

When Reduction Becomes Lossy: An Intervention-Relative Test of Explanatory Boundaries

Keywords: causal abstraction; coarse-graining; Markov lumpability; intervention; computational cognition; agent–environment coupling; explanatory sufficiency

Abstract

Can a restricted computational model predict well while omitting distinctions required by its explanatory task? We define boundary sufficiency relative to an outcome, a representation, and a declared family of input interventions. Exact sufficiency requires a common response law on every fiber of the retained representation throughout that family; observational predictive accuracy alone does not establish it. We prove that, under a fixed distribution, the excess optimal log loss from discarding distinctions equals conditional mutual information, and separate this irreducible within-regime loss from the transport loss of a predictor fitted under another regime. Three finite-state synthetic systems implement the test. A strongly lumpable four-state chain is an exact negative control: aggregation incurs zero regret under every tested intervention. A non-lumpable chain has observational regret 0.080 nat per outcome, but regret 0.368 nat after aggregate refitting and 0.733 nat when the observational predictor is frozen under uniform microstate intervention. In an agent–context model with the same observational regret, changing context to oppose the agent gives 1.460 nat frozen-predictor regret. Thirty seeded simulations reproduce these patterns with finite data; sensitivity sweeps include the zero-contrast counterexample. The resulting criterion diagnoses a failure of a specified computational boundary, not an ontological failure of complete microdescription. No empirical claim about neural or cognitive mechanisms follows from these constructed examples.

1. Introduction

Computational models of cognition necessarily omit features of their targets. They discretize time, compress neuronal populations into state variables, treat environmental inputs as fixed, or replace a network by a tractable effective process. Such omissions are often epistemically productive. Their legitimacy cannot be decided simply by counting represented components. It depends on what the model purports to explain and on the circumstances in which its explanatory claims are meant to remain true. A model that predicts behavior in a stable laboratory regime may fail when an experimenter changes an input that was previously correlated with an internal state. Conversely, an aggregation may preserve precisely the transitions that matter, even while discarding many distinctions.

This paper gives a narrow, testable meaning to an *explanatory boundary error*. A candidate model retains a representation B of a more complete state S and addresses an explanandum Y . Its intended domain contains an observational law D and a family I of interventions on the assignment of inputs. The model commits a boundary error for this task if distinctions it excludes make different contributions to the response of Y under admissible interventions, despite its asserted invariance or adequacy. The qualification “for this task” is essential. Missing variables may be irrelevant to one outcome and indispensable to another; interventions admissible in one science may be physically or ethically inaccessible in another. We therefore report a multidimensional sufficiency profile rather than compressing explanation into a numerical score.

The question is not whether complexity in general defeats reduction. In one of our three systems reduction succeeds exactly. Nor is the question whether a macrodescription can predict an aggregate of a fully known microscopic law better than that microscopic law under the same loss and sampling distribution. It cannot: the complete state contains the aggregate. The defensible issue concerns a *restricted* representation or claim. This correction matters because effective-information comparisons

across scale can use different intervention distributions and different outcomes (Hoel et al. 2013; Hoel 2017). Those comparisons illuminate a distinct question and cannot establish that complete microscale knowledge has lost information about its own deterministic image.

The contribution is threefold. First, we formulate an exact fiberwise condition for outcome- and intervention-relative sufficiency, and distinguish it from average predictive adequacy. Second, we separate three estimands that are frequently collapsed: observational optimal regret, interventional optimal regret after refitting, and the transfer regret of the original restricted predictor. Third, we implement a controlled sequence of a sufficient reduction, a scale partition that ceases to be Markov-sufficient, and an agent–environment boundary that hides a contextual interaction. Exact finite-state enumeration makes the positive and negative cases auditable; independent Monte Carlo runs show what a fitted version of the comparison looks like. The systems are intentionally small so that failure cannot be attributed to optimization opacity.

2. Reduction, explanatory aims and prior work

“Reduction” names several claims that should not be inferred from one another. *Ontological reduction* concerns the dependence of higher-level phenomena on lower-level entities and laws. *Explanatory reduction* claims that a lower-level account suffices for a specified phenomenon. *Methodological reduction* isolates components to make inquiry tractable. *Computational reduction* constructs a limited representation by retaining selected variables, states or relations. Coarse-graining maps fine states to aggregated states; abstraction suppresses differences that are unnecessary for a specified purpose. Coarse-graining can therefore be an excellent abstraction. In the present work, “lossy” means failure relative to a declared explanatory commitment, not an intrinsically defective simplification or a denial of microphysical realization.

Marr's distinct computational, algorithmic and implementation levels remain a useful warning against equating one level of modeling with all levels of understanding (Marr 1982). Mechanistic accounts require that model elements be plausibly related to components, activities and organization of the target mechanism (Craver 2007; Kaplan and Craver 2011). Our finite systems have explicit engineered transition mechanisms, but a statistical response law alone cannot satisfy that empirical mapping requirement for a real brain. Potochnik (2017) emphasizes that idealizations serve different scientific aims. We add a formal diagnostic for one particular aim: retaining a specified outcome-response relation across selected manipulations. Woodward's account of intervention and invariance motivates asking which counterfactual changes an explanatory generalization should withstand (Woodward 2003). An invariant predictive response is a necessary condition for an explanation that claims that kind of invariance, not a complete theory of explanation.

Markov lumpability supplies the cleanest control. A partition of a finite Markov chain is strongly lumpable if the probability of entering each next-state block is identical for all source states in a given present-state block (Kemeny and Snell 1960). Strong lumpability ensures that the aggregate is Markov for every initial distribution. The logic is directly relevant to restricted dynamical models: a macrostate can be an exactly sufficient process for a chosen next-macrostate outcome. It is not automatically a mechanistic explanation of every property of the underlying system. Shalizi and Moore (2025) likewise connect the construction of macrostates to observed distinctions and objective dynamics; our treatment fixes a partition and tests it, rather than claiming to discover a unique optimal one.

Causal-abstraction work asks whether relationships between low- and high-level models commute under mapped interventions (Rubenstein et al. 2017; Geiger et al. 2021, 2025). Our criterion is deliberately less ambitious: it addresses a binary outcome and a declared family of input assignments. It can be regarded as an outcome-projected condition of interventional consistency, with an operational loss decomposition. It does not solve the general problem of mapping all interventions between causal models. The response probabilities in the constructed systems are stipulated, not

inferred by causal discovery. Pearl (2009) provides the distinction between observational conditioning and interventions that informs this separation.

Work on causal emergence asks other questions. Hoel, Albantakis and Tononi (2013) and Hoel (2017) formalize comparisons of effective information at different scales. Rosas and colleagues (2020) characterize modes of information that are not reducible to partwise predictive information within their own multivariate framework. These are substantive results, and neither the existence of an information gain in their measure nor its absence in ours resolves all forms of explanatory sufficiency. Dewhurst (2021) cautions against moving without argument from a specific effective-information comparison to stronger metaphysical emergence. Our target is narrower: whether a stipulated restricted boundary supports the outcome and intervention family it claims to support. Under a uniform microstate intervention and an explicitly uniform selection within macrostate fibers, our scale case has *less* macro than micro effective information. We do not seek macro-over-micro emergence; the same case nevertheless exposes a failure of coarse-grained sufficiency under within-block interventions. The two results are compatible because they answer different questions.

An agent–environment case addresses another dimension. Situated approaches stress that cognition can depend on ongoing relationships between agent and environment (Chemero 2009; Clark 2013). This motivates including context in a model's candidate boundary. Our toy input–output mechanism illustrates the logical possibility that an internal-only model performs well in one context distribution but cannot sustain a claim about interventions on the environment. It is not evidence that any particular cognitive theory actually makes that mistake. The point is to specify an empirical burden that a situated explanatory claim would have to meet.

3. Formal framework: a boundary sufficiency profile

Let S be a finite complete input state, $Y \in \{0,1\}$ a next-step outcome, and $B=b(S)$ a retained representation. In a scale model, b aggregates microstates; in a boundary model, b removes an external context variable. A structural response mechanism assigns $p_s = P(Y=1 | do(S=s))$. Each admissible policy i in I supplies a distribution $\pi_i(s)$ over input states; it changes input assignment and leaves p_s invariant. An observational policy $D=\pi_0$ may correlate components of S . Positivity is stated explicitly when conditional comparisons are made: $\pi_i(s) > 0$ for relevant states, or else exact statewise interventions are considered separately. The full-state oracle predicts p_s . The best B -measurable probability under policy i is

$$q_i(b) = \sum_{\{s: b(s)=b\}} \pi_i(s) p_s / \sum_{\{s: b(s)=b\}} \pi_i(s).$$

This is a distribution-weighted aggregate, not an autonomous mechanism merely by virtue of being a conditional probability. The B -only model fitted under observation predicts $q_D(B)$; the B -only model refitted under policy i predicts $q_i(B)$. In real applications both must be estimated; here we first compute them exactly and then fit them from finite samples.

For the task (Y,I) , call B *strongly boundary-sufficient* if $p_s = p_t$ whenever $b(s)=b(t)$ for every pair of states that the intervention family can place in the same retained state. The response then factors through b : there exists h such that $p_s = h(b(s))$. A single response h works across all policies in I , and $q_i(b) = h(b)$ whenever that block is reachable. This is a deliberately demanding, outcome-specific condition. It is weaker than preserving the full causal structure of S , but stronger than having an acceptable average loss under D . If I cannot vary the composition within a block, the observable implication can be weaker: sufficiency is then explicitly relative to the restricted I , and universal fiberwise claims cannot be identified from that family alone.

We use the Bernoulli log loss $L(p,a) = -p \ln a - (1-p) \ln(1-a)$, in natural logarithm units, and the Brier risk $p(1-a)^2 + (1-p)a^2$. Both are strictly proper (Brier 1950; Gneiting and Raftery 2007). Define the *observational optimal regret* $R_D^{\text{obs}} = E_D[L(p_S, q_D(B)) - L(p_S, p_S)]$. Conditional entropy algebra gives

$$R_D^{\text{obs}} = H_D(Y|B) - H_D(Y|S) = I_D(Y;S|B) \geq 0.$$

This is a within-distribution identity for the stipulated response law; it is neither a universal measure of explanation nor a proof that a predictor will transport. Divide nats by $\ln 2$ to obtain bits (Cover and Thomas 2006). The second quantity, *interventional optimal regret*, is $R_i^{\text{refit}} = E_i[L(p_S, q_i(B)) - L(p_S, p_S)] = I_i(Y;S|B)$. The third, *frozen-predictor regret*, is $R_i^{\text{frozen}} = E_i[L(p_S, q_D(B)) - L(p_S, p_S)]$. All are nonnegative with exact probabilities. The chain-rule decomposition is

$$R_i^{\text{frozen}} = R_i^{\text{refit}} + E_{\{B \sim i\}} [KL[\text{Bern}(q_i(B)) \parallel \text{Bern}(q_D(B))]].$$

The first term captures distinctions unrecoverable from B even if the restricted model is refitted to that particular regime. The second captures failure to transport the observationally fitted response. A zero R_i^{refit} in each regime individually need not imply a common invariant B -only function: distinct q_i can give individually perfect but mutually incompatible responses. We therefore also inspect the statewise response diameter $\max_{\{s,t: b(s)=b\}} |p_s - p_t|$, the largest difference between controlled outcomes at the same retained state, and worst-state total variation $|p_s - q_D(b(s))|$. These tests prevent a favorable mixture from hiding opposing effects.

The result is a *profile*, not a scalar index. It records observational log loss and Brier performance, refitted and frozen interventional regret, regime robustness, conditional information relative to the chosen state description, and a separate discussion of mechanistic correspondence. A scientist must also declare a scientifically justified tolerance if approximate adequacy rather than exact factorization is the goal. We do not select a universal numerical threshold. A large conditional information value can be irrelevant to an outcome outside the intended explanandum; a small average regret can be unacceptable if a rare state is central to a clinical or safety-sensitive intervention. This normative choice cannot be outsourced to an information metric.

4. Constructed systems and computational methods

4.1 Design shared by the scale systems

Systems A and B use the same four-state Markov transition family. Write $S_t \in \{0,1,2,3\}$, $M_t = g(S_t) = \lfloor S_t/2 \rfloor$, and $Y = M_{\{t+1\}}$. Let p_s be the probability that the next state belongs to macro-block 1, and distribute the next state uniformly between the two members of its selected block. Thus the transition row is $P_s = [(1-p_s)/2, (1-p_s)/2, p_s/2, p_s/2]$. This complete matrix, rather than an informal simulation recipe, is included in the supplement and machine-readable check file. The observational input distribution gives each macro-block probability one half, with conditional mass 0.95 on microstates 0 and 2 and 0.05 on microstates 1 and 3. It is a one-step experimental input assignment, not a claimed stationary distribution of the chain. This distinction matters: calling a deliberately sampled state distribution stationary would introduce a false dynamical claim.

System A has $p = (0.15, 0.15, 0.85, 0.85)$. The next-block probabilities are identical for every member of each present block. It is strongly lumpable for g , so M alone is sufficient for Y under all state assignments, including uniform input and four point interventions $do(S=s)$. The reduced predictor is $h(0)=0.15$, $h(1)=0.85$. This is the prespecified control in which reduction succeeds.

System B has $p = (0.10, 0.90, 0.90, 0.10)$. Its four states can also be written as (M, R) , with $R = S \bmod 2$; the response depends on their interaction. Under D , $q_D(0)=0.14$ and $q_D(1)=0.86$, because rare within-block states are assigned mass 0.05. We compare the complete microdescription S , equivalently (M, R) , against the coarsened M alone. There is no assertion that a complete micro model fails. The prespecified stress policy assigns probability one quarter to each S , and the four point interventions reveal the within-block contrast. Under stress, $q_i(0)=q_i(1)=0.50$; both blocks mix strongly divergent micro responses. An observationally successful M -only predictor is exposed by the changed composition of the states within its fibers.

4.2 Agent–environment boundary

System C has binary agent state X , binary context C , and next response Y . State s enumerates (X,C) as $(0,0),(0,1),(1,0),(1,1)$. The stipulated input–output mechanism is $P(Y=1|X,C)=0.10$ when $X=C$ and 0.90 otherwise. Under observation X is balanced and C matches X with probability 0.95 ; the joint probabilities are $(0.475,0.025,0.025,0.475)$. The boundary $B=X$ excludes the environment. Its observational response is $q_D(0)=q_D(1)=0.14$. The context-augmented model (X,C) is the full input representation. Interventions $do(C=0)$ and $do(C=1)$ set context while holding $P(X=0)=P(X=1)=0.5$; a policy $do(C=1-X)$ opposes the observed match; a matching-context policy $do(C=X)$ is a failure case for a crude “all shifts harm” slogan. An intervention is represented as a replacement of the context assignment, not as a causal effect estimated from observational correlation. The structural response function is invariant by construction.

4.3 Estimation, robustness and figures

All reported central values are exact expectations computed by enumerating the finite states. A robustness grid varies the rare within-block or mismatching fraction $r \in \{0.01, 0.05, 0.10, 0.25, 0.50\}$ and response contrast $\delta \in \{0, 0.20, 0.40, 0.80\}$, with $p = (0.5 - \delta/2, 0.5 + \delta/2, 0.5 + \delta/2, 0.5 - \delta/2)$ for B and C . The $\delta=0$ row is an explicit counterexample in which no boundary loss is possible. For finite-sample verification we run 30 independently seeded datasets per system (base seeds 1000–1029, with system-specific offsets). Each training set has 5,000 observational inputs; each observational or intervened test set has 20,000 independent outcomes. Groupwise Bernoulli frequencies use Beta($1/2, 1/2$) smoothing; we compare a fitted B -only predictor to a fitted complete-state predictor and report the median and interquartile range of their held-out loss difference. The sample variability is algorithmic Monte Carlo variability, not uncertainty about a population of humans or brains. No significance test is attached to an algebraic identity.

For comparison with the emergence literature, we calculate $I(S_t; S_{t+1})$ under uniform four-state input and $I(M_t; M_{t+1})$ under a macro intervention that selects a microstate uniformly within each macro fiber. We disclose this intervention convention and do not use the numbers as a general measure of intrinsic causal power. Their purpose is to show concretely that our loss criterion and an effective-information-like scale comparison need not have the same sign or philosophical interpretation.

5. Results

5.1 Exact negative control

The control is sufficient in the strongest tested sense. For A , both within-block response differences are zero and the macro transition is invariant to state reallocation. Observational regret, uniform-intervention refitted regret and uniform-intervention frozen regret are each exactly zero, as are Brier regret and worst-state total variation. The full-state and macrostate mutual information under the disclosed uniform intervention are both 0.3902 bits. An additional redundant micro distinction improves neither the target response nor the forecast. This is a substantive counterexample to any thesis that coarsening is automatically explanatorily destructive.

5.2 Loss across an aggregation

For B , the observational oracle risk is $H_{\text{Bern}}(0.1)=0.325083$ nat per outcome, since each microstate has Bernoulli uncertainty 0.1 or 0.9 . The macro predictor's risk is $H_{\text{Bern}}(0.14)=0.404964$ nat; the difference is 0.079881 nat (0.11524 bit). This is much smaller than the maximum divergence of the controlled responses within a block, $|0.9-0.1|=0.8$. Under uniform state input, a newly refitted macro predictor must assign 0.5 to either block and suffers 0.368064 nat excess log loss. Retaining the observational predictors 0.14 and 0.86 instead gives 0.733385 nat excess. Their difference, 0.365321 nat, is the distributional transport term in the KL decomposition. At a microstate for which the

observational prediction is 0.14 but the true probability is 0.9, the worst-state total variation is 0.76. These quantities separately exhibit aggregation loss and failure of a frozen aggregate to transport.

A binary macro Markov transition is not defined invariantly for all initial distributions in B: two source states with the same M have different distributions over Y. A fitted macro transition can be defined for a particular mixture of microstates, but changing that mixture changes the transition. The uniform-fiber information comparison yields 0.5310 bit for the micro transition and 0 for the macro transition, because p averaged within each macro block is 0.5. This is not a “macro beats micro” example. It is an example where an observation-weighted aggregate conceals a micro distinction relevant to controlled future block membership. The condition of lumpability explains the outcome without relying on the magnitude of an information statistic.

5.3 An observationally indistinguishable sufficient twin

The previous comparison leaves an important identification question: could a sufficiently accurate observational study of the retained macrostate alone determine which boundary is sufficient? A derived construction shows that it could not, even with unlimited observations from D. Construct A^* , a second chain with the same four states and observation policy as B, but set its response vector to (0.14, 0.14, 0.86, 0.86), exactly the B aggregate response q_D copied onto each microstate in its corresponding block. The joint observational law of (M,Y) is then identical in A^* and B: $P_D(M=0)=P_D(M=1)=0.5$, with $P_D(Y=1|M)=(0.14, 0.86)$. Yet A^* is strongly lumpable and its M-only response remains correct under every reallocation within the blocks; B is non-lumpable and the same observational predictor incurs 0.733385 nat regret under uniform microstate intervention. The extra A^* construction is an analytic consequence obtained after the prespecified three-case run, not an additional preregistered simulation; the exact verification is saved separately as `matched_twin.csv`.

The construction generalizes. For any response vector p_s and observation law D with positive block masses, define a new response $p^*_s = q_D(b(s))$ while retaining the same assignment of S under D. The two systems induce the same joint law of (B,Y) under D, but the new system is boundary-sufficient and the original may be insufficient if some p_s differ within a reachable fiber. Thus no test of the observational (B,Y) distribution alone identifies strong boundary sufficiency without assumptions constraining the unobserved distinctions. This is an identification statement about restricted data, not an assertion that full observations of S would be uninformative. It clarifies why a robust intervention design or independent mechanistic knowledge is indispensable when an invariant B-only explanation is claimed.

5.4 Context exclusion and policy dependence

C shares B's observational optimal regret of 0.079881 nat, yet the excluded variable is external context rather than a microstate distinction within a Markov partition. Under $do(C=0)$ or $do(C=1)$, the frozen X-only predictor has 0.733385 nat excess log loss. Under the policy that makes context oppose agent state, the true response is 0.9 for both X values while the frozen predictor remains 0.14, giving 1.459501 nat excess. The worst-state total variation under this policy is 0.76. A context-augmented predictor specifies all four responses and remains correct because the mechanism was held fixed.

The refitted X-only model has zero regret under each of those context policies *taken separately*: given an intervention that deterministically fixes C, X alone is enough to predict Y within that regime. This qualification strengthens rather than weakens the conceptual diagnosis. Under $do(C=0)$, the optimal X-only probabilities are (0.1, 0.9); under $do(C=1)$, they are (0.9, 0.1); under the opposing-context policy they are (0.9, 0.9). No single X-only response function is invariant across the specified family. In particular, for fixed $X=0$, the two controlled values of C generate responses 0.1 and 0.9. Hence a claim that X alone is *uniformly sufficient across the intervention family* fails, although one may accurately say it is sufficient after policy-specific refitting. Matching context $do(C=X)$ is a mild counterexample: frozen regret is only 0.007269 nat, not a generic catastrophe for every change of regime.

Case and policy	Observational regret (nat)	Interventional regret, refit (nat)	Interventional regret, frozen (nat)	Worst-state TV
A, uniform states	0.0000	0.0000	0.0000	0.00
B, uniform states	0.0799	0.3681	0.7334	0.76
C, opposing context	0.0799	0.0000	1.4595	0.76

5.5 Robustness and finite-sample checks

The response-contrast grid returns exactly zero observational, refitted and frozen regrets at $\delta=0$ for all five rarity values. As rare states become rarer, the observation-weighted regret can shrink while the consequences of a test policy that reaches them remain substantial. For B at $r=0.01$, $\delta=0.8$, observational regret is 0.01723 nat, while uniform-stress refitted and frozen regrets are 0.36806 and 0.84487 nat. For C at the same settings, observational regret is also 0.01723 nat and opposing-context frozen regret is 1.68941 nat. At $\delta=0$, there is no hidden effect to discover; at a policy confined to an outcome-equivalent subset, the reduction can be sufficient. This pattern locates the source of the loss in a conjunction of response heterogeneity and the reach of the declared intervention family, rather than in complexity as such.

Finite-sample checks follow the exact results. Across 30 seeds, B's fitted restricted-minus-full test log-loss difference has median 0.0799 nat under D (interquartile interval 0.0776–0.0814) and median 0.7308 under uniform stress (0.7149–0.7378). C has median 0.0787 (0.0775–0.0804) under D and 1.4463 (1.4315–1.4735) under opposing context. A's difference under uniform stress is slightly negative, median -0.0016 (-0.0032 to -0.0007): the full model estimates four probabilities, including rare cells, and pays a finite-sample variance penalty even though the macro model is exactly sufficient. This is not a contradiction of the nonnegative *optimal exact* regret theorem, which compares Bayes-optimal predictors rather than two noisy fitted estimators. It is a useful practical reason to retain a sufficient coarse representation when sampling is limited.

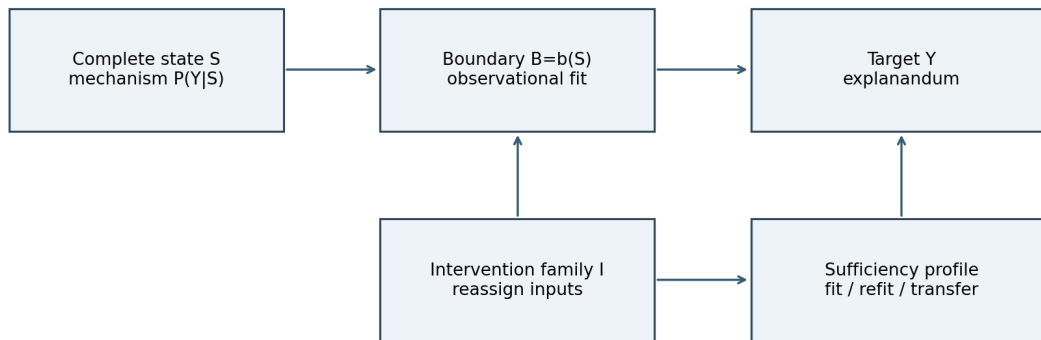


Figure 1. The target Y and intervention family I determine whether a retained boundary B preserves the response law of complete state S . The profile separates observation, regime-specific refitting and transport.

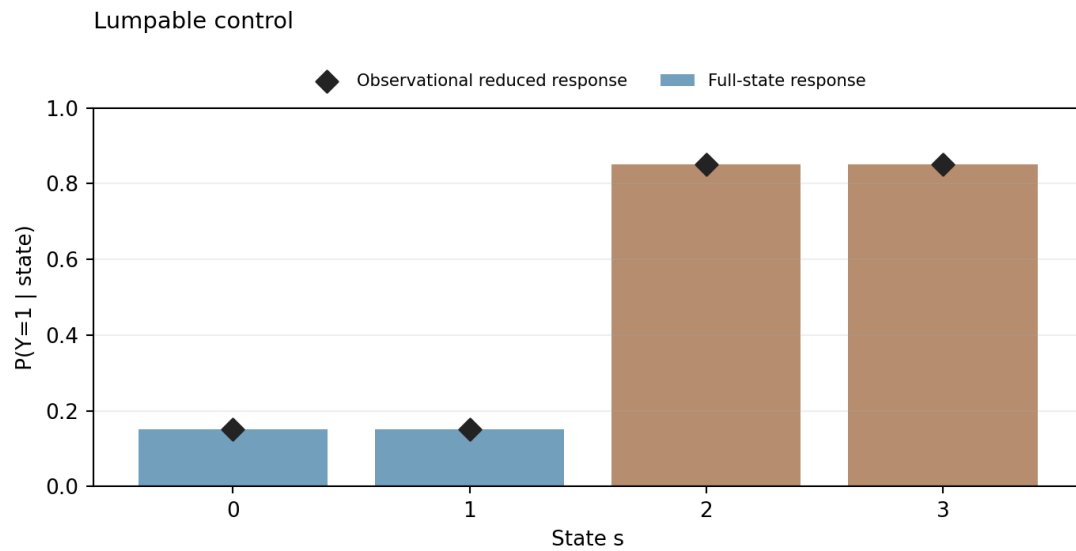


Figure 2. A: the four full-state response probabilities (bars) collapse exactly to two macro responses (diamonds); reassigning microstates within a block has no effect on the next-block target.

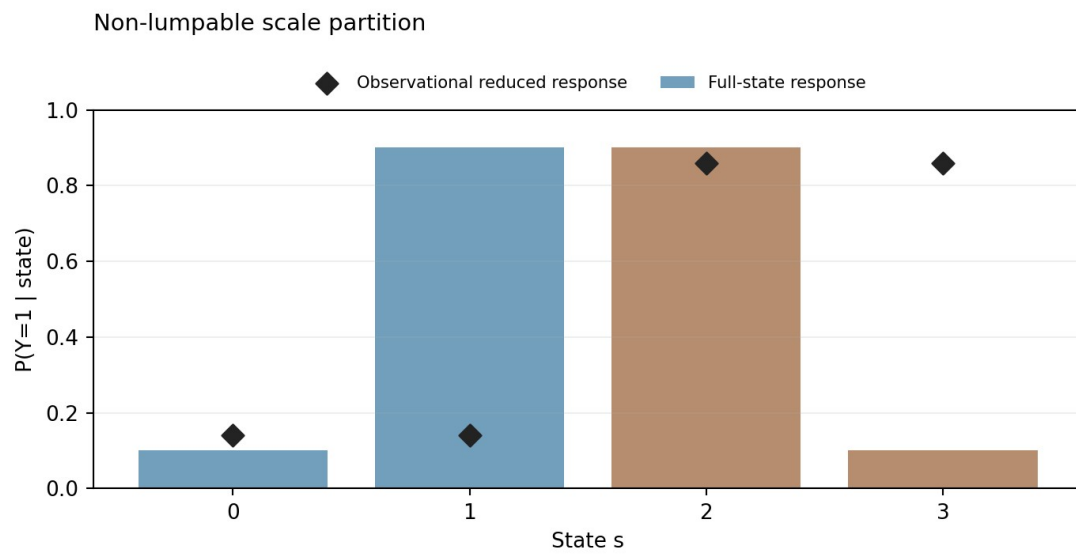


Figure 3. B: microstates in the same macro block have opposing response probabilities. Observation-weighted macro forecasts appear near the responses of common states and fail for rare states made reachable by intervention.

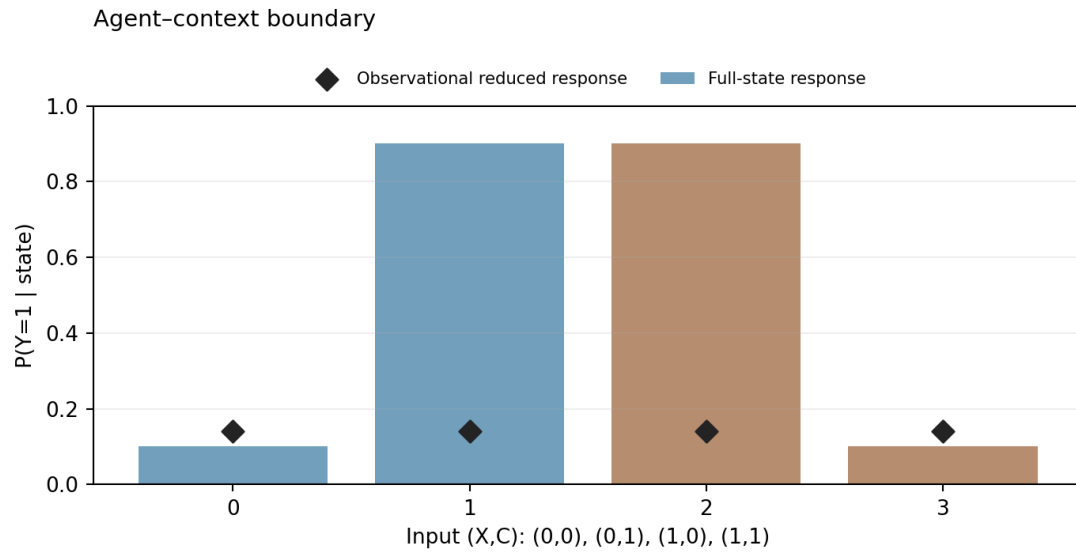


Figure 4. C: context changes the response at either fixed agent state. The observational X-only forecast is close to the common match states but cannot represent both controlled context values.

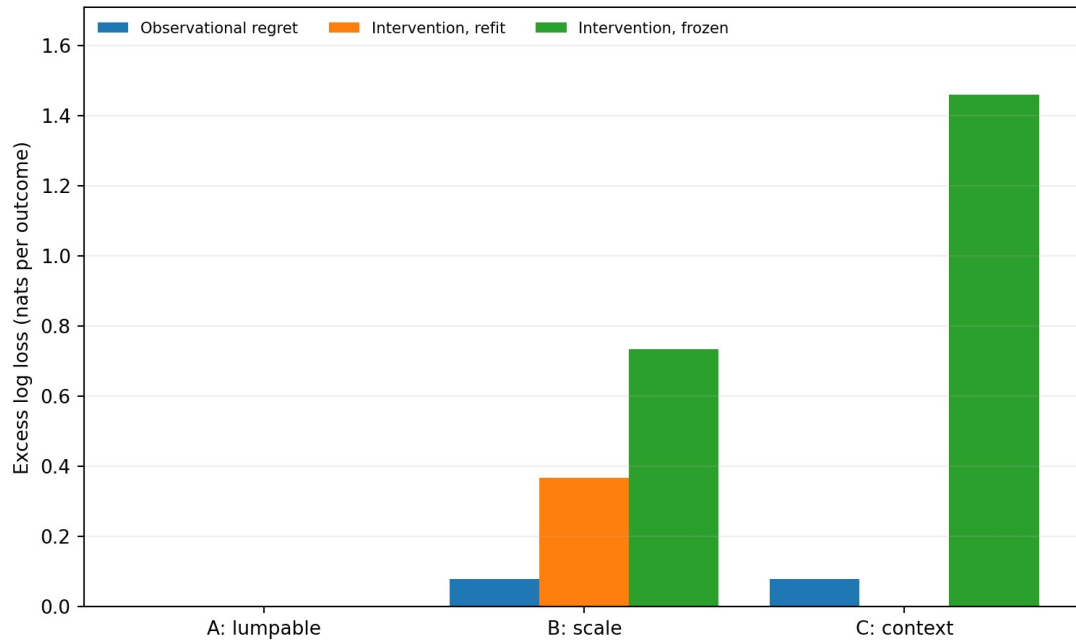


Figure 5. Exact excess log loss for the prespecified control, scale and context cases. The refitted and frozen bars refer to different estimands; the context case has policy-specific refit sufficiency but cross-policy failure.

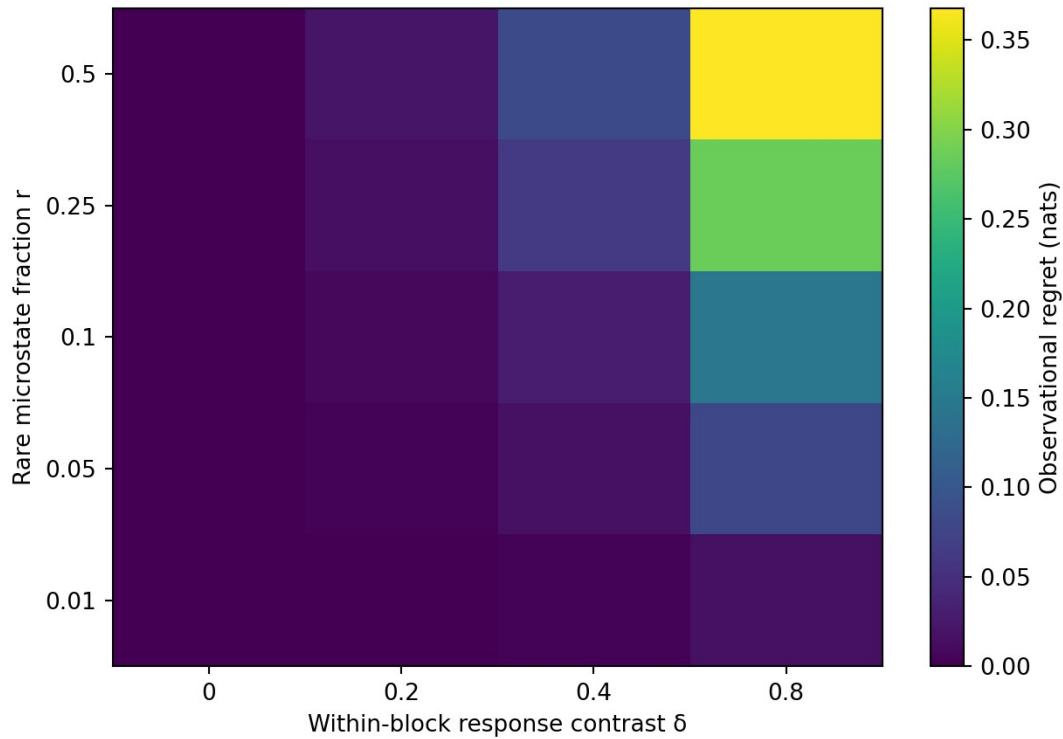


Figure 6. Observational aggregation regret in system B as a function of response contrast and rarity. A small observed regret can coexist with a large controlled-state contrast.

6. What the test establishes, and what it does not

The mathematical lesson is a conditional equivalence claim. If the outcome response is constant on every accessible fiber of the retained representation, the aggregation is exactly sufficient for the declared intervention family. If controlled states sharing one B value have different response laws, a single B-only invariant mechanism cannot represent those controlled contrasts. Given a fixed policy, the optimal aggregate may nevertheless achieve a small mean risk because the contrasting state is rare. Under changed state composition it may suffer intrinsic within-policy loss, transport loss, or both. These are different mechanisms of failure and require different repairs: recover an omitted distinction, restrict the claim to a policy, or recalibrate to a new population if recalibration is scientifically permissible.

The relationship to familiar claims about the micro and macro deserves care. Complete microstate knowledge, with its true transition law, has access to the macrostate as a deterministic function; under a common target, intervention distribution and proper score it cannot be Bayes-inferior to an aggregate. Our B case shows that *moving from complete micro to a coarser state* loses information about next-block outcomes. It does not show that macro causes are ontologically independent of micro causes. The causal-emergence literature can choose interventions and effectiveness measures under which a macrodescription is more informative according to its specified comparison (Hoel 2017). We neither refute nor reproduce that result. Our profile makes the chosen target and policies visible so that such comparisons do not silently change their estimand.

Likewise, conditional mutual information is a sharp characterization of predictive regret under one distribution, not a free-standing measure of explanatory depth. It is zero when the discarded states make no difference under the sampled regime, even when absent support prevents testing other states. Its magnitude depends on the base rate, response noise and assignment distribution. Mechanistic claims require additional evidence about the mapping between formal variables and real components, the manipulability and independence of interventions, and the organization that produces the measured response (Craver 2007; Kaplan and Craver 2011). The profile has a

mechanistic-adequacy dimension precisely because numerical predictive and interventional criteria do not entail that mapping.

The framework is useful for computational cognition when a model asserts generality beyond its training regime. One might define an internal state X from behavioral or neural measurements, a task context C , and a target action Y ; specify interventions on C or on internal components; then ask whether the proposed state representation supports one response law across those interventions. On real systems this requires identification assumptions, intervention feasibility, measurement validity, and a principled choice of target. A formal toy demonstration supplies no evidence that a specific neural area, cognitive process, therapy or diagnostic practice is affected by the kind of boundary error defined here. Nor does it imply that all cognition is essentially distributed. It specifies a way to make a limited claim about explanatory sufficiency vulnerable to a test.

A further distinction concerns temporal depth. A one-step aggregate that is sufficient for Y under an input distribution may fail for multi-step trajectories, especially if a coarse process retains hidden history. Conversely, strong lumpability in A entails a macro Markov process across time for the next-block target and, by iteration, its block-valued trajectories. Our B macro process lacks that uniform guarantee. The passage from one-step loss to arbitrary pathwise prediction requires separately testing the Markov or abstraction properties needed at each horizon. Shalizi and Moore (2025) motivate thinking about macrostates through dynamics rather than static labels; a future extension could examine memory, partial observability and policies that respond to past aggregate states.

7. Limitations and falsifiability

All three systems are constructed examples whose parameters make their conditions transparent. They demonstrate logical possibilities and calibration of the test, not prevalence or effect sizes in cognition. The state spaces are tiny; the response mechanism is known by design; point interventions are available; measurement is perfect; selection and interventions do not alter the response mechanism. In observational cognitive data, each assumption could fail. A policy that changes the environment may also change the agent state, introduce unmeasured pathways, or alter the mechanisms one hoped to hold fixed. The distinction between a “context” and an “internal variable” may itself depend on the timescale and research question. Our formal input assignment deliberately avoids claiming that $\text{do}(S=s)$ is always feasible in living systems.

The paper's principal criterion could be falsified as a useful diagnostic if it systematically flags as errors abstractions that preserve all scientifically relevant responses under feasible interventions, or misses differences that investigators have independently established as explanatory for the intended task. The constructed B case would cease to be a loss case if its within-block probabilities became equal, if the outcome changed to one insensitive to those differences, or if the admissible intervention family could not reach the contrasting states and the explanatory claim were correspondingly restricted. C would cease to challenge an internal-only account if context had no response effect, if the account explicitly promised only the observational regime, or if the augmented variable lacked a defensible causal interpretation in real data. These are potential limits on the claim, not failures to be hidden in favorable averages.

The near equality of observational regrets in B and C is engineered. It illustrates that one number does not identify the causal organization behind a failure; it does not estimate a distribution of boundary errors. Our finite-sample simulations check a straightforward fitted implementation and its variance, but 30 seeds are no substitute for benchmark replication on independently constructed systems. The robustness sweep varies two parameters in the same functional family rather than exhaustively exploring model misspecification, feedback, high-dimensional causal discovery or alternative partitions. Effective-information-like calculations depend on explicitly chosen intervention distributions and remain auxiliary. A full mechanistic test would need independently validated variable mappings and, probably, a richer causal model than a binary response table.

8. Conclusion

A reduction is neither legitimate nor illegitimate merely because it retains one scale or omits an environmental variable. For a specified outcome and intervention family, it is sufficient when the response law factors through the retained boundary; it is lossy when excluded distinctions matter to claims the model undertakes to explain. The lumpable control shows exact sufficiency. The non-lumpable scale case separates modest observational regret from much larger intervention and transport losses. The agent–context case shows that even individually refittable regimes can lack a common invariant internal-only response. This is a tractable formal test of a limited explanatory commitment. Whether a particular cognitive or neural model violates that commitment remains an empirical question requiring appropriate interventions and mechanistic evidence.

Declarations

Data and code availability. All data are synthetic. The accompanying archive contains the complete generator, exact calculations, finite-sample analyses, environment requirements, machine-readable outputs and figures. No external dataset was used.

AI assistance. OpenAI ChatGPT contributed to literature research, synthesis analysis, and editorial review. The analysis is fully detailed in the provided code.

References

- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)
- Chemero, A. (2009). *Radical Embodied Cognitive Science*. MIT Press. <https://doi.org/10.7551/mitpress/8367.001.0001>
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204. <https://doi.org/10.1017/S0140525X12000477>
- Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley. <https://doi.org/10.1002/047174882X>
- Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199299317.001.0001>
- Dewhurst, J. (2021). Causal emergence from effective information: Neither causal nor emergent? *Thought: A Journal of Philosophy*, 10(3), 158–168. <https://doi.org/10.1002/tht3.489>
- Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34. <https://papers.nips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html>
- Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., & Icard, T. (2025). Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83), 1–64. <https://www.jmlr.org/papers/v26/23-0058.html>
- Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378. <https://doi.org/10.1198/016214506000001437>
- Hoel, E. P. (2017). When the map is better than the territory: Measuring causal emergence. *Entropy*, 19(5), 188. <https://doi.org/10.3390/e19050188>

- Hoel, E. P., Albantakis, L., & Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49), 19790–19795. <https://doi.org/10.1073/pnas.1314922110>
- Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, 78(4), 601–627. <https://doi.org/10.1086/661755>
- Kemeny, J. G., & Snell, J. L. (1960). *Finite Markov Chains*. Van Nostrand. <https://link.springer.com/book/9780387901923>
- Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman. <https://mitpress.mit.edu/9780262288989/vision/>
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press. <https://www.cambridge.org/core/books/causality/B0046844FAE10CBF274D4ACBDAEB5F5B>
- Potochnik, A. (2017). *Idealization and the Aims of Science*. University of Chicago Press. <https://press.uchicago.edu/ucp/books/book/chicago/I/bo27128726.html>
- Rosas, F. E., Mediano, P. A. M., Jensen, H. J., Seth, A. K., Barrett, A. B., Carhart-Harris, R. L., & Bor, D. (2020). Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLOS Computational Biology*, 16(12), e1008289. <https://doi.org/10.1371/journal.pcbi.1008289>
- Rubenstein, P. K., Weichwald, S., Bongers, S., Mooij, J. M., Janzing, D., Grosse-Wentrup, M., & Schölkopf, B. (2017). Causal consistency of structural equation models. *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*. <https://www.auai.org/uai2017/proceedings/papers/11.pdf>
- Shalizi, C. R., & Moore, C. (2025). What is a macrostate? Subjective observations and objective dynamics. *Foundations of Physics*, 55, article 2. <https://doi.org/10.1007/s10701-024-00814-1>
- Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press. <https://doi.org/10.1093/0195155270.001.0001>