

The world behind the number

What a model must remember before it can predict.

Shubham Attri · Independent researcher and founder, Phi9
2 October 2026 · Candidate book chapter · edition 1

An exact elementary argument, followed by a proposed experiment. No new general law, empirical diagnostic advantage or novelty of the closure criterion is claimed.

-
1. Two boxes, one number
 2. When tomorrow asks a different question
 3. The equation earns its place
 4. A rule can age; a measurement can fail
 5. What no observer can distinguish
 6. A fair experiment
 7. From a chapter to a scientific result
-

1 / Two boxes, one number

There are ten counters on a table. Some lie in a box on the left; the rest lie in a box on the right. You cannot see inside either box. A display tells you only the total: ten.

For the question “How many counters are there?”, this is an excellent instrument. It gives the answer exactly. A photograph of every counter would contain more information, but none of that extra detail is needed for this particular answer.

Now someone doubles the contents of the left box. What will the display show?

If the boxes began with four and six, the new total is fourteen. If they began with seven and three, it is seventeen. The original display was correct in both cases. Yet the same number, ten, now leads to different answers.

The measurement did not lie. It answered a question that was smaller than the one we eventually asked.

This is a useful place to begin thinking about models. A model does not receive the world itself. It receives measurements, summaries, records or features. Even a very rich measurement leaves something out. The important question is not whether a description is complete in every respect. It is whether the omitted distinctions can matter to the questions and interventions we permit.

The distinction is physical before it is computational. Our intervention acts on one box, while our measurement merges the boxes. No faster calculation of the total can reconstruct a separation the total never recorded. A different measurement, a different intervention, or an assumption about how the boxes were filled could change that conclusion. A larger optimizer alone could not.

SAME OBSERVATION / DIFFERENT SUCCESSOR

$$\text{Left } 4 + \text{Right } 6 = 10 \quad \longrightarrow \quad 8 + 6 = 14$$

double left

$$\text{Left } 7 + \text{Right } 3 = 10 \quad \longrightarrow \quad 14 + 3 = 17$$

Constructed example • deterministic arithmetic • no measured data

Constructed arithmetic example, not experimental measurements.

2 / When tomorrow asks a different question

Write the contents as two nonnegative numbers, u and v . Our instrument produces $z = u + v$. The operation that doubles the left box changes the state to $(2u, v)$, so the next total is

$$z' = 2u + v = z + u.$$

The difficulty is visible in the last expression. We know z , but not u . If u is four, z' is fourteen; if u is seven, z' is seventeen. A function that takes only z and the instruction “double left” cannot return both correct values for the same input.

Compare an operation that doubles both boxes. Then $z' = 2u + 2v = 2z$. For this operation, the total is enough. The same representation can therefore be adequate for one family of actions and inadequate for another. “Enough information” is not an absolute property of a vector. It is a relationship among the representation, the allowed dynamics and the required answers.

We can repair the first example in more than one way. Record u alongside z . Or restrict the actions to ones whose effect depends only on the total. Or impose a known constraint such as $u = v$, provided the dynamics preserve it. Each repair has a different cost and a different domain of validity.

That last qualification matters. An assumption does not become a physical fact because it makes the mathematics convenient. If equal box contents are assumed, an operation that doubles only the left box destroys the equality. A model built on the old constraint must say when its guarantee ends.

A useful summary is thus not merely short. It is short in a way that stays compatible with the evolution we want to study. In mechanics, thermodynamics or learning systems, one repeatedly encounters this tension: a tractable description is valuable precisely because it suppresses detail, but suppressed detail can return as a cause of prediction error.

3 / The equation earns its place

We can now state the exact question without asking the symbols to do the explaining first. Let X contain the possible states, A the allowed actions and $F(x,a)$ the deterministic successor state. Let $q(x)$ be the answer we need, and $\varphi(x)$ the representation we keep. Its possible values form $Z_0 = \varphi(X)$.

We want an answer map h and a represented transition G . They must make these two statements true for every state and allowed action:

$$\begin{aligned}h(\varphi(x)) &= q(x) \\ G(\varphi(x), a) &= \varphi(F(x, a)).\end{aligned}$$

The first equation says that our summary answers today's question. The second says that it can update itself correctly after an action, without looking behind the summary. Together they allow us to answer the same observable after any finite allowed sequence of actions.

Such maps exist exactly when states that receive the same representation also have the same answer and, under each allowed action, the same represented successor:

$$\begin{aligned}\text{For all } x, x' \in X \text{ with } \varphi(x) = \varphi(x') : \\ q(x) &= q(x'), \text{ and} \\ \varphi(F(x, a)) &= \varphi(F(x', a)) \text{ for every } a \in A.\end{aligned}$$

Here is the proof. If h and G exist, identical represented inputs give identical outputs. Both conditions are necessary. In the other direction, take any represented value z and choose a state x that represents it. Define $h(z) = q(x)$ and $G(z, a) = \varphi(F(x, a))$. Choosing a different state with the same representation cannot change either result, by the two conditions. The maps are therefore well-defined. Apply this argument at each step and induction gives the represented trajectory and required observable for every finite action sequence.

This proof is short because the statement is exact and limited. It is an elementary quotient or closure argument, closely related to established abstraction and behavioral-equivalence ideas. It is not offered as a new theorem. Nor does it tell us how to discover a useful φ cheaply, learn G from finite observations, quantify approximate error or survive an unannounced change in F .

It is also stronger than merely preserving one task's outputs over time. A representation could contain a redundant coordinate that fails to update consistently while another coordinate still answers the task perfectly. Our criterion concerns exact closure of the entire specified representation. Failure of this condition does not prove that every task-specific predictor, richer history or alternative representation must fail.

History can help, but only if it contains a distinguishing signal. If previous actions and totals reveal u , then a stateful observer can keep that information even though the instantaneous total cannot. We must name whether a predictor receives one measurement or a history; otherwise an apparent comparison of models is actually a comparison of information.

Stochastic dynamics require another statement. A common sufficient construction asks equal represented states to have equal task outputs and equal distributions of next represented states for every action. Even that construction must specify the observation and reward process and the policy class. The deterministic result above should not be silently promoted into a universal stochastic claim.

4 / A rule can age; a measurement can fail

Suppose a predictor used to work and now misses. The simplest response is to fit it again. Sometimes that is precisely right. If both boxes used to double and now both triple, the total still determines the next total. The old rule $z' = 2z$ is stale; the new rule is $z' = 3z$. The measurement remains sufficient.

But suppose the operation changes from doubling both boxes to doubling only the left. The old rule is again wrong. This time no replacement rule of the form $z' = g(z, a)$ can be correct for all arrangements. We have changed not just the coefficients but the information required for prediction.

These are two different failures: a fitted mapping that no longer matches the dynamics, and a representation that no longer supports the required mapping. They suggest different responses—refit the rule or acquire a distinguishing observation. In practice both failures may occur together, and noise, optimization failure or limited samples introduce further explanations. The clean pair is a controlled starting point, not a taxonomy of every real modeling error.

Even missing information does not make useful statistical prediction impossible. Under a known distribution of u given z , we could predict an average next total. Its unavoidable uncertainty would depend on that distribution. The impossibility here is exact prediction across all states in a merged class, not the impossibility of estimating a conditional mean.

The research question at Phi9 is whether a diagnostic procedure can choose a useful response under a limited sensing and computation budget. Does an extra observation repay its cost? When is refitting enough? When should the procedure admit that the available evidence does not select one explanation?

Those are hypotheses to test. Calling the procedure a selector does not demonstrate that it selects well. A selector can waste its budget, overfit its diagnostic signals or simply reproduce a cheaper periodic policy.

5 / What no observer can distinguish

Imagine two hidden explanations, C_0 and C_1 . A diagnostic agent sees a history H : observations, actions, rewards if applicable, and any permitted probe outcomes. Let π denote a policy for choosing actions and probes using the available history. If every allowed π induces the same distribution of visible histories under the two explanations, the explanations are observationally indistinguishable to this agent.

$$P(H \mid C_0, \pi) = P(H \mid C_1, \pi)$$

for every permitted probing policy π .

The agent's answer is a function of H and perhaps its own random seed. Equal input distributions give equal output distributions. Under equal prior probabilities for the two explanations, no such binary test can achieve expected classification accuracy above one half. With unequal priors, always choosing the more likely explanation can do better than one half, but this is not evidence that the observation distinguishes the causes.

For a fixed policy and a finite history window, a sharper finite-sample formulation uses total variation distance between the two history distributions. With equal priors, the best binary accuracy is $(1 + TV)/2$. Exact equality gives $TV = 0$. Small but nonzero separation permits some advantage; it does not promise that our chosen algorithm can realize it from the available samples. This is a standard hypothesis-testing relation, not a new Phi9 result.

The way out is not rhetorical confidence. It is additional structure: a probe with different consequences, a useful prior assumption, a longer informative history or another measurement. Each must be stated. If we read the left box, we have changed the information available, not refuted the earlier impossibility.

That is why probe parity belongs at the center of the experiment. If a diagnostic method receives a revealing sensor while a refit-only comparator is forbidden to use it, a win may only show that more information helps. It does not establish the value of diagnosis. The strongest comparison gives competing adaptive methods the same opportunities and charges them for what they actually use.

An indistinguishable case is therefore a valuable negative control. A method that fabricates certainty there has exposed a defect. An honest system may still act—perhaps choose a conservative intervention—but its causal label should remain uncertain.

6 / A fair experiment

Begin with small systems whose hidden states and transitions can be audited by the evaluator. Keep those hidden states out of the learner's input. Start from matched pre-change predictive performance, then draw changes from a held-out collection rather than choosing the cases after seeing which ones favor the selector.

Use three conditions. In the stale-fit condition, the retained observation still supports prediction but the fitted rule is wrong. In the information-loss condition, states merged by the observation require different future answers. In the negative-control condition, competing hidden explanations have identical permitted visible-history laws. A separate probe-enabled condition should state exactly which action creates observable separation.

The box example supplies an intuition and a deterministic check, but it is not yet a fully specified stochastic benchmark. For example, changing from $2z$ to $3z$ is not a matched hard diagnostic case for changing to $z + u$. A careful benchmark must specify reset distributions, noise, persistence of hidden variables, change times and probe dynamics. Matching initial error alone does not match post-change identifiability.

Compare refit-only, fixed-schedule sensing, random sensing, acquire-then-refit and an adaptive budgeted selector. Include a simple uncertainty or residual-threshold policy as a strong inexpensive competitor. Use the same action space, history window, resets, change detector and total budget wherever the comparison permits. Keep a full-state learner as a labeled oracle, not an ordinary baseline that secretly sees more.

Record prediction error after change, recovery delay, probe count, diagnostic mistakes and calibration. Charge sensing, fitting, selector evaluation and inference. Report curves across budgets and cost assumptions rather than hiding all tradeoffs in one unexplained weighted sum. If the selector is slower but more accurate, the reader should be able to see both facts.

Predefine a primary comparison and an evaluation horizon. Use independent test regimes and seeds; select thresholds and models on training or validation regimes only. Report variability and unsuccessful cases. Include ablations that remove probing, remove diagnosis or replace adaptation with a fixed schedule. An implementation test only verifies mechanics; it is not a scientific replication.

Abstention needs its own accounting. A method that abstains on every case might be well calibrated but useless. Measure coverage alongside conditional diagnostic accuracy, and measure the task cost of its fallback action. In the exact indistinguishable condition, evaluate whether the system avoids unjustified confidence, not whether it guessed the hidden label by luck.

The practical hypothesis survives only if the selector improves the stated error-cost or recovery tradeoff against the strongest matched baseline on held-out regimes. It fails if benefits disappear after including overhead, matching probe access or controlling for an easier change distribution. A negative result would still tell us which proposed distinction did not buy enough useful information.

7 / From a chapter to a scientific result

At this stage we have a teaching example, an exact limited proposition and an experiment proposal. We do not have a demonstrated diagnostic advantage or a general physics of learning. A research program may ask broad questions; a paper must state a claim that its evidence can carry.

Related work already studies abstraction across tasks, compression-performance tradeoffs and task-specific memory. Abel and colleagues' lifelong abstraction work considers a specified distribution of tasks; it should not be described as ignoring change. Booker and Majumdar examine reducing memory for robot control tasks. V-JEPA evaluates learned representations on downstream tasks. These are useful places to look for constraints on our proposal, not labels to attach to an untested method.

Our earlier single-sample latent compression comparison belongs in a separate evidence record. It does not establish the representation's downstream sufficiency. Our artifact pipeline's engineering tests belong to a software record, not a scientific-results section. Keeping these distinctions visible is part of writing clearly.

A book chapter can lead the reader through discovery. A blog can make the question memorable. A paper must add something inspectable: a new supported argument, a carefully measured finding, a benchmark that reveals a real failure or an experimentally justified method. The same prose should not be relabeled as all four accomplishments.

For general physics modeling, the small lesson is demanding enough: say what the system is, what can be measured, which interventions are allowed and which statements the evidence supports. Only then ask how learning changes the description.

Before asking whether a model can learn the answer, ask whether its measurements leave the answer available.

Reading paths

[Technical statement and paper scaffold](#) · [Short companion essay](#) · [Evidence and source ledger](#)