a,N}(h')| \le 3N^{-1/4}$. Use the connected components as the new classes. Until a pair has observations, use any fixed convention. An arbitrary fixed ordering of raw records can assign names to components.

Theorem 14 (Consistent conditional separation)**.** *Under these finite-support assumptions, q_N eventually equals the partition q_* almost surely. No hidden-mechanism labels are needed. The class names can be revised during formation, while the original operation records are retained.*

Proof. There is a positive minimum probability over the finitely many (h, a). Each cell therefore receives a linear number of observations eventually almost surely. Hoeffding's inequality and a union bound imply that

$$\max_{h,a} |\widehat{m}_{a,N}(h) - m_a(h)| \le N^{-1/4}$$

eventually almost surely: the probability bounds, including the event that a cell count is too small, are summable in N. Apply the Borel–Cantelli lemma. For histories in the same true class, the estimated response vectors differ by at most $2N^{-1/4}$, so an edge is present. Among distinct true classes within an initial class, the nonzero response-vector distances have a positive minimum, since there are finitely many pairs. For large enough N, none of these pairs has an edge. The connected components are therefore exactly the true classes. Coarseness follows because any refinement of q_0 supporting every mean must separate any pair with a different response vector. □

For a finite return alphabet, apply the proof to its finitely many indicators. This is an explicit state-refinement construction in the tradition of predictive state splitting (Shalizi & Shalizi, 2004); it is not a claim to invent state splitting. Its role here is to supply a data-derived formation map in the causal return.

Role	Finite door process	Neural response completion
Adopted reference	PRESS or WAIT	Probe setting A
Received symbol	OPEN or CLOSED	Second return Y_2
Observed conditions	Reading h, operation a	$H = (A, Z)$ after first return Z
Initial restriction	q_0 merges the readings	$E = \mathrm{span}\{1, A, Z\}$ lacks the interaction
Reconstructed relation	$p(\sigma \mid h, a)$, $q_*(h)$	$g(H)$ and its new component ψ
Actual deployment	Install predictor and condition groups	Train the new feature; execute selected preparation

Table 1: The same event roles have different concrete realizations. Source references, returned symbols, and learned condition labels are distinct. Z is observed after assigning A, so $H = (A, Z)$ is available before predicting Y_2, not before the whole encounter. The door is an analytical instance; Table 2 tests the neural realization.

For a fixed evaluation distribution supported on the sampled pairs (H, A) and with the same response means,

$$m_0(q_0(H), A) = \mathbb{E}[m_A(H) \mid q_0(H), A].$$

Conditional expectation is the best predictor for half squared loss. Orthogonality therefore gives the half squared-risk difference

$$\frac{1}{2}\,\mathbb{E}\Big[\big(m_A(H) - \mathbb{E}[m_A(H) \mid q_0(H), A]\big)^2\Big]. \tag{12}$$

Because $m_A(H)$ factors through $(q_*(H), A)$, the new partition attains the full-record conditional-mean risk. For an implemented predictor, pool the empirical returns within each current class of q_N separately for each operation a; output zero until that cell has a sample. Once q_N has stabilized to q_*, the pooled means converge to the common $m_a(h)$ in each class. These predictions lie in $[-M, M]$. Finite-domain convergence and boundedness therefore give convergence of expected prediction risk by dominated convergence. This is a gain for the declared consequences, not identification of every latent cause or sufficiency for arbitrary multi-step control.

The available pre-action record may already distinguish situations that the active description merges. The construction reorganizes that available information. If a needed difference is absent from the entire permitted record, an additional informative interaction is required; internal regrouping cannot manufacture it. Finally, forming q_N does not itself prove that a neural update uses it. Appendix D.3 gives the separate installation and finite-sample gain result.

B.17 Concrete roles in the two learning realizations

Table 1 relates the finite controlled example to the neural response-completion example. They share event roles but impose different representational restrictions: one merges observed conditions, and the other lacks a function over retained conditions.

C Separation: full statements and proofs

C.1 Reception and information

Let $\mathcal{Y}$ be the space of legal returns, $e : \mathcal{Y} \to \mathcal{S}$ the current expression, and $\tau : \mathcal{Y} \to \mathcal{T}$ the required response. The following statement gives the reception requirement used in Section 2.

Proposition 15 (Reception requirement)**.** *Fix the pre-reception state and let $\tau(y)$ be a required subsequent response. If τ does not factor through e, no reception map of the form $\Phi = \varphi \circ e$ followed by internal processing can realize τ. For random processing with no additional informative input, equal received states also imply equal output laws.*

Fix a pre-reception state, a space of legal returns $\mathcal{Y}$, and a required response $\tau : \mathcal{Y} \to \mathcal{T}$. If reception has the form $\Phi = \varphi \circ e$ and the subsequent computation is B, success requires $\tau = B \circ \varphi \circ e$. This contradicts the hypothesis that τ does not factor through e, proving Proposition 15. For random

computation, replace B by a conditional output kernel. Equal received inputs have equal output laws whenever the remaining random sources contain no additional information about the return. A protocol with an accessible raw record or informative external query must include that path in Φ; it is not subject to a restriction it can bypass.

Let H be the complete usable record, $R = Q(H)$ a restricted record, and $g(H) = \mathbb{E}[Y \mid H]$ for square-integrable Y. Let $\mathcal{R}_H^*$ and $\mathcal{R}_R^*$ be the best half squared prediction risks with those records. Then

$$\mathcal{R}_R^* - \mathcal{R}_H^* = \tfrac{1}{2}\mathbb{E}\|g(H) - \mathbb{E}[g(H) \mid R]\|^2. \tag{13}$$

To prove this identity, conditional expectation is the orthogonal projection in L^2 onto functions measurable with respect to the available record. For every H-measurable predictor f,

$$\tfrac{1}{2}\mathbb{E}\|Y - f\|^2 = \tfrac{1}{2}\mathbb{E}\|Y - g\|^2 + \tfrac{1}{2}\mathbb{E}\|g - f\|^2.$$

The cross term vanishes because $\mathbb{E}[Y - g \mid H] = 0$. The best R-measurable predictor is $\mathbb{E}[g \mid R] = \mathbb{E}[Y \mid R]$. Subtracting the two optimal risks proves the identity. It is an information restriction on the complete permitted record, rather than a lower bound for an arbitrarily selected decoder.

C.2 Why a self-generated reference can matter

Let θ be an unknown environment, B existing information, S a self-generated condition, and Y its return. Suppose S is generated without new knowledge of θ, so $I(\theta; S \mid B) = 0$. The chain rule, when the quantities are finite, gives

$$I(\theta; S, Y \mid B) = I(\theta; Y \mid B, S) = I(\theta; Y \mid B) + I(\theta; S \mid B, Y).$$

Thus S need not initially identify the environment to become relevant after the return. Preserving the pair can provide information absent from the anonymous return. The relevant preserved object can be smaller than S itself: any statistic sufficient for the required posterior or decision is enough.

Suppose internal updates continue during this inference. If c_t recovers the required adopted class, retention means

$$c_{t+1}(U_t z) = c_t(z). \tag{14}$$

When the decoder changes, it belongs to the complete state or is explicitly indexed by time. If interpretation can take two different values under the same recovered adopted class, reference and current interpretation cannot be the same function of the state. This is the functional separation used in the main text. It does not assert a global product decomposition of all internal states.

C.3 An interaction-wide modification lower bound

Let two environments have equal initial observable laws. Assume their next-observation kernels agree conditional on every reachable observable history and every action allowed by the preparation protocol. An induction over interactions shows that the complete history laws agree under any adaptive policy. The policy's internal randomization can be incorporated into the history as independent randomness. Internal self-modifications are functions or kernels of this history, so they cannot break the equality.

Consequently, the final modification has the same law ν in both environments. Let $\ell_\theta(p)$ be the expected loss of an admissible successor modification under the specified future task and adaptation protocol. Define

$$\gamma = \inf_p \tfrac{1}{2} \sum_{i=0}^{1} \left(\ell_{\theta_i}(p) - \inf_{p'} \ell_{\theta_i}(p') \right) > 0. \tag{15}$$

Its equal-prior excess successor loss is

$$\int \tfrac{1}{2} \sum_{i=0}^{1} \left(\ell_{\theta_i}(p) - \inf_{p'} \ell_{\theta_i}(p') \right) d\nu(p) \geq \gamma.$$

The condition $\gamma > 0$ is essential. Different named optima do not suffice if there is a common near-optimal sequence. The admissible class also includes any future adaptation explicitly permitted inside a candidate. If one common candidate can identify the environment and achieve both optima within the evaluation period, the claimed positive gap does not apply.

C.4 Reference-preserving plasticity

For an adopted reference, let c_t recover its identity from the usable state z, and let U_t be an actual update. Preservation means

$$c_{t+1}(U_t z) = \mathrm{id}_S(c_t(z)) = c_t(z).$$

The identity map acts on the space S of retained references. It does not assert $U_t z = z$, unchanged interpretation, or a physical state causing itself. The time-indexed recovery maps allow re-encoding, provided the same adopted reference remains recoverable. New references can be added without asserting that their identities existed before formation. An interpretation readout ℓ_t may simultaneously satisfy $\ell_{t+1}(U_t z) \neq \ell_t(z)$. Thus continued reference and revised knowledge are compatible. This is the reuse requirement of Proposition 12, not a consequence of writing the tautology $s = s$.

Let the complete representational state belong to a smooth manifold M. Let $c : M \to \mathbb{R}^r$ recover a continuous reference and $\ell : M \to \mathbb{R}^k$ describe an actual response. Assume c has constant rank locally. At z, write $C = Dc(z)$ and $J = D\ell(z)$. A first-order response change b can preserve the reference if and only if

$$b \in J(\ker C).$$

Indeed, the required velocity must solve $Cv = 0$ and $Jv = b$. Choosing a basis for $\ker C$ and a complementary basis gives

$$\operatorname{rank}(J|_{\ker C}) = \operatorname{rank}\begin{bmatrix} C \\ J \end{bmatrix} - \operatorname{rank} C.$$

This measures response freedom under the reference constraint. A discrete reference can be locally constant, in which case its derivative has rank zero. One cannot insert the number of discrete labels into this differential formula.

When C has full row rank and the actual plasticity metric is $G \succ 0$, define

$$P = G^{-1} - G^{-1}C^\top (CG^{-1}C^\top)^{-1} CG^{-1}.$$

Then $P \succeq 0$, $CP = 0$, and $\operatorname{range} P = \ker C$. For an acquired response target y, the flow

$$\dot{z} = -PJ^\top(\ell - y)$$

preserves c and satisfies

$$\dot{\ell} = -K(\ell - y), \qquad K = JPJ^\top, \qquad \frac{d}{dt}\tfrac{1}{2}\|\ell - y\|^2 = -(\ell - y)^\top K(\ell - y).$$

Changing G within positive definite metrics leaves $\operatorname{range} K = J(\ker C)$ unchanged. It can change rates but cannot create a missing instantaneous direction. Global reachability requires examining the reachable part of the reference fiber, not extrapolating this local statement.

C.5 Smooth representations and projected dynamics

The same condition has a geometric form when the state and representation spaces admit smooth structure. Let $\pi : M \to N$ be a smooth surjective submersion between manifolds, and let V_a be a smooth vector field for a permitted operation. The fiber $\pi^{-1}(r)$ consists of implementation states that share a representation. No assumption here identifies an arbitrary discrete partition with a smooth manifold.

Proposition 16 (Projected operation flows)**.** *A smooth field $\bar{V}_a$ on N satisfying*

$$D\pi_x V_a(x) = \bar{V}_a(\pi(x))$$

exists if and only if $D\pi_x V_a(x)$ is constant on every fiber, with both vectors compared in the common tangent space $T_{\pi(x)}N$. For the corresponding local flows,

$$\pi \circ \Phi_a^t = \bar{\Phi}_a^t \circ \pi$$

whenever both sides are defined. Conversely, this flow identity implies the field identity.

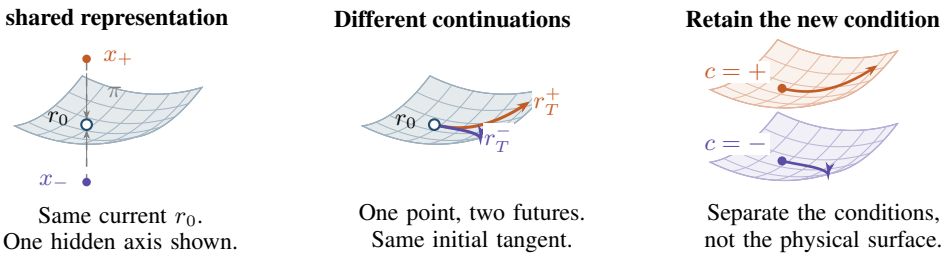


Figure 4: Different realizations can share one point on the representation surface but have different deterministic continuations. Gray arrows apply the representation map π; colored curves show time evolution. The curves share an initial tangent before separating. The right panel retains a condition distinguished from recorded probe returns (Appendix C.6). Curvature and sheet spacing are display choices.

Proof. Necessity follows from the field identity. For sufficiency, define $\bar{V}_a(r) = D\pi_x V_a(x)$ for any x with $\pi(x) = r$. Fiberwise constancy makes it well defined. Local sections of the submersion show that it is smooth. The chain rule makes $\pi(\Phi_a^t(x))$ a solution of the projected differential equation with initial value $\pi(x)$. Uniqueness gives the flow identity. Differentiating that identity at $t = 0$ gives the converse. □

This is the standard relation between related vector fields and their flows (Goldman, 2021); its use here is to express the same abstraction test as Proposition 9. In local coordinates $\pi(u, v) = u$, write

$$\dot{u} = f_a(u, v), \qquad \dot{v} = g_a(u, v).$$

A closed operation on u requires its velocity to be independent of the omitted coordinate v. Globally, the comparison must cover the entire fiber, not only one coordinate neighborhood. A change along an unrepresented direction can therefore leave the current representation fixed but change how it will move next.

Learning can also change the representation map itself. For $\pi_\theta(x)$, the correct derivative is

$$\frac{d}{dt}\pi_\theta(x) = D_x\pi_\theta\,\dot{x} + D_\theta\pi_\theta\,\dot{\theta}.$$

Equivalently, include θ in the full state and apply the proposition to the joint dynamics. The extra term cannot be discarded merely because the representation is called a projection.

No metric or curvature is required by this argument. A stochastic process requires its transition-law analogue, not only a condition on the drift. Unresolved variables can also appear through history-dependent and forcing terms in exact projected descriptions, as in the Mori–Zwanzig representation (Chorin et al., 2000). Those terms are not automatically known to the learning system. The geometric statement describes when a reduction is sufficient; the finite statistical theorem separately supplies a process that can learn a sufficient refinement.

C.6 A SURFACE VIEW OF HIDDEN RESPONSE DIFFERENCES

A causal graph and a state surface encode different objects. Graph nodes denote variables and graph edges denote parent dependencies in structural mechanisms (Bongers et al., 2021). A state-space point assigns values to the variables. A vector field or executed map, together with its causal and intervention semantics, specifies their change. A curved surface alone therefore does not specify a causal structure. One direct geometric representation of a smooth structural mechanism $Y = f(X, E)$ is its solution graph

$$\mathcal{G}_f = \{(x, e, y) : y = f(x, e)\}.$$

It is a smooth manifold because $(x, e) \mapsto (x, e, f(x, e))$ is a smooth parametrization with the coordinate projection as inverse on its image. The assignment to Y and the allowed interventions still carry the causal orientation. An unoriented solution set alone does not specify which equation changes under intervention. The figure below instead displays the retained-state space and its continuations, which makes failure of a representation easier to see.

Let $\pi : M \to N$ be a smooth surjective submersion from implementation states to retained representations. For the two-dimensional illustration, the drawing uses an embedding $\iota : N \hookrightarrow \mathbb{R}^3$ to display N as a surface. The height, normal direction, and curvature of that display have no intrinsic causal status. In particular, all displayed projected velocities $D\iota\, D\pi\, V_a$ are tangent to the surface. Predictive sufficiency asks whether the projected continuation is single-valued on each fiber, not whether the arrow is tangent.

For a declared family of experiments $\mathcal{I}$, let $\rho_i(x)$ be the required response under experiment i. Within one fiber, define

$$x \sim_{\mathcal{I}} x' \quad \Longleftrightarrow \quad \rho_i(x) = \rho_i(x') \text{ for every } i \in \mathcal{I}.$$

In a stochastic setting, specify whether ρ_i is a response law or a required statistic such as its mean. This classifies unrepresented variations by their effects from the representation's viewpoint. One fiber may contain several such response classes. Any refinement adequate for the declared responses must distinguish states in different classes. This is the factorization test of Proposition 8; it is not a claim that the physical world has torn into components.

A finite-time test can detect effects hidden from instantaneous motion. Let Φ_a^T be a smooth operation flow at a fixed time T, and define

$$\mathcal{K}^a_{T,x} = D(\pi \circ \Phi_a^T)_x\big|_{\ker D\pi_x}.$$

Its domain consists of changes that leave the initial representation unchanged. If $\mathcal{K}^a_{T,x} v \neq 0$, the projected finite-time response changes to first order along that hidden direction. Indeed, the submersion gives a smooth fiber curve $\gamma(\epsilon)$ with $\gamma(0) = x$ and $\gamma'(0) = v$, and

$$\pi\Phi_a^T(\gamma(\epsilon)) = \pi\Phi_a^T(x) + \epsilon\mathcal{K}^a_{T,x} v + o(\epsilon)$$

in a target coordinate chart. Thus arbitrarily close states in the same initial fiber have different continuations, precluding an exact prediction from the initial representation alone. If the task only requires a readout O, use $D(O \circ \pi \circ \Phi_a^T)_x v$ instead. A zero first-order response at one point does not establish exact equivalence or absence of a later effect. No orthogonal decomposition of hidden causes is assumed.

Analytic construction of Figure 4. Consider the causal probe dynamics on $M = \mathbb{R}^4$,

$$\dot{u} = 1, \qquad \dot{v} = u^2 w, \qquad \dot{z} = -z, \qquad \dot{w} = 0, \qquad \pi(u, v, z, w) = (u, v).$$

The coordinate u records the phase of the probe, v its measured response, and z, w omitted coordinates. These equations describe the illustrated probe, not a complete autonomous learner. Their solution is

$$u_T = u_0 + T, \qquad v_T = v_0 + \frac{w}{3}\big[(u_0 + T)^3 - u_0^3\big].$$

Changes in z do not affect these responses. At $u_0 = 0$, changes in w also leave the instantaneous projected velocity unchanged. Nevertheless, $w = \pm 1$ gives $v_T = v_0 \pm T^3/3$ for every $T > 0$. The displayed paths therefore start at the same point with the same tangent and then separate.

The display embedding is $\iota(u, v) = (u, v, 0.18(u^2 + v^2))$. The trajectories are plotted from the analytic solution with $T = 1.3$. In the left panel one omitted coordinate is displayed as height relative to the surface; the coordinate z is suppressed in the drawing. The right panel shows two slices of the sufficient refinement $\pi^+(u, v, z, w) = (u, v, w)$. Its dynamics are $(1, u^2 w, 0)$, and z can still be omitted. The separated sheets are a coordinate display of a new retained condition, not a physical crack.

This refinement must be obtained from accessible experience. For the noiseless probe with $u_0 = 0$, retaining the initial response, final response, and elapsed time permits $\widehat{w} = 3(v_T - v_0)/T^3$. The drawing does not grant w as a hidden-state label. With noisy data, reliable discrimination requires its own observation and estimation conditions. To connect to the finite automatic theorem, one may restrict to two reproducibly distinguishable contexts, fixed-duration experiments, and a genuine reset-and-record protocol satisfying its assumptions. The smooth illustration alone does not prove that all continuous systems learn automatically.

More generally, a geometric difference between zero-probability points is not itself a statistical risk bound. For half squared prediction loss, let H include all available source and operation information and let $g(H)$ be the response mean. The existing risk identity gives

$$\mathcal{R}_q^* - \mathcal{R}_H^* = \tfrac{1}{2}\mathbb{E}\Big[\|g(H) - \mathbb{E}[g(H) \mid q(H)]\|^2\Big].$$

The gap is positive when relevant conditional differences survive in the available records with positive statistical weight but are merged by q. This connects the displayed failure of a reduced description to C1. C2 separately supplies conditions under which an implemented learning process acquires a sufficient representational separation.

C.7 The observed environment can have unresolved conditions

A simple smooth hierarchy is

$$\dot{y} = f(y, z), \qquad \dot{z} = g(y, z, w).$$

Two states with equal (y, z) but different w have the same y and $\dot{y}$, while

$$\Delta\ddot{y} = D_z f(y, z)\, \Delta g(y, z, w).$$

Thus the current apparent environment need not determine its own continuation. Whether this separation is learnable depends on observations that preserve its trace, as required in the main text.

More generally, let the full physical state be x, with autonomous vector field F and observed output $h(x)$. Let $J_k(x) = (h, L_F h, \ldots, L_F^k h)$. Suppose a proposed smooth autonomous representation $z = \pi(x) \in \mathbb{R}^d$ has $D\pi F = f_z \circ \pi$ and $h = \bar{h} \circ \pi$, with smooth $\bar{h}$. Induction gives $L_F^j h = (L_{f_z}^j \bar{h}) \circ \pi$. Hence $J_k = \bar{J}_k \circ \pi$, and

$$\operatorname{rank} DJ_k(x) \le d.$$

Unbounded ranks preclude any finite-dimensional exact smooth autonomous realization of these observations. They do not preclude finite approximate models, and a representation can be sufficient without being identical to the physical world. This hierarchy explains one source of continuing separation without assuming that every environment forces endless learning.

D Dynamics, formation, and actual adoption

D.1 Coupled response and environmental return

Assume F is continuously differentiable near a reference trajectory on a finite time interval, with bounded Jacobians. Consider an implementation admitting smooth local coordinates $x = (s, \ell, e)$ for the two internal roles and the immediate environment, with further external conditions w. Here s denotes symbolic-role coordinates, which can include adopted-reference and received-response fields; it does not identify c_t with σ_t. This coordinate realization is additional structure, not a consequence of the functional separation alone. If the roles are readouts of an underlying state rather than independent coordinates, variations must be restricted to those induced by that state. For $\dot{x} = F(x, w)$, the trajectory variation satisfies

$$\delta\dot{x} = A(t)\delta x + G(t)\delta w, \qquad A = D_x F, \quad G = D_w F.$$

Let U be the propagator generated by A. Variation of constants gives

$$\delta x(T) = U(T, 0)\delta x(0) + \int_0^T U(T, u)G(u)\delta w(u)\, du. \tag{16}$$

For a continuously differentiable fixed-time adoption $s^+ = \mathcal{A}(\ell(T))$, keep the initial joint state fixed and set $\mathcal{U}_{\ell \leftarrow w}(T, t) = P_\ell U(T, t)G(t)$. Applying $D\mathcal{A}$ to the relational component gives

$$\delta s^+ = D\mathcal{A}(\ell(T)) \int_0^T \mathcal{U}_{\ell \leftarrow w}(T, t)\delta w(t)\, dt. \tag{17}$$

The adoption map is specified by the implementation under analysis. The empirical constructions in Appendix B.16 and below provide explicit formation maps under their stated sampling conditions. A

discrete class-creation rule is not automatically differentiable; the formula applies only to a smooth implementation or to a smooth region between adoption events. This uses no direct access to the hidden w by the learning algorithm. If the outer condition w responds to x, δw is its actual trajectory variation; computing a closed prediction requires adding those dynamics. It is not an independently available control or noise source by assumption.

The propagator has the Peano–Baker expansion

$$U(T,u) = I + \sum_{k=1}^{\infty} \int_{u<t_k<\cdots<t_1<T} A(t_1)\cdots A(t_k)\, dt_k \cdots dt_1.$$

If $\|A(t)\| \le L$, the kth term has norm at most $L^k(T-u)^k/k!$, proving uniform absolute convergence. Let P_s, P_ℓ be role projections. The four operators $P_i U P_j$, $i,j \in \{s,\ell\}$, collect all propagation paths between each pair of roles, including excursions through environmental coordinates. They classify response endpoints, not four compulsory temporal stages.

Writing $r = (s,\ell)$, the partitioned variational equations are

$$\delta\dot r = A_{rr}\delta r + A_{re}\delta e + G_r \delta w, \qquad \delta\dot e = A_{er}\delta r + A_{ee}\delta e + G_e\delta w.$$

Solving the second equation using the propagator Φ_e of A_{ee} alone gives

$$\delta e(t) = \Phi_e(t,0)\delta e(0) + \int_0^t \Phi_e(t,u)\{A_{er}(u)\delta r(u) + G_e(u)\delta w(u)\}\, du.$$

Substitution gives

$$\delta\dot r(t) = A_{rr}(t)\delta r(t) + \int_0^t M(t,u)\delta r(u)\, du + \xi(t). \tag{18}$$

where $M(t,u) = A_{re}(t)\Phi_e(t,u)A_{er}(u)$, and

$$\xi(t) = A_{re}(t)\Phi_e(t,0)\delta e(0) + G_r(t)\delta w(t) + \int_0^t A_{re}(t)\Phi_e(t,u)G_e(u)\delta w(u)\, du.$$

The full propagator's ee block is generally not Φ_e. Confusing them would count some return paths twice.

D.2 Retained representatives and adoption events

A useful special case retains an adopted s during a measurement window:

$$\dot\ell = F_\ell(\ell, s, t), \qquad s^+ = \mathcal{A}(\ell(T)).$$

If U_ℓ propagates $D_\ell F_\ell$, differentiation gives

$$D_s s^+ = D\mathcal{A}(\ell(T)) \left[U_\ell(T,0) D_s \ell(0) + \int_0^T U_\ell(T,u) D_s F_\ell(u)\, du \right].$$

The first term vanishes only when the initial relational state is independent of s. A direct dependence of $\mathcal{A}$ on s adds $D_s\mathcal{A}$. External consequences that depend on s must be included in the effective relational dynamics or in the joint state. This supplies the precise scope of the restricted return integral.

At a fixed-time adoption $z^+ = R(z^-)$, propagate variations by DR. For an autonomous event surface $h(z) = 0$ crossed transversely, with smooth reset R, the derivative across the event at a common comparison time is

$$\mathcal{S} = DR + \frac{(f^+ - DRf^-)\, Dh}{Dh\, f^-}.$$

This follows by differentiating the crossing time: $\delta t = -Dh\,\delta z^-/(Dh\, f^-)$. Time-dependent resets or event surfaces can be handled by augmenting the state with time. A held historical reference and a newly adopted representative are distinct events; a reference-preserving path cannot rewrite which earlier event occurred.

D.3 Formation and minimal response completion

Assume H has finite support with distribution $\mu(h) = \mathbb{P}(H = h) > 0$, and $Y \in \mathbb{R}^m$ is square-integrable. Let $E \subseteq L^2(\mu)$ be a linear function space containing constants, and let P_E be its orthogonal projection in the μ inner product. Define $g_j(h) = \mathbb{E}[Y_j \mid H = h]$ and $g = (g_1, \dots, g_m)$. Write

$$u_j = (I - P_E)g_j, \qquad \mathcal{N} = \operatorname{span}\{u_j\}, \qquad E^+ = E \oplus \mathcal{N}. \tag{19}$$

Theorem 17 (Reproduction by response completion)**.** *For half squared loss, the optimal active predictor in E has excess risk $\frac{1}{2}\sum_j \|u_j\|_\mu^2$ relative to the complete record. The space E^+ is the smallest linear extension of E that removes this excess for the declared returns. Conditional sample averages consistently estimate g_j from independent paired episodes under the fixed protocol. If an estimated new function is retained and installed in the prediction update, it changes the update's response as characterized below.*

The conditional-variance decomposition gives

$$\inf_{f_j \in E} \tfrac{1}{2}\mathbb{E}\|Y - f(H)\|^2 = \tfrac{1}{2}\mathbb{E}\|Y - g(H)\|^2 + \tfrac{1}{2}\sum_j \|(I - P_E)g_j\|_\mu^2.$$

Any linear extension E' containing E and attaining the full-record risk must contain every g_j. It therefore contains $g_j - P_E g_j = u_j$. Conversely, $E + \operatorname{span}\{u_j\}$ contains all g_j. This proves the minimality statement in Theorem 17.

Corollary 18 (The missing response becomes an update direction)**.** *In the scalar case of Theorem 17, let $g(h) = \mathbb{E}[Y \mid H = h]$, $f_0 = P_E g$, and $u = (I - P_E)g \neq 0$. Use population gradient flow for half squared loss in orthonormal readout coordinates. The old flow $\dot f = P_E(g - f)$ remains at f_0. After installing $E^+ = E \oplus \operatorname{span}\{u\}$, retaining the old coefficients and initializing the added coefficient to zero, the same readout gradient rule gives*

$$f(t) = P_E g + (1 - e^{-t})u, \qquad \dot f(t) = e^{-t}u. \tag{20}$$

For $\mathcal{R}(f) = \frac{1}{2}\mathbb{E}[(Y - f(H))^2]$, its remaining excess risk is

$$\mathcal{R}(f(t)) - \mathcal{R}(g) = \tfrac{1}{2}e^{-2t}\|u\|_\mu^2.$$

Proof. Orthogonality gives $P_E(g - f_0) = P_E u = 0$, so f_0 is stationary for the old readout. Since $g = P_E g + u$ belongs to E^+, the installed flow is $\dot f = P_{E^+}(g - f) = g - f$ for $f \in E^+$. Solving this linear equation from f_0 gives Equation 20. Conditional-mean orthogonality yields $\mathcal{R}(f) - \mathcal{R}(g) = \frac{1}{2}\|f - g\|_\mu^2$, proving the risk expression. □

Here the old class has already fitted everything it can express; its remaining prediction error supplies no further gradient within that class. Reconstruction and installation make the missing relation an available direction of the existing update rule. Every subsequent population-gradient change of this readout then lies along u, while the previously fitted component $P_E g$ is retained. This is a consequence of the response-completion construction, not a separate recovery controller. The exact trajectory assumes population gradients and the exact reconstructed u; the formation estimates and finite-sample comparison below quantify the use of an estimated relation.

The system can construct these functions without receiving hidden-factor labels. From independent episodes (H_i, Y_i), form the empirical conditional average at each observed h. Every state is visited infinitely often almost surely, and its average converges to $g(h)$. On a finite domain this gives simultaneous convergence and hence convergence in $L^2(\mu)$. Set the estimate to zero until the first visit. With known μ, the projection onto E is computable from its dictionary. The empirical residual then converges to u. An unknown μ requires consistent estimation of the Gram matrix as well.

For an explicit scalar-return confidence radius, suppose $|Y| \leq M$ and every state has at least $n_{\min}$ observations. Conditional Hoeffding bounds and a union bound over d record states give

$$\max_h |\widehat{g}(h) - g(h)| \leq M\sqrt{\frac{2\log(2d/\delta)}{n_{\min}}}$$

with probability at least $1 - \delta$. The orthogonal residual projection is a contraction in $L^2(\mu)$, so the same right side is a valid residual norm error bound. This connects the formation accuracy used in Equation 21 to actual samples.

D.4 KERNEL CHANGE AND FINITE-SAMPLE IMPROVEMENT

For this subsection take a scalar return Y and $g(H) = \mathbb{E}[Y \mid H]$. For the exact finite-sample comparison, take μ to be known and use independent formation and training data. Suppose $u = (I - P_E)g$ and formation supplies $\widehat{u} \in E^{\perp}$ with $\|\widehat{u} - u\| \le \eta < \|\widehat{u}\|$. Set $\widehat{\psi} = \widehat{u}/\|\widehat{u}\|$ and $d = \langle g, \widehat{\psi}\rangle$. Compare one batch update of size $\alpha \in (0, 1]$, from zero, in old and expanded orthonormal dictionaries on the same N fresh samples. Their expected true-risk improvement is

$$\alpha(1 - \alpha/2)d^2 - \frac{\alpha^2}{2N}\operatorname{Var}(\widehat{\psi}(H)Y). \tag{21}$$

If $|Y| \le M$, it is positive for $N > \alpha M^2/((2 - \alpha)(\|\widehat{u}\| - \eta)^2)$. The following computation proves both the update change and this improvement.

For a linearly independent dictionary ϕ, $G_E = \mathbb{E}[\phi\phi^{\top}]$ is positive definite. The update

$$a^{+} = a - \alpha G_E^{-1}\phi(H)(a^{\top}\phi(H) - Y)$$

changes a prediction at h' in proportion to $\phi(h')^{\top}G_E^{-1}\phi(H)$. If ψ is normalized and orthogonal to E, the expanded Gram matrix is block diagonal with blocks G_E and 1. This proves Equation 6.

Condition on a formed function $\widehat{\psi}$ independent of the subsequent training batch. Use orthonormal coordinates for both dictionaries, zero initialization, and the same batch. The old coordinates after the update are identical. The new coefficient is

$$b = \frac{\alpha}{N}\sum_{i=1}^{N}\widehat{\psi}(H_i)Y_i, \qquad \mathbb{E}b = \alpha d, \quad \operatorname{Var} b = \frac{\alpha^2}{N}\operatorname{Var}(\widehat{\psi}(H)Y).$$

Orthogonality gives the true-risk difference $b\,d - b^2/2$ for each realized batch. Taking expectations proves Equation 21. Moreover,

$$d = \frac{\langle u, \widehat{u}\rangle}{\|\widehat{u}\|} \ge \|\widehat{u}\| - \eta, \qquad \operatorname{Var}(\widehat{\psi}Y) \le \mathbb{E}[\widehat{\psi}^2Y^2] \le M^2.$$

These inequalities give the stated sufficient sample size. They show why an uncertain new relation is not automatically beneficial at a fixed training budget.

Population gradient flow from zero in orthonormal coordinates has $f_E(t) = (1 - e^{-t})P_E g$. Installing $\mathcal{N}$ therefore improves the loss at the same adaptation time τ by

$$\tfrac{1}{2}(1 - e^{-2\tau})\sum_j \|u_j\|_{\mu}^2.$$

This result is for the stated readout update. It does not replace the separate ordinary two-layer gradient-flow result in Appendix E.

D.5 ONE EXTERNAL PROCESS SUPPLIES THE FULL FORMATION CHAIN

The system issues a uniform sign A, encoding its probe setting. Let Z be an independent uniform sign and B an independent hidden sign with mean β. The device returns two coordinates

$$Y_1 = Z, \qquad Y_2 = AZB.$$

The record $H = (A, Y_1)$ is uniform and

$$g(H) = \mathbb{E}[Y_2 \mid H] = \beta A Y_1.$$

The hidden sign B and its mean are not supplied. The prediction task uses the setting and first coordinate as input and the second coordinate as an observed target. Conditional on a fixed mechanism θ in the noiseless neural protocol of Appendix E, take $A = q$, $Z = \zeta$, and $B = -\theta$, so $\beta = -\theta$. The formation law is fixed within each run; the equal-prior mixture over mechanisms is used only for the between-system comparison. A measurement flip replaces B by $-\theta$ times the flip sign. Thus the formation example and the neural calibration example use the same observation law. If A and every

means of recovering it are erased before the pair is used, (Y_1, Y_2) are independent uniform signs for all β. The optimal half squared risk is $1/2$.

Even with the complete record, the active class $E = \operatorname{span}\{1, A, Y_1\}$ has the same floor. The interaction AY_1 is orthogonal to all three old functions. Empirical conditional averages form its coefficient from actual returns. Adopting this function expands the update through Equation 6. For the exact added feature, one batch update has improvement

$$\alpha(1 - \alpha/2)\beta^2 - \frac{\alpha^2(1 - \beta^2)}{2N}$$

over the paired old update. To improve on the old class's optimum, rather than on that noisy old update, it suffices that

$$N > \frac{\alpha(4 - \beta^2)}{(2 - \alpha)\beta^2}, \qquad \beta \neq 0.$$

The extra $3\alpha^2/(2N)$ accounts for noise fitted by the three old coordinates. If $\beta = 0$, there is no population relation for this task to acquire. This example is analytical, not a benchmark experiment.

D.6 Nonlinear role changes and return geometry

For a four-part differentiable implementation $z_j = T_j(z_{j-1}, \rho_j)$, let $A_j = D_z T_j$ and $B_j = D_\rho T_j$ along the actual trajectory. With a fixed initial state,

$$\delta z_4 = A_4 A_3 A_2 B_1 \delta\rho_1 + A_4 A_3 B_2 \delta\rho_2 + A_4 B_3 \delta\rho_3 + B_4 \delta\rho_4.$$

An effect entering early is transformed by the subsequent operations. Boundary interactions can be inserted as additional maps. This identity holds for a declared implementation; it does not deduce that every learning system needs four physical instructions. Likewise, a nonzero Lie bracket describes order dependence, not automatically knowledge or improvement. Appendix E supplies an instance where external evidence selects a useful noncommuting preparation.

E Self-modification through ordinary neural gradient flow

E.1 Protocol and permissions

The network is $f_{W,V}(x) = WVx$, with $W \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{n\times p}$ and $\mathbb{E}[xx^\top] = I_p$. For a matrix target T, the loss is $L_T = \|WV - T\|_F^2/2$. Preparation may use targets computed by the system itself. The target is treated as constant by differentiation, so the parameter update is always

$$\dot{W} = EV^\top, \qquad \dot{V} = W^\top E, \qquad E = T_{\text{int}} - WV. \tag{22}$$

This is an explicit internal control permission. The environment need not supply arbitrary positive or negative training signals. Preparation targets are not additional evidence about the external mechanism.

Set $W_0 = D \succ 0$ diagonal and $V_0 = 0$. Preparation ends with $WV = 0$, before the future target is revealed. Both comparison branches have the same control permissions and a sufficient common preparation budget. An additive fixed predictor f_0 or a fixed whitened nonlinear feature map can be included: $f(x) = f_0(x) + WV\phi(x)$. Only the two factors are trained in this extension. The results do not assume an arbitrary end-to-end network has the same reachable set.

E.2 The complete attainable response orbit

Theorem 19 (Attainable learning responses)**.** *For any $p \geq 1$, the preparation endpoints of Equation 22 with $WV = 0$ are exactly $(QD, 0)$, $Q \in SO(n)$. Every such endpoint is attainable by finitely many continuous feedback segments. The successor's initial response is $K_Q[E] = QD^2Q^\top E$. For a future target T and preparation information $\mathcal{I}$, the best expected initial loss decrease is*

$$\mathcal{V}(\mathcal{I}) = \mathbb{E} \max_{Q\in SO(n)} \operatorname{tr}\big(\mathbb{E}[TT^\top \mid \mathcal{I}]QD^2Q^\top\big). \tag{23}$$

Differentiation gives

$$\frac{d}{dt}(W^\top W) = VE^\top W + W^\top EV^\top = \frac{d}{dt}(VV^\top).$$

Thus $W^\top W - VV^\top = D^2$ throughout preparation. This is the familiar balancedness invariant of homogeneous networks (Du et al., 2018), used here to constrain all admissible feedback paths. Since $W^\top W = D^2 + VV^\top \succ 0$, W stays invertible. At a returned function $WV = 0$, one has $V = 0$ and $W^\top W = D^2$. Continuity preserves the positive determinant, hence $W = QD$ with $Q \in SO(n)$. Differentiating the current function $R = WV$ gives

$$\dot{R} = WW^\top E + EV^\top V.$$

At a returned endpoint its response is therefore $K_Q[E] = QD^2Q^\top E$.

It remains to prove attainability, rather than only a necessary invariant. For $p = 1$, write $V = v$ and $r = Wv$. The matrix

$$K = WW^\top + \|v\|^2 I$$

is positive definite. For a prescribed output velocity u, use the internal error $e = K^{-1}u$. Then

$$\dot{W} = ev^\top, \qquad \dot{v} = W^\top e, \qquad \dot{r} = u.$$

The target $\mathrm{stopgrad}(r + K^{-1}u)$ realizes this error through the same gradient law.

Let h_i be the resulting vector field for $u = e_i$. With the convention $[X, Y] = DY\,X - DX\,Y$, direct differentiation at $(D, 0)$ gives

$$[h_i, h_j]_v = 0, \qquad [h_i, h_j]_W = \frac{E_{ji}}{d_i d_j^2} - \frac{E_{ij}}{d_i^2 d_j} = \frac{(E_{ji} - E_{ij})D}{d_i^2 d_j^2}.$$

To turn this local calculation into an exact finite path, fix a sufficiently small $\delta \neq 0$ and traverse an output rectangle with sides $\delta e_i, be_j, -\delta e_i, -be_j$. The output returns exactly. The state stays within the corresponding two-dimensional block, so the endpoint is $R_{ij}(\varphi_\delta(b))D$. At $b = 0$ the path retraces itself and $\varphi_\delta(0) = 0$. Smooth dependence and the mixed derivative above give

$$\varphi'_\delta(0) = \frac{\delta}{d_i^2 d_j^2} + O(\delta^2) \neq 0.$$

The inverse function theorem therefore realizes every sufficiently small positive or negative angle by a finite rectangle.

The gradient system is equivariant under $(W, r, e) \mapsto (QW, Qr, Qe)$. The same small-angle construction can be transported to every returned endpoint. A finite repetition gives any plane rotation, and a finite Givens decomposition gives every element of $SO(n)$. For $p > 1$, act on the first column of V and keep the others zero. The same construction proves the claimed orbit for all $p \geq 1$.

A common finite preparation bound can be chosen for this orbit. There are finitely many coordinate planes. Choose a common small-angle interval and split each bounded Givens angle into a fixed sufficient number of pieces. Equivariance preserves the local bounds. A measurable decomposition rule and the continuous local inverse give measurable feedback selections. Finite numerical implementations approximate these continuous paths; their error is included in the successor consistency result.

E.3 Information value on the attainable orbit

At $(QD, 0)$, the loss on a future target has initial derivative

$$-\dot{L}_T(0) = \mathrm{tr}(TT^\top QD^2Q^\top).$$

Let $\mathcal{I}$ be the complete preparation information. For $\mathbb{E}\|T\|_F^2 < \infty$, conditional expectation reduces the optimal expected initial decrease to the maximum in Equation 23. Compactness of $SO(n)$ and continuity yield a maximum; measurable maximizing choices exist for standard Borel information spaces. The reachability construction realizes those choices.

If λ_i and μ_i are the eigenvalues of D^2 and $H = \mathbb{E}[TT^\top \mid \mathcal{I}]$ in decreasing order,

$$\max_Q \operatorname{tr}(HQD^2Q^\top) = \sum_i \lambda_i \mu_i.$$

In eigenvector coordinates the trace is $\sum_{ij} \mu_i \lambda_j U_{ij}^2$. The squared entries form a doubly stochastic matrix, and ordered pairing attains the maximum. Changing one eigenvector's sign enforces determinant one without changing the conjugated matrix.

For $\mathcal{G} \subseteq \mathcal{F}$, define $\rho(H) = \max_Q \operatorname{tr}(HQD^2Q^\top)$. It is convex as a maximum of linear functions, so conditional Jensen gives $\mathcal{V}(\mathcal{F}) \geq \mathcal{V}(\mathcal{G})$. Equality holds exactly when a coarse-information maximizing orientation is also optimal almost surely for the merged fine-information cases. This follows by taking the conditional expectation of the nonnegative difference $\rho(H_\mathcal{F}) - \operatorname{tr}(H_\mathcal{F} K_\mathcal{G})$. Additional information has strict value only when it changes a useful attainable choice.

E.4 A SOURCE-DEPENDENT ENVIRONMENT

Here is the full two-dimensional protocol described in Section 4. Set $D = \operatorname{diag}(1, 2)$, $u = (1, -1)^\top$, $v = (1, 1)^\top$, and $A_+ = [u, v]$, $A_- = [v, u]$. The mechanism $\theta \in \{+1, -1\}$ has equal prior probabilities. A preparation request S is uniform on $\{e_1, e_2\}$ and returns $Y = \zeta A_\theta S$, with an independent uniform sign ζ. The future target is $T = \zeta' A_\theta e_1$, with a fresh independent sign, revealed after preparation. Preparation requests are generated in an independent uniform batch before returns are processed. If the source identity is erased, every return is uniform over $\{u, -u, v, -v\}$ under both mechanisms. The anonymous record includes the common initialization, all anonymous returns, and independent internal randomness. There is no retained seed, order convention, parameter trace, or external receipt from which the individual requests can be recovered. If such a path exists, it belongs to the source-preserving case.

For retained sources, $q = 1$ when $S = e_1$ and $q = -1$ when $S = e_2$. The two return coordinates satisfy $Y_1 = \zeta$ and $Y_2 = -\theta q \zeta$, so $-qY_1Y_2 = \theta$. The relation is inferred from the request and its actual outcome. The later target $T = \zeta' A_\theta e_1$ shares the mechanism but has a fresh independent sign. Even after learning θ, the Bayes prediction before that sign is revealed is zero.

Every returned endpoint has $\operatorname{tr} K_Q = 5$. Since $uu^\top + vv^\top = 2I$, any anonymous choice has equal-prior expected initial decrease

$$\tfrac{1}{2}\mathbb{E} \operatorname{tr}((uu^\top + vv^\top)K_Q) = 5.$$

In contrast, aligning the eigenvalue 4 with the relevant target line attains decrease 8 under either mechanism. For the usual two-dimensional rotation $Q(\alpha)$, these decreases are $5 + 3\sin(2\alpha)$ and $5 - 3\sin(2\alpha)$. Thus opposite orientations are required.

E.5 FIXED-PERIOD LOSS AND CONSISTENT SELECTION

Theorem 20 (Unlearnable and learnable successor selection)**.** *Let $\ell_\theta^t(Q)$ be the expected loss after the successor at $(QD, 0)$ follows the original gradient flow for a fixed period t. In the preceding two-dimensional protocol: (i) there exist $C < \infty$ and $t_0 > 0$ such that every anonymous preparation has expected excess loss at least $3t - Ct^2 > 0$ for $0 < t < t_0$, regardless of the number of preparation observations; (ii) with retained source relations, a consistent feedback estimator, increasingly dense rotation candidates, and controlled implementation accuracy give*

$$\mathbb{E}\,\widetilde{\ell}_{\theta,N}^t \longrightarrow \mathbb{E} \min_{Q \in SO(2)} \ell_\theta^t(Q) \tag{24}$$

for every fixed $t > 0$. Here $\widetilde{\ell}_{\theta,N}^t$ is the loss of the actually implemented successor. The estimator and accuracy schedule are constructed in Appendix E.5, including a noisy-observation version.

Initial-speed optimality alone does not prove fixed-period optimality. For the finite target set and compact endpoint orbit, smooth ODE dependence gives a uniform expansion

$$\ell_\theta^t(Q) = 1 - g_\theta(Q)t + R_\theta^t(Q), \qquad |R_\theta^t(Q)| \leq C_0 t^2$$

on a common short interval. The common interval exists because initial states and targets lie in compact sets. Solutions remain bounded on every finite interval: loss decreases, $\|Wv\| \leq 2\|T\|$, the invariant gives $\sigma_{\min}(W) \geq \min_i d_i$, and hence both v and W are bounded.

The anonymous average loss is at least $1 - 5t - C_0t^2$. A mechanism-specific orientation has loss at most $1 - 8t + C_0t^2$. Consequently,

$$\inf_{\text{anonymous preparation}} \mathbb{E}\left[\ell_\theta^t(Q) - \min_{Q'} \ell_\theta^t(Q')\right] \geq 3t - 2C_0t^2 > 0$$

for sufficiently small positive t. This inequality is uniform over all permitted anonymous preparations and all preparation sample sizes.

For a noisy preparation version, independently multiply the observed second coordinate by a sign B with known flip probability $\varepsilon < 1/2$. The return is $(\zeta, -\theta q \zeta B)$. Anonymous returns remain uniform under both mechanisms. Paired returns give $Z = -qY_1Y_2 = \theta B$. The sign of the average of N such values estimates θ with error

$$\alpha_N \leq \exp\bigl(-N(1-2\varepsilon)^2/2\bigr).$$

The future target is the uncorrupted environmental response specified above. Preparation measurement noise is part of the common observation protocol, not a supplied target for the optimal modification.

For each fixed adaptation period $t > 0$, choose increasingly dense finite grids $\mathcal{Q}_N \subset SO(2)$. Uniform continuity of ℓ_θ^t on the compact orbit gives grid error $\omega_N \to 0$, uniformly over the two mechanisms. For each grid point, integrate the known successor gradient equations for the two possible future signs. Control the loss-evaluation error by $\nu_N \to 0$. Choose the best estimated grid candidate $\widehat{Q}_N$ and implement its feedback path. Let $\widetilde{\ell}_{\theta,N}^t$ be the actual successor loss, with $|\widetilde{\ell}_{\theta,N}^t - \ell_\theta^t(\widehat{Q}_N)| \leq \delta_N \to 0$. The finite-dimensional smooth flow and constructive local inverses permit these accuracy schedules. Numerical precision is used to select candidate angles and rectangle side lengths. Each implemented preparation remains an exactly closed output path under the continuous feedback law above, so its actual endpoint is $(\widetilde{Q}_N D, 0)$ with $\widetilde{Q}_N \in SO(2)$. The implementation error measures the resulting rotation error, not a relaxation of the return constraint. A fully discrete gradient implementation would require an additional discretization analysis.

All exact-endpoint losses lie in $[0, 1]$. On a correct mechanism estimate, the excess loss is at most $\omega_N + 2\nu_N + \delta_N$. On an incorrect estimate it is at most $1 + \delta_N$. Therefore

$$0 \leq \mathbb{E}\left[\widetilde{\ell}_{\theta,N}^t - \min_{Q} \ell_\theta^t(Q)\right] \leq \alpha_N + \omega_N + 2\nu_N + \delta_N \longrightarrow 0.$$

This proves Equation 24 and the main-text Equation 8, with modifications identified by the attainable rotation and $\tau = t$. Increasing data alone with a permanently coarse search or implementation would not give this conclusion. The specified schedule increases those accuracies as well.

E.6 What the neural result tests

Source preservation is necessary here for selecting the useful modification, not for making the geometric path exist. The path is available to both comparison branches. If D^2 is scalar, its conjugation orbit has one response matrix and the information advantage vanishes. If the future target is disclosed before preparation, the anonymous branch can use that disclosure. If arbitrary parameter overwrites or a different probing policy are allowed, the attainable class or information protocol changes. These are distinct, testable changes of conditions.

The result reconfigures an actual learning response through unchanged gradient dynamics. It is stronger than observing two parameter settings with the same prediction. It remains a bounded preparation-and-adaptation theorem, whose recursive use is covered next.

E.7 Executed numerical mechanism test

The program accompanying this construction is `experiments/causal_return/self_modification.py`. It uses NumPy and SciPy for estimation and integration, and CPU PyTorch to check the analytical gradient against automatic differentiation. The test has 128 independent seeds, with both mechanisms evaluated using paired exogenous randomness within each seed. Formation

Condition	Prediction loss	Successor loss
Complete process	0.187	0.525
Source correspondence removed	0.504	0.688
Original feature class fixed	0.506	0.688
Formed relation, no selected deployment	0.187	0.688

Table 2: Forming a relation and deploying it affect different losses. Means over 128 seeds with two paired mechanisms, 10% measurement noise, 256 formation and 256 independent training returns per branch, and adaptation period 0.08. Losses are half squared errors; preparation costs are excluded.

and coefficient training use independent sample splits, each of size $N \in \{16, 64, 256\}$. Measurement flips have probability 0.1. The future adaptation period is 0.08, and the selection grid contains 41 angles from $-\pi/2$ to $\pi/2$.

In the scalar coordinates of Section 4, each calibration encounter assigns a uniform probe sign $A \in \{-1, 1\}$, receives an independent uniform sign Z, and then receives Y_2. With probability $1-\epsilon$, the latter equals $-\theta AZ$; measurement noise reverses it otherwise. One hidden sign $\theta \in \{-1, 1\}$ connects these observations to the future task:

$$\mathbb{E}_\theta[Y_2 \mid A, Z] = -\rho\theta AZ, \qquad y_{\mathrm{next}} = \xi(1, -\theta)^\top, \qquad \rho = 1 - 2\epsilon > 0.$$

The measurement-flip rate ϵ is known. The independent uniform sign ξ is revealed only after preparation. Paired records of A, Z, Y_2 identify the target line without disclosing this future sign. The learned line tells the system where to direct its stronger response. Removing the source A leaves only (Z, Y_2), whose law is the same under both values of θ. Those records cannot identify the target line. The program estimates the relation through conditional means, compares attainable orientations over the fixed adaptation period, and executes the selected path by integrating the original gradient field.

The learner receives the finite observed alphabet, initial dictionary, known flip rate, and declared future query. It estimates a response residual from records, rather than receiving the hidden mechanism or the product feature as a prescribed formula. The four-record domain has a one-dimensional complement to the old dictionary, so this is not unrestricted symbolic structure discovery. The old-class ablation skips new-feature formation. The anonymous branch receives no source column. The no-deployment branch learns the same feature but executes a zero-net-rotation preparation.

Candidate endpoints are used only to simulate the known future gradient equations. Actual deployed parameters are obtained by DOP853 integration of the feedback-controlled original gradient field. All branches use the same number of output rectangles; their energy and search cost are not equalized. Across all evaluated preparations, the largest output-return error is 6.50×10^{-11} and the largest balancing-invariant drift is 1.12×10^{-10}. The analytical gradient differs from automatic differentiation by at most 1.39×10^{-17} in the stated check. The measured initial loss decrease uses the actual residual $T - Wv$, not an assumption of exact numerical return.

Table 2 uses $N = 256$. Prediction-loss standard errors across independent seed-pair means are 0.000421, 0.000369, 0.000418, and 0.000421, respectively. Successor-loss standard errors are numerically zero for the complete and no-deployment branches at the displayed precision, and 0.01412 for the other two branches. The same controls are compared on equal observation counts. The noisy-return Bayes half squared risk is 0.18; the old active class has population risk floor 0.5. The complete branch reaches the best evaluated grid loss in these runs. This finite-grid fact does not certify the continuous asymptotic optimum in Theorem 20.

The test realizes response formation and neural deployment. It does not numerically test Theorem 3's stable-transition partition algorithm or multi-generation improvement. The latter claims retain their separately stated proofs and conditions.

F FURTHER LEARNING AND RSI CONSEQUENCES

F.1 COMPRESSION RELATIVE TO FUTURE OPERATIONS

Let $\mathcal{P}$ be a family of permitted successor operations, and let $m_p(h)$ be the required conditional consequence of operation p after record h. An exact compressed record $R = Q(H)$ supports all these predictions if and only if each m_p factors through Q on the supported records. Define

$$h \sim_{\mathcal{P}} h' \quad \Longleftrightarrow \quad m_p(h) = m_p(h') \text{ for every } p \in \mathcal{P}.$$

The quotient by this equivalence is the coarsest exact record for this family. If $\mathcal{P} \subseteq \mathcal{P}'$, then $\sim_{\mathcal{P}'}$ refines $\sim_{\mathcal{P}}$. The proof follows directly from factorization. Supporting additional modifications can therefore require preserving separations irrelevant to the old ones.

For selecting an optimal operation, rather than predicting every consequence, a weaker record can suffice. In finite candidate and record spaces, a coarse record has zero decision loss exactly when every positive-probability coarse class admits a common optimizer for its fine records. Otherwise there is a strictly positive gap. Thus the relevant compression criterion depends on the required future use.

For a language system, matching its current textual answer is not a sufficient test of this criterion. A compression should also be tested on its ability to use later evidence to revise relations and choose new operations. Continuous summaries, structured relations, recoverable source records, and retrieval can each preserve such separations. Soft summaries already have concrete implementations (Chevalier et al., 2023). Neither text nor hidden state alone determines whether the criterion holds.

F.2 GENERALIZATION WITH UNREPRESENTED SUPPORT

Fix a finite declared family of available contexts $\mathcal{H}^\dagger$, operations $\mathcal{I}$, and actual consequence laws $\kappa_i(h)$. The family specifies the source and target uses under discussion. Let $q : \mathcal{H}^\dagger \to \mathcal{R}$ be the representation used for these predictions. A context may encode an observed configuration without expressing the mechanisms that make its consequence law hold. This distinguishes representing a condition from knowing its complete causal realization.

Corollary 21 (Validity and error of representational reuse)**.** *A response rule using only $q(h)$ can predict all the declared laws exactly if and only if*

$$q(h) = q(h') \implies \kappa_i(h) = \kappa_i(h') \quad \textit{for all } i \in \mathcal{I}.$$

Let h_0, h_1 share a representation. Suppose $\widehat{\kappa}_i$ approximates the source law with total-variation error at most ϵ. Let Δ_i be the total-variation distance between the source and target laws. The transferred prediction obeys

$$\mathrm{TV}(\widehat{\kappa}_i, \kappa_i(h_1)) \le \epsilon + \Delta_i.$$

If both contexts require prediction and $\Delta_i > 0$, every common predicted law has error at least $\Delta_i/2$ on at least one of them.

Proof. The first statement is Proposition 8 on the declared source–target family. The second is the triangle inequality through $\kappa_i(h_0)$. For any common prediction ν, the same inequality gives $\Delta_i \le \mathrm{TV}(\kappa_i(h_0), \nu) + \mathrm{TV}(\nu, \kappa_i(h_1))$, which proves the lower bound. □

The upper bound separates source estimation error from a difference between source and target consequence laws. Changes in unrepresented causal support can produce this difference. Under invariance, the subject can reuse the relation without identifying all the causal mechanisms behind it. When the two declared contexts require different laws, improving a single common estimate cannot make it correct for both; the operative representation must retain a distinguishing condition to do so. Causal transportability provides a broader account of which cross-environment differences permit causal conclusions to transfer (Pearl & Bareinboim, 2014). Here the condition is attached to the formation, adoption, and subsequent revision of a representation.

This validity statement does not itself estimate an unobserved target law. The finite formation theorems learn on their declared sampling support; additional structural or empirical information is

required to identify arbitrary new contexts. A different kind of reuse is already available within the learned response family: the later-cost result below supports objectives not specified during formation. Both cases explain how reconstruction can acquire a scope beyond a particular past encounter, while making explicit what continues to be supplied by unrepresented support.

F.3 Coordinated changes of symbols and relations

Equation 14 gives the invariant for joint updates. If an invertible change of encoding T transforms the state and the decoder becomes $c' = c \circ T^{-1}$, the reference is preserved exactly. Freezing, slow updates, version mappings, and suitable joint transformations can all implement this requirement. The theorem does not say that simultaneous updates must fail.

Target-network and dictionary methods provide related engineering mechanisms (He et al., 2020; Tian et al., 2021). The present test concerns whether the required reference remains recoverable under the actual update. It must be checked for the downstream inference at issue, not inferred from a method's name or from a nominal stop-gradient operation.

F.4 Learning without labels for the new factor

In the construction of Appendix D.3, the observed pair and the system's own adopted condition suffice to estimate $\mathbb{E}[Y_2 \mid A, Y_1]$. No supplied label identifies the hidden sign B, its mean β, or the correct feature. These are not unsupervised discoveries from an information-free state: the actual interaction supplies observations and the system has reconstruction operations. The distinction is between supplying a correct latent concept and learning a predictive relation from the effects it produces.

This can support later objectives not specified during formation. For finite outcome spaces, suppose estimated response laws satisfy $\sup_a \mathrm{TV}(\widehat{p}_a, p_a) \le \epsilon$. For every later cost c with $0 \le c \le 1$,

$$|\mathbb{E}_{\widehat{p}_a} c - \mathbb{E}_{p_a} c| \le \epsilon.$$

Choosing the estimated best action in a finite admissible set incurs excess expected cost at most 2ϵ. Thus forming a response model need not require a terminal task reward. Selection for a particular later purpose still needs that purpose and an appropriate comparison. The claim concerns reusable response predictions, not unique identification of every latent cause.

F.5 Allocating capacity to useful unrepresented relations

For response coordinates j, let $u_j = (I - P_E)g_j$ and let $\omega_j \ge 0$ be declared task weights. Define the finite-rank positive operator

$$S = \sum_j \omega_j u_j \otimes u_j \quad \text{on } E^{\perp}.$$

For an added space $W \subseteq E^{\perp}$, population adaptation from zero for time τ improves the weighted loss by

$$\tfrac{1}{2}(1 - e^{-2\tau}) \operatorname{tr}(P_W S).$$

Under a budget of b dimensions, the optimum uses the leading b eigenvectors of S. This is the standard variational principle for eigenvalue sums applied to the response residual. Variance unrelated to later consequences does not enter S.

Different return families and task weights therefore favor different representational changes. This gives a measurable account of different learning emphases. Assigning a language model, world model, or action model to a fixed phase requires evidence about its response and adoption operations; the architecture name alone does not supply it.

F.6 Selecting and continuing self-modification

A system allocating training between frequent errors and rarely visited conditions supplies an example. Each choice changes the model and later evidence. Recorded starting configurations and

consequences let one executed modification guide the next. Learning operations themselves enter the causal return.

Under the controlled-process and continuation conditions of Section 3.2, the learned outcome and transition relations define a decision process over the reconstructed categories. The system estimates and solves this process internally. Consistent estimates and vanishing planning error make the selected rule approach the optimal discounted value available in the declared operation family (Appendix A.8). This result concerns the system's executed choices, including their later consequences, rather than prediction accuracy alone.

Recursive application requires recoverable source correspondences, informative interaction, reconstruction and deployment in each successor. Features and update rules can then change through relations formed from earlier modifications. Some rounds improve estimates without adding categories; a sufficient representation can stabilize. The selected-rule guarantee concerns the declared operation family; continued exploration incurs a separate cost.

F.7 Recursive use and its stopping conditions

At generation n, let h_n be the actual history, $\mathcal{P}_n(h_n)$ the executable modification class, and $\ell_n(h_n, p)$ the expected successor loss. Let $\widehat{\mathcal{P}}_n(h_n) \subseteq \mathcal{P}_n(h_n)$ be the constructed candidate class. Suppose its optimal true loss exceeds the infimum over $\mathcal{P}_n(h_n)$ by at most a_n. The empirical loss estimate is uniformly accurate to ϵ_n on $\widehat{\mathcal{P}}_n(h_n)$, selection is within η_n of the empirical infimum on that class, and execution adds at most δ_n loss. Successive replacements give

$$\ell_n(h_n, \widetilde{p}_n) - \inf_{p \in \mathcal{P}_n(h_n)} \ell_n(h_n, p) \le 2\epsilon_n + \eta_n + \delta_n + a_n.$$

If these bounds hold conditionally along actual histories with conditional expectation at most e_n, then

$$\mathbb{E} \sum_{n=1}^{G} \left[\ell_n(h_n, \widetilde{p}_n) - \inf_{p \in \mathcal{P}_n(h_n)} \ell_n(h_n, p) \right] \le \sum_{n=1}^{G} e_n.$$

If $e_n \to 0$, the mean per-generation excess tends to zero. This compares decisions along the actual history, not with a different globally optimal sequence that generates a different world. Including the unchanged system as a candidate gives approximate non-degradation relative to that contemporary alternative. Strict improvement requires a strict opportunity.

A linear active-response realization makes a different stopping mechanism explicit. Let B map adopted conditions to their identifiable external responses, and C map executable representations to adopted conditions. For a current finite-dimensional space E set

$$U = CE, \quad E^+ = E + BU, \quad U^+ = CE^+.$$

There are induced surjections

$$E^+/E \xrightarrow{\bar{C}} U^+/U \xrightarrow{\bar{B}} (E^+ + BU^+)/E^+.$$

They are well-defined because $CE = U$ and $BU \subseteq E^+$. Consequently, the next response increment cannot exceed the deployed increment, which cannot exceed the current representation increment. An added relation annihilated by deployment creates no new condition. A newly deployed condition can also have a response already expressible in E^+. In nonlinear or changing environments these linear rank bounds need not persist, but the reception, deployment, and response questions remain separate.

F.8 The meaning of an unexpressed relation

A relation can affect a system without being independently reusable. The operational test is whether later allowed comparisons, queries, or updates can recover and use the separated information. This captures part of the motivating notion of intuition without asserting that every felt or latent tendency is reliable. Reconstruction can produce errors; reliability comes from its connection to actual consequences.

A hypothetical Truman-style experiment has the same criterion. If two worlds induce equal complete observation laws for all permitted interactions, internal reasoning cannot distinguish them. If further

external conditions produce different available consequences, the system can test and represent those differences. The framework predicts an identifiability boundary, not a guarantee that every simulated world contains detectable inconsistencies.

G Scope and relation to the learning literature

G.1 The connected learning claim

The paper follows an unrepresented causal influence through its symbolic projection, relational reconstruction, and use in self-modification. The causal projection identifies why the current internal relations cannot determine every enacted transition. Returns enrich the associations of the projected effect; comparison and reconstruction connect it to usable conditions. The formation constructions make these relations operative, and the successor analysis evaluates an executed change. Within each construction, the observation, reference, and execution interfaces describe one continuing process. Reception limits and source preservation supply conditions at particular points in that return; they are not the cause of the initial separation. The controlled-partition realization proves automatic formation under its sampling and reconstruction assumptions.

Organizational closure supplies a direct foundation for identifying maintained activity without equating its boundary with physical isolation (Montévil & Mossio, 2015). Our causal formulation includes computational use and its unrepresented support in the same process. The boundary then supports two response comparisons, inward and outward. Their factorization criterion specifies which differences cannot be merged when preserving the respective effects. The subsequent formation and execution results supply learning conclusions beyond this structural statement.

The direct foundations have distinct roles. Binding theory supports identity and compositional use (Greff et al., 2020). DORA provides a mechanism for relation formation through comparison (Doumas et al., 2008). DreamCoder makes learned abstractions available to future generation (Ellis et al., 2021). Our response-completion construction specifies the required increment and its effect on a learning update. The finite-data bound connects the accuracy of formation to the benefit of adoption.

Training-dynamics results support a different part of the argument. Self-supervised analyses study the effects of asymmetric updates (Tian et al., 2021). Function-preserving parameter transformations can alter later optimization (Zhao et al., 2024). Here the full response orbit follows from the original gradient law, and the information obtained from the environment determines which orbit element can be selected. The two branches have identical control access. The difference is their ability to learn the useful control from the available consequences.

Conditional expectation, variational response, and spectral alignment are established mathematical tools. Projection methods also account for unresolved effects through memory and forcing (Chorin et al., 2000). Their use here supports explicit learning statements: the loss from a restricted record, the missing response in an active update, and the value of forming and adopting that response. The causal-boundary decomposition and its connection to relational reconstruction organize these tools around the separation and return of the same operative representative.

G.2 Relational architectures and reuse

Relational architectures provide concrete ways to compare identifiable objects. ESBN binds keys to perceptual values, and Abstractors organize relations through object-identifying symbols (Webb et al., 2021; Altabaa et al., 2024). For the cycle to continue, the resulting relation must become usable in a subsequent prediction, action, or update. Skill–symbol loops and active exploration implement such reuse in action (Konidaris, 2015; Andersen & Konidaris, 2017). For self-improvement, the relevant use is a change to the learning process.

G.3 Self-reference and the Gödel machine

The Gödel machine permits self-modification of its own software, including its search process (Schmidhuber, 2006). Its switching theorem compares a candidate against continuing the current program under its axioms and utility. Our results characterize empirical identifiability and

execution in a concrete learning process. They do not assert a contradiction in that switching theorem.

A Gödel-machine acceptance proof would additionally need to compare the actual continuation, future search, timing, and costs. The self-reference studied here is operational: the system changes processes that determine evidence for further changes. No logical paradox is needed for the source-dependent learning limit.

G.4 System boundary and learning signals

Symbolic and relational roles are determined by use, rather than by storage format. For each result, the record, observation channel, and update permissions are explicit. A reward generator or optimizer can influence the represented organization without its generating rule being represented there. The system can model such a rule from its effects, but learning the rule does not by itself grant permission to change it.

The four positions describe the formation and reuse of a particular relation. They can overlap, recur, or be implemented at different timescales. The four source-to-target classes of a variational propagator answer a different question about influence. Neither description proves a universal four-instruction implementation. The constructive results specify their actual reception, comparison, reconstruction, and adoption operations.

G.5 Analytical scope

The paper reports analytical results, constructive algorithms, and a bounded numerical mechanism test. Automatic stable separation concerns finite accessible records, policy-invariant controlled laws, and continuing internal estimation and deployment. Its proof is distinct from the response-feature construction tested in the neural experiment. The general response analysis applies locally to smooth coupled dynamics, with event corrections for adoption. Response completion uses a finite record space and an active linear span of possibly nonlinear functions. The exact neural guarantee uses a two-layer linear factorization, controllable internal targets, and fixed-period future evaluation. The proof of consistency supplies the estimation, candidate coverage, and implementation schedules.