

The world a model must keep

Representations · motion · learning · evidence

Shubham Attri

Independent researcher and founder, Phi9

Consolidated research edition 1

2 October 2026

This is a complete current reading edition, not a claim that the research programme is finished. It brings the existing Phi9 writings into one argument: a model should keep the distinctions its decisions need, and its evidence should say when that promise fails. Classical mathematics, constructed examples, retained experiments and future applications are identified separately. Nothing in this edition establishes a universal learner, a new general law, clinical readiness or a peer-reviewed result.

Contents

Before we begin	3
1 The world behind the number	4
2 Coordinates and operations	6
3 A contract for the future	8
4 Learning a rule is not inventing information	11
5 The state, the motion and the readout	13
6 Fourier is a coordinate choice, not a fundamental law	15
7 Memory, context and belief	18
8 Programs, proofs and roundoff	21
9 Learning interaction rules	24
10 A map of the model's own mistakes	27
11 The price of observing more	30
12 Refit the rule, or ask another question?	33
13 Calibration and the price of choosing	36
14 From representations to useful agents	39
Appendix A The technical contract	42
Appendix B Evidence ledger, not a combined benchmark	45
Appendix C Terms that should stay distinct	48
Appendix D Sources, coverage and editorial method	49

Before we begin

Before consolidation, there was one explicitly numbered candidate book chapter, The world behind the number. The state-motion-readout companion had eleven sections; the foundation notebook had ten numbered teaching units. Those were not eleven or ten completed book chapters. The remaining artifacts were essays, technical studies, empirical reports and supplements. PDF page counts are not chapter counts.

This edition defines fourteen chapters editorially. It merges repeated explanations, restores the negative results beside the positive ones, and introduces a small amount of new teaching material connecting them. Those new explanations are not new empirical evidence. The technical appendices retain the proof boundary, experiment ledger and references. The original files remain unchanged.

Our recurring picture is two boxes. Our recurring question is more demanding: after a measurement compresses the world, can the model still predict the consequences of what we do? We will move from counters to vectors, from vectors to motion, and from motion to decisions. Along the way, the experiments repeatedly correct our first intuition.

The diagrams in this book are original explanatory drawings. Schematic arrows are not measurements. Charts identified as retained results redraw reported values without running the experiments again. Numerical comparisons belong to their stated protocols, not to a common league table. References in brackets point to the source guide at the back.

A reader can follow the examples before the equations. A researcher can inspect the equations before accepting the examples. Both should arrive at the same boundary: information, computation and evidence are different things.

1 | The world behind the number

Two boxes, one excellent measurement There are ten counters on a table. Some are in the left box and the rest in the right. A display shows their total. For the question “how many?”, this is an excellent instrument. It answers exactly without showing every counter.

Now double the contents of the left box. If we began with four and six, the new total is fourteen. If we began with seven and three, it is seventeen. The original display was right in both worlds. It simply answered a smaller question than the one we eventually asked.

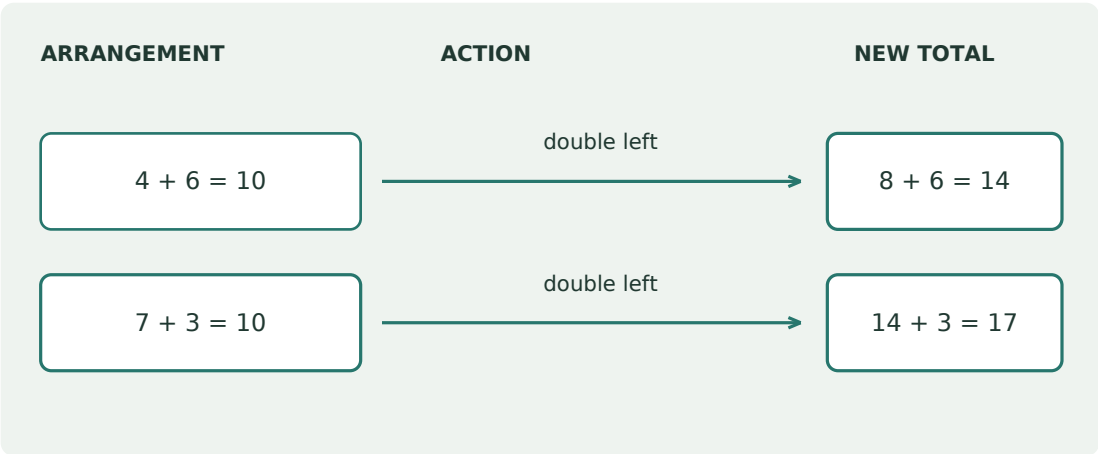


Figure 1. Constructed example. The same total conceals two arrangements; the same left-only intervention produces different successors.

Write the hidden arrangement as $x = (u,v)$. Keep only $z = u + v$. After the intervention, the arrangement becomes $(2u,v)$. Its total is:

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

Knowing z does not tell us u . A deterministic function receiving only the number ten and the instruction “double left” cannot produce both fourteen and seventeen. Faster computation of ten cannot recover a separation that ten never recorded. This is an information obstruction, not a shortage of optimization steps.

Enough is relative to an action Change the instruction to “double both”. Now the new total is $2u + 2v = 2z$. The same display that failed for one action succeeds for another. A representation is not sufficient merely because it is large, small or accurate on yesterday’s task. It is sufficient relative to the outputs and actions we permit.

We can repair the left-only example by storing u alongside z . We can instead restrict the actions to ones whose effects depend only on z . Or we can impose a known constraint, such as $u = v$, while checking that the permitted actions preserve it. These repairs are not interchangeable: they change the information, the task or the assumptions.

If equal box contents are assumed, doubling only the left box destroys that equality. An assumption does not become a physical fact because it makes the algebra convenient. The model must know where its guarantee ends.

Prediction need not be exact to be useful Suppose the possible left counts are equally likely to be four or seven. Conditional on total ten, the next total is fourteen or seventeen. Its mean is 15.5. An average prediction can be useful, even though exact prediction across both arrangements is impossible. The uncertainty is part of the answer, not an embarrassment to remove.

Try the distinction yourself. Does total ten determine whether the left box is empty? No: (0,10) and (5,5) give the same display and different answers. Does it determine the result of doubling both? Yes. The important thing is the collision: two worlds have been merged, and the requested operation may separate them again.

The first job of a model is therefore not to remember everything. It is to state what it promises to remember enough of. This example is constructed arithmetic, not a measured experiment. Its role is to make the later contract visible. [P1, P2]

2 | Coordinates and operations

A vector keeps distinctions separate A scalar is one number. A vector is an ordered collection of coordinates with a meaning attached to each. For our boxes, $x = (3,7)$ means three on the left and seven on the right. It is not the same physical arrangement as $(7,3)$, even though both have total ten.

Adding $(1,0)$ gives $(4,7)$: one counter has been added to the left. Adding $(1,-1)$ transfers one from right to left, provided there is one to transfer. The physical states are nonnegative integers. Extending the formulas to real vectors helps us reason algebraically; it does not turn fractional or negative counters into physical objects.

$$P = [1 \ 1] \quad z = Px = u + v$$

The row matrix P describes what we retain. Its kernel contains all differences of the form $(t, -t)$. These changes move counters between boxes while leaving the display unchanged. A kernel is an invisible direction, not necessarily a legal physical state.

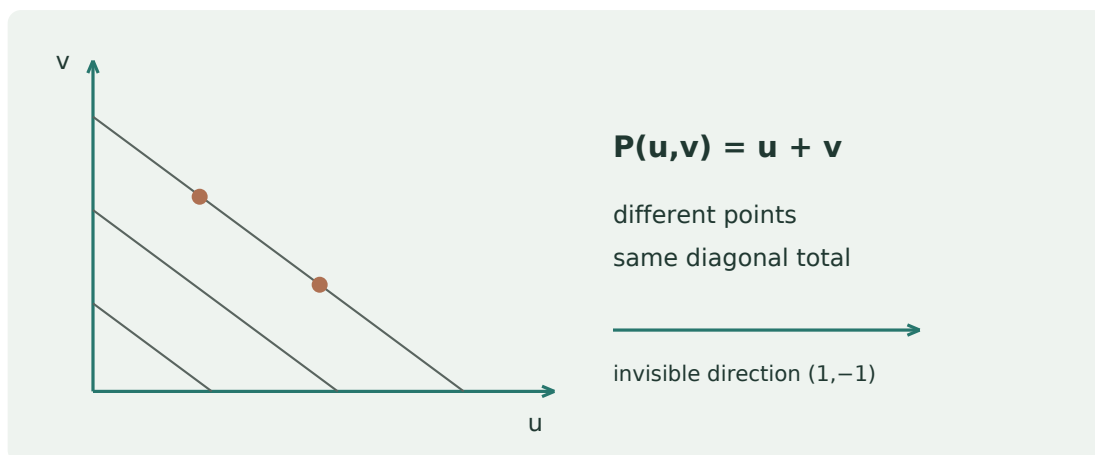


Figure 2. Schematic state space. Each diagonal is a fibre of the total map. Motion along a fibre changes the hidden arrangement, not the total.

A matrix is a linear transformation. Doubling both boxes uses $A = \text{diag}(2,2)$; doubling only the left uses $B = \text{diag}(2,1)$. Swapping uses $S(u,v) = (v,u)$. Apply B after S to (3,7) and obtain (14,3). Apply S after B and obtain (7,6). Composition has an order.

What multiplication means The expression ABx applies B first and A second. Associativity lets us regroup a fixed exact product, $(AB)C = A(BC)$. It does not allow us to reorder its factors. A map adding a counter is affine rather than linear: $T(x) = Ax + b$, with the offset b recording the addition.

The dimension of a representation is the number of its coordinates, not the number of vectors appearing in an equation. A two-dimensional state can have one current state vector, one velocity vector and many candidate actions. Listing more arrows does not change the dimension of the state space.

For motion in a constrained world, ambient dimension and usable freedom may differ. A point moving around a circle occupies a two-coordinate plane, but locally has one degree of freedom along the circle. A robot's joints, object pose and contacts introduce different coordinates and constraints. We must name those meanings before counting them.

A worked compressed update If A doubles both boxes, $PA = [2 \ 2] = 2P$. Thus we can update the total by multiplying it by two. For B, $PB = [2 \ 1]$. There is no scalar g such that $gP = [2 \ 1]$, because gP has equal entries. A compact rule exists in the first case and fails in the second.

Try an affine version: double both boxes and add one to the left. The total obeys $z' = 2z + 1$. The offset did not disappear during compression; it was mapped into the retained coordinate. This elementary linear example is the doorway to a much more general closure condition. [P3, P4]

3 | A contract for the future

Two routes should meet Imagine two ways to predict tomorrow's display. One updates the full world and then measures it. The other measures today's world and updates only the measurement. We want them to agree.

Let X be the admitted states, A the admitted actions, $F(x,a)$ the deterministic successor, $q(x)$ the answer required and $\varphi(x)$ the representation retained. We seek a readout h and a represented update G , defined on the image $\varphi(X)$:

$$h(\varphi(x)) = q(x)$$

$$G(\varphi(x),a) = \varphi(F(x,a))$$



Figure 3. The closure square. Updating then representing must agree with representing then updating, on every admitted state and action.

The first equation preserves the answer. The second closes the represented dynamics: the summary can update itself without looking behind itself. Together they preserve the requested output after every finite admitted action word, provided the assumptions hold at each visited state. For actions chosen during execution, the policy and action-legality conditions must also preserve the contract.

The collision test Such maps exist precisely when equal representations imply equal answers and equal represented successors under each action. Necessity is immediate: a function applied to identical inputs has identical outputs. For sufficiency, choose any state representing z and define $h(z) = q(x)$, $G(z,a) = \varphi(F(x,a))$. The collision conditions make those definitions independent of the state chosen. An induction extends the one-step relation along any finite action sequence.

This is a classical elementary quotient argument, closely related to automata and state abstraction. It is not our new theorem. Set-theoretic existence does not supply an efficient algorithm, a learnable representation, a finite state set or behavior on values outside $\varphi(X)$. [R1, R4]

Closure is stronger than task-only prediction Suppose the required answer is always zero. A representation also carries an irrelevant label. Two states share label 0, but their successors have labels 1 and 2. The answer remains zero forever; it is perfectly predictable. The entire chosen representation nevertheless fails to close, because $G(0)$ cannot be both 1 and 2.

That example matters when interpreting a failed representation test. Failure of whole-state closure does not prove that every predictor of one task must fail. A coarser representation may remove the inconsistent irrelevant label. A richer history may preserve a needed distinction. A different task may never ask about it.

The action chooser can hide information too Even if each fixed action closes, a policy using hidden information may break the closed loop. Let $\varphi(x,h) = x$. “Advance” increments x ; “stay” leaves it unchanged. Both close individually. A scheduler advances when hidden h is zero and stays otherwise. States $(0,0)$ and $(0,1)$ then have equal representations and different represented successors.

A sufficient safeguard is for the policy to depend only on $\varphi(x)$. It is not always necessary: different actions can have identical represented effects. The essential check is the resulting closed-loop transition. Action legality, time, scheduler state and changing context must be included or explicitly preserved when they matter. [P4, R4]

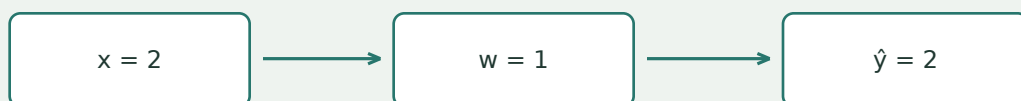
4 | Learning a rule is not inventing information

What a fit changes We supplied the rule “double both”. A learner can instead estimate the relation from examples. With observed totals z and successors z' , fit $z' \approx gz$ by minimizing squared error. Differentiating the sum of squared residuals gives:

$$g = \Sigma(zz') / \Sigma(z^2), \text{ when } \Sigma(z^2) > 0$$

In the executed foundation toy, 66 exact training pairs gave $g = 2$. On 55 withheld starting states, the maximum exact error was zero. These are new inputs under the same constructed rule, not independent environments or evidence of general intelligence. Finite testing and an all-state algebraic argument have different roles. [P3]

Follow one gradient all the way Consider a deliberately small new teaching example, not a new training run. A predictor uses $\hat{y} = wx$. On $x = 2$, desired $y = 6$, current $w = 1$, it predicts two. Set $L = (\hat{y} - y)^2/2$. The error is -4 and the derivative is $dL/dw = (wx - y)x = -8$. A step of size 0.1 gives $w' = 1 - 0.1(-8) = 1.8$. The new prediction is 3.6 and the new loss is 2.88, down from eight.



target 6 → error -4 → derivative -8

step 0.1 → $w' = 1.8$ → $\hat{y}' = 3.6$

L: 8 → 2.88

Figure 4. Worked arithmetic, not a new training run. A derivative tells how loss responds locally to the parameter; the chosen step changes the prediction.

The arithmetic shows why the gradient points toward a local reduction in this example. It does not prove that arbitrary large steps improve every loss, that a neural network finds a global optimum, or that one example determines the right rule.

For a chain of operations, backpropagation applies the chain rule. If $h = w_1x$ and $\hat{y} = w_2h$, then $dL/dw_1 = (\hat{y} - y)w_2x$ and $dL/dw_2 = (\hat{y} - y)h$. The forward pass computes predictions; the backward pass distributes derivatives; an optimizer changes parameters. These are distinct stages.

Expressiveness is not a guarantee A network stacks parameterized maps and nonlinearities. Without nonlinearities, several linear layers compose into one linear map. Nonlinearities permit richer functions, but the information available at the input still limits the output. If two hidden arrangements produce the identical retained total, a deterministic network seeing only that total cannot separate them.

Training changes a fitted function. It may also change a learned encoder. Neither operation makes discarded distinctions recoverable without assumptions or additional observations. Approximation results about network families do not say that finite data, an optimizer and a deployment distribution will deliver every desired function.

Keep evaluation out of the fit We learn on training data, choose settings using a declared validation procedure, and measure generalization on held-out data. After repeated choices, validation performance itself can become optimistic. Later chapters show this in a retained experiment.

A learned model has a domain, a target, a loss and a resource budget. A beautiful explanation should make those visible rather than describing “learning” as a single magic verb. The useful question is what changed: weights, features, memory, available sensing or the evaluator’s rules? [P5, R17]

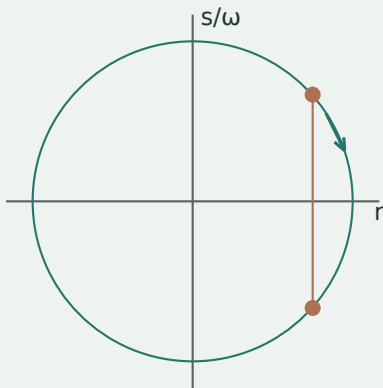
5 | The state, the motion and the readout

A point is not a trajectory A state $x(t)$ is where the system is at one instant. A trajectory is the path $t \mapsto x(t)$. A velocity is its instantaneous direction and rate of change. The equation $dx/dt = b(x,a)$ specifies a vector field, possibly depending on action a . It is not itself a decoder or a learned representation.

For an oscillator, let r be position and s velocity. With frequency $\omega > 0$:

$$dr/dt = s \quad ds/dt = -\omega^2 r$$

Starting at $r(0) = 1, s(0) = 0$ gives $r(t) = \cos(\omega t), s(t) = -\omega \sin(\omega t)$. The phase portrait is an ellipse in (r,s) coordinates and a circle in $(r,s/\omega)$ coordinates. Reading position only merges moments with different velocities. The same position can be moving left or right.



same r , opposite velocity

$$r(t) = \cos(\omega t)$$

$$s(t) = -\omega \sin(\omega t)$$

readout: position only

Figure 5. Analytic schematic, not a rerun. Position alone merges opposite directions of motion. The scaled state $(r,s/\omega)$ follows a circle.

The decoder answers a specified question An encoder maps an observation or state to z . A decoder maps z to a requested output. In a generative model it may reconstruct pixels or parameters of a distribution; in our closure notation h reads the task output. “Decoder” in a Transformer also names an architectural stack. These meanings are related by data flow, not identical definitions.

A readout producing an image is not the transition law moving the state. A low reconstruction error does not alone establish that the latent state predicts actions and futures. Conversely, a good task predictor may deliberately lose visual details that the task does not need. VAE, Transformer, Dreamer and Neural ODE papers provide different examples of these roles. [R19–R23]

Skipped states can mean different things Sampling records points at chosen times; the continuous trajectory can pass through intermediate states not recorded. A numerical step approximates motion between samples. A projection can merge different points. A physically unreachable point cannot be visited under the admitted dynamics. These are four different reasons for a picture to appear to “skip” something.

The retained oscillator demonstration ran Euler and RK4 for ten periods at three step sizes, comparing six trajectories with an analytic solution. Three deterministic assertions passed. At the middle resolution, final scaled-state error was approximately 3.67 for Euler and 3.04×10^{-6} for RK4. This is a numerical-method demonstration, not a learned world model. [P6]

Explicit Euler increases oscillator energy by a factor $1 + (\omega\Delta t)^2$ per step in exact arithmetic. Smaller steps reduce that error; they do not turn Euler into the exact flow. The earlier demo checked this identity to roundoff. No trajectories were recomputed for this book.

The important division is now visible: state describes what is, motion describes how it changes, and readout describes what we ask to see. Mixing those jobs produces impressive-looking diagrams and imprecise claims.

6 | Fourier is a coordinate choice, not a fundamental law

A wave can become a rotating number Fourier analysis expresses suitable functions in oscillatory components. A cosine and a sine of the same frequency can be combined into one complex phase. For the oscillator $\dot{r} = s$, $\dot{s} = -\omega^2 r$ with constant $\omega > 0$, define $c = r - is/\omega$. Then $dc/dt = i\omega c$. Instead of following two coupled real coordinates, we follow one complex rotation.

This is an elegant representation of an already specified dynamics. The Fourier transform did not create the oscillator's physical law. It made that law easier to express.

$$\hat{f}(\omega) = \int f(t)e^{-i\omega t} dt$$

This convention uses angular frequency; inverse reconstruction has a $1/(2\pi)$ factor. Transform and inverse statements require an appropriate function space and convergence interpretation. Under suitable differentiability and decay, transformation sends f' to $i\omega f$. Periodic versions use Fourier series and periodic boundary conditions.

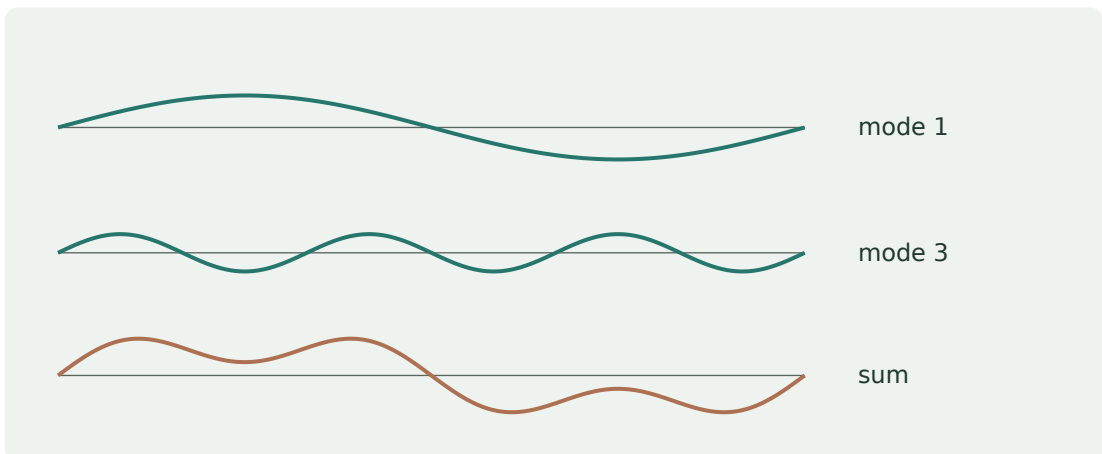


Figure 6. Schematic superposition. Components add to a signal. Retaining all coefficients can change coordinates; dropping a component changes the retained information.

What a finite transform preserves For N samples, the full discrete Fourier transform is an invertible linear change of coordinates when all N coefficients and the inverse convention are retained. Keeping only a few frequencies is compression and may lose information. A finite Fourier reconstruction is continuous as a function of time; that does not prove the sampled source was continuous or uniquely recover its behavior between samples.

Band-limited reconstruction needs sampling and signal assumptions. A sharp event, unobserved contact or missing sensor distinction does not become known because the retained numbers have been written as waves. Fourier cannot invert the information-losing sum of our two boxes unless other assumptions supply the missing separation.

Nonlinearity couples modes If a motion includes a cubic term, cubing a cosine produces both its original frequency and a third harmonic:

$$\cos^3\theta = (3\cos\theta + \cos 3\theta)/4$$

A basis that diagonalizes one linear operator need not diagonalize a nonlinear dynamics. Modes can interact. Multiplication in one domain becomes convolution in the other under the relevant transform conventions. Oscillatory coordinates are a candidate tool, not a certificate that every representation becomes independent continuous motion.

Koopman moves observables For a discrete deterministic transition F , define $Kg = g \circ F$ on a suitable invariant space of observable functions. This operator is linear in g even when F is nonlinear. That trades nonlinear state dynamics for potentially infinite-dimensional linear observable dynamics. A finite chosen dictionary closes exactly only if its span is invariant. A fitted matrix can approximate its evolution without proving invariance. [R24]

For continuous-time dynamics, the corresponding generator acts as $Lg = \nabla g \cdot b$. The following example closes under that generator. A worked special case is $\dot{x} = \mu x$, $\dot{y} = \lambda(y - x^2)$. Retain $\psi = (x, y, x^2)$. Differentiation gives $\dot{x} = \mu x$, $\dot{y} = \lambda y - \lambda x^2$ and $(x^2) = 2\mu x^2$. This particular three-coordinate lifted system is linear and exact. It adds a coordinate; it is not automatically compression. For $\dot{x} = x^2$ on an unrestricted real domain, the dictionary $(x, \dots, x^n)$, $n \geq 1$, generates nx^{n+1} when its highest-degree member is differentiated and fails this closure test. That does not rule out every other transform.

The research question is which coordinates simplify the admitted dynamics while preserving the desired output at acceptable cost. It is not whether Fourier is the equation behind every intelligence.

7 | Memory, context and belief

History can reveal what a snapshot hides A total alone may be ambiguous. If previous actions and measurements reveal the left count, a stateful observer can retain it. History helps only when it contains a distinguishing signal. Repeating an uninformative measurement forever does not necessarily reveal the hidden arrangement.

A fair comparison must say whether a predictor sees one observation, a fixed window or the full history. Giving one method memory and denying it to another is a comparison of information access as well as model architecture.

Under partial observation, a belief state records possible worlds and their probabilities. In a controlled hidden-state model, one schematic Bayesian update is:

$$b'(x') \propto p(o' | x') \sum_x p(x' | x, a) b(x)$$

It predicts possible successor states and reweights them by the new observation. Continuous states replace the sum with an integral. Normalization, the transition law and the observation likelihood must be specified. A practical approximation may be much smaller than this full distribution, without carrying an exact guarantee.

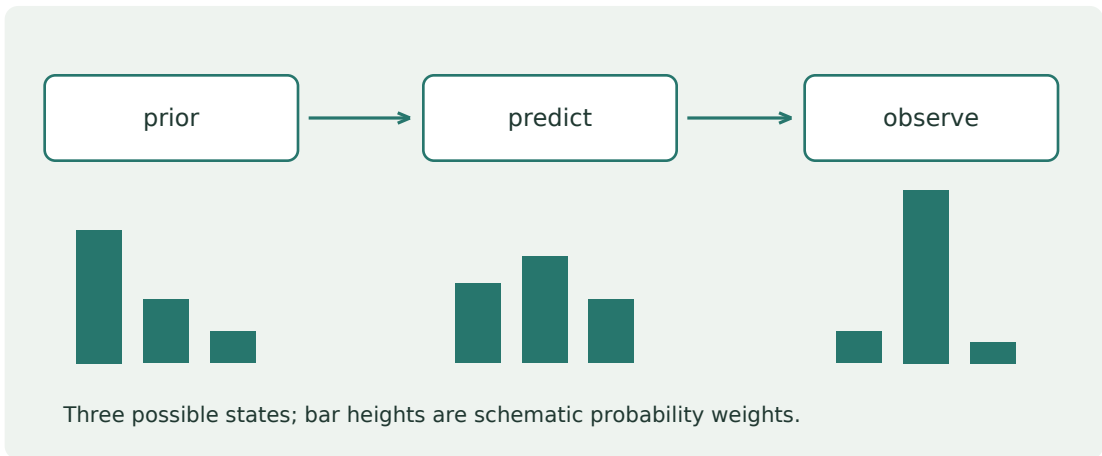


Figure 7. A belief update first predicts possible worlds and then reweights them using the observation. This schematic is not a fitted filter.

Context belongs in the model An observation can be numerically identical while its meaning changes. Millimeters are not meters. A sensor may have moved. A coefficient may refer to an old version. A task may now ask about duration rather than endpoint. Context determines how the values should be interpreted.

The foundation implementation guarded shape, nonnegative counts, an expected context string, source ID and an age bound. These checks rejected declared malformed or stale inputs. The age bound was a teaching policy, not a calibrated scientific threshold. A supplied provenance label is not authenticated provenance. A guard cannot detect a world change nobody records. [P3]

In-context learning changes the computation A model with fixed weights can produce different predictions after receiving examples in its context. The forward computation depends on those examples. That does not mean every weight has been updated, nor that the model has inferred the true law of an arbitrary world.

For the box problem, context might provide examples revealing the left count or the intervention rule. If every demonstration remains compatible with two hidden explanations, uncertainty survives. Results on Transformers learning function classes in context show a capability under specified distributions, not universal scientific identification. [R22]

A distribution is a different promise Statistical sufficiency for a model parameter is not automatically predictive sufficiency for a future target. A strong predictive statement is $P(Y | H) = P(Y | Z)$ for $Z = \phi(H)$, under a specified law. Equality of conditional means is weaker: it preserves a squared-loss average, not every tail event or multimodal outcome.

In controlled systems, prediction under one behavior policy need not preserve counterfactual effects of other actions. Predictive-state representations organize information through action-conditioned future tests; their existence results do not guarantee that finite observations learn an optimal representation of every changing world. [R2, R3, R6]

The safe reading of “the model understands the context” is operational: the declared context changes a specified calculation. It is not evidence that the model has access to every hidden fact.

8 | Programs, proofs and roundoff

Agreement has more than one layer A mathematical rule, an executable program and a physical system are different objects. A proof about one layer does not certify all three. The exact foundation compared a step-by-step interpreter with composed matrices over 85 action sequences of length zero through three on 121 starting states: 10,285 agreements. It also passed 31 unit tests. [P3]

These checks establish the retained cases. The all-state mathematical argument comes from the specified linear identities, not from multiplying the test count. Exact integer or rational arithmetic can remove one source of numerical uncertainty without making the world’s measurements exact.

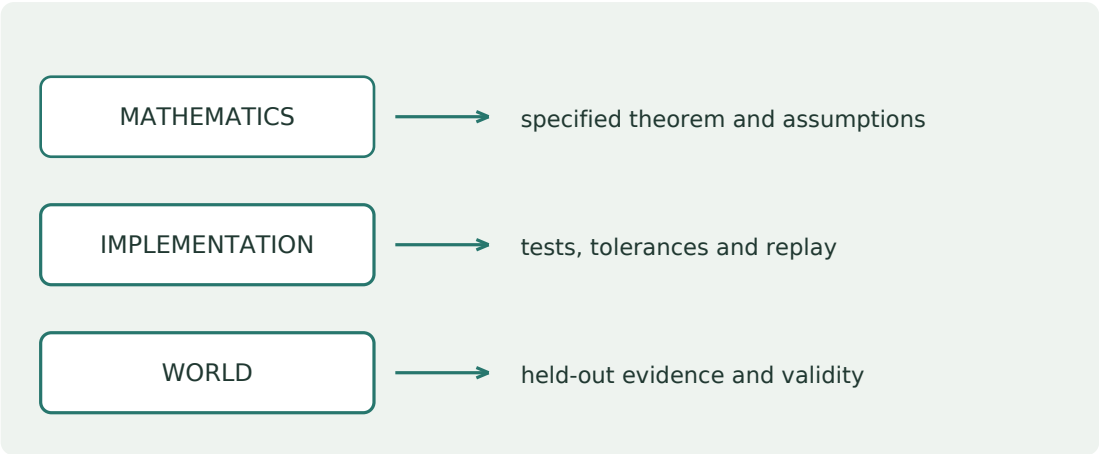


Figure 8. Three evidence layers. A mathematical theorem, runtime agreement and external validity require different evidence; none automatically certifies the next.

What the formal artifacts established The retained Lean reports include a necessary collision obstruction and fixed-sequence commutation given one-step premises. A separate affine supplement covers exact integer identities, composition and compilation of lists of supplied actions. Their proofs do not discover a representation, construct every possible reduced rule, establish stochastic closure or certify a learned world model. They were reviewed in retained sources, not rerun for this edition. [P4, P7]

The Bend source check, JavaScript compilation and fixture execution were reported as working engineering steps. Independent optional BendTT/kernel validation remained unresolved. The book does not silently promote the working compiler path into a verified end-to-end chain.

Associativity can fail in the executable numbers In binary64 floating arithmetic, rounding changes $(1 + 2^{53}) - 2^{53}$ into zero, while $1 + (2^{53} - 2^{53})$ is one. The exact integer expressions are equal; the evaluated floating expressions are not. The retained F32 fixture gives the analogous issue at 2^{24} .

Thus algebraic simplification may change numerical behavior. A compressed execution can agree to a measured tolerance rather than exactly. Its error needs a bound or a protocol, not a slogan about identical mathematics.

A derivative condition needs a uniqueness condition For smooth ϕ and full dynamics $\dot{x} = b(x)$, a proposed latent field V must satisfy $D\phi(x)b(x) = V(\phi(x))$. This expresses consistent projected velocity. It does not alone guarantee that the latent initial-value problem has a unique solution.

Take $\dot{x} = 1$ and $\phi(x) = x^3$. Then $\dot{z} = 3|z|^{2/3}$. From $z(0) = 0$, the projected full trajectory is $z(t) = t^3$. The latent equation also permits waiting at zero until any $c \geq 0$ and then following $(t - c)^3$. Its field is not locally Lipschitz at zero. The full trajectory and its actual projected flow remain uniquely specified here; the latent differential equation alone admits additional solutions. The projected velocity relation did not select a unique latent flow. Suitable regularity, a specified solution class or the actual projected flow supplies additional structure. [P6]

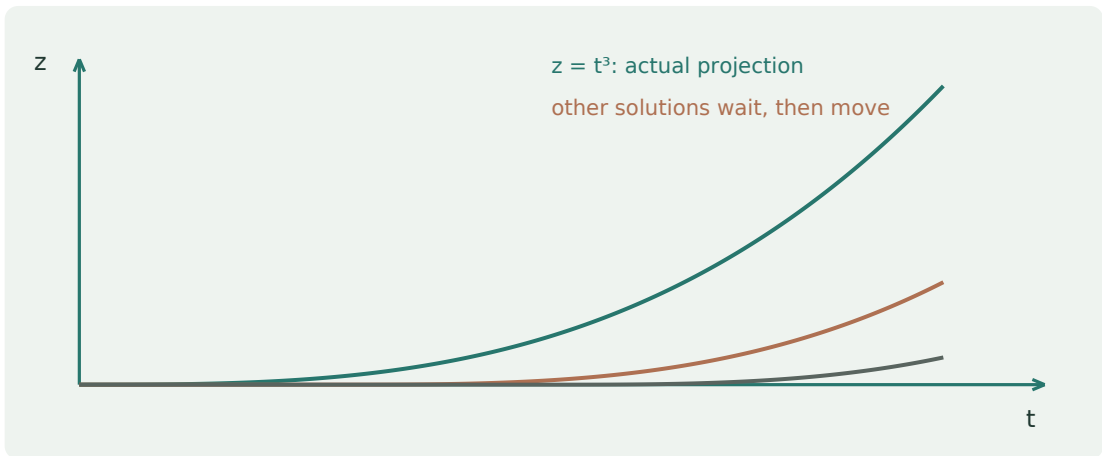


Figure 9. Exact constructed solutions of the latent equation. The actual projected full flow is t^3 ; the differential equation also permits arbitrary waiting at zero. These curves are not numerical runs.

What no general procedure can promise For effectively specified computable total program-state transitions, an output flag can remain zero until a simulated program halts. Deciding whether it agrees forever with an always-zero output would decide nonhalting. A universal all-program equivalence decider is therefore unavailable. Finite horizons, explicit finite systems and restricted algebraic families are practical escape routes. This is a prose reduction, not a new machine-checked theorem. [P5, R14]

9 | Learning interaction rules

A compact formula can carry a lot of work Consider coupled particles. Their next acceleration may depend on position, velocity, fields and interactions. A learner can build a library of candidate terms, fit their coefficients and retain a small active subset. This is different from being handed the correct physical terms.

The retained interaction study used a 105-term library and thresholded ridge fitting. In a clean interpolation condition, sparse-rule rollout RMSE was approximately 0.00012015, compared with 0.038743 for the tested generic dense fit. It nearly matched the hand-selected physics model in that setting. [P8]

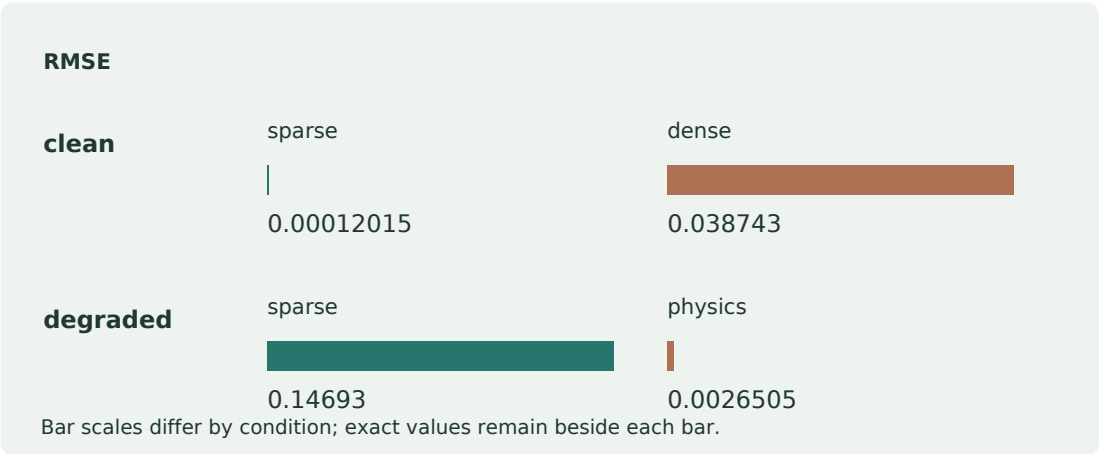


Figure 10. Retained result values. Clean interpolation and jointly degraded extrapolation are different conditions, not one pooled benchmark. Lower RMSE is better.

The interesting part is the reversal. Under the jointly noisy, sparsely observed extrapolation condition, sparse RMSE was approximately 0.14693 while the physics-selected model achieved 0.0026505. The strong prior helped. It was not learned from scratch, and its supplied family makes it a different comparator.

A fair bill includes the search The study retained 160 fits, twenty independent training seeds and 10,240 model trajectory evaluations across its conditions. Sparse fitting searched twelve settings while the cited baseline search used three. The resulting accuracy comparison is not an equal-learning-budget contest.

A separate execution comparison held the learned rule fixed and changed how it was evaluated. Compact execution was faster in the reported local CPU timings, with median speed ratios about 3.20 in clean and 2.35 in degraded conditions; maximum discrepancy was around 1.23×10^{-13} . This supports a bounded execution optimization. It does not show that sparse learning is universally faster, because the fitting search has a separate cost.

The dense generic model had 163 divergent rollouts assigned penalty fifty. A method's failure policy materially affects averages. Nonfinite trajectories cannot simply disappear from the report, but the chosen penalty is not a physical error unit.

Differentiation amplifies sensing noise Estimating velocity or acceleration by differencing nearby noisy measurements can be unstable. Roughly, central derivative noise grows as measurement noise divided by time spacing. An integral-window fit avoids explicit point differentiation but averages over a wider temporal support and changes the regression design.

A retained scalar comparison ran 360 fits and 7,200 held-out rollouts across nine noise/spacing cells. In its finest-spacing, highest-noise cell, failures fell from 80/400 to 13/400, with penalized-RMSE gain 8.4032 and ordinary paired interval [1.9095,14.8968]. Five of six noisy-cell intervals crossed zero. A descriptive decomposition assigned 8.375 of the highlighted gain to failure penalties. That is 99.66% of the mean contrast, not a causal attribution proving why the algorithms differ. [P4]

The conclusion is narrower than “integral methods solve noise”. Averaging support, overlapping rows, wider held-out support and failure accounting all changed the comparison. A stronger experiment would match what it can and expose the remaining difference.

What the model actually learned It learned coefficients and selected support inside supplied families. It did not learn every measurement convention, simulator law, detector threshold or search strategy. Compactness is a property of the resulting program. Sufficiency is a promise about its future use. One does not establish the other.

10 | A map of the model’s own mistakes

Predicting error is a task too A world model predicts an outside event. An error monitor predicts where that world model may be wrong. In the retained study, descriptors included ensemble spread, distance from observed support, predicted motion and initial velocity. A fitted regression used those descriptors to rank candidate trajectories. [P9]

The monitor accepted 32 of forty trajectories: 80% coverage. It did not improve every underlying prediction. All selection policies shared the same numerical predictor. A lower accepted error can result from withholding harder cases, so coverage must stay beside the error.

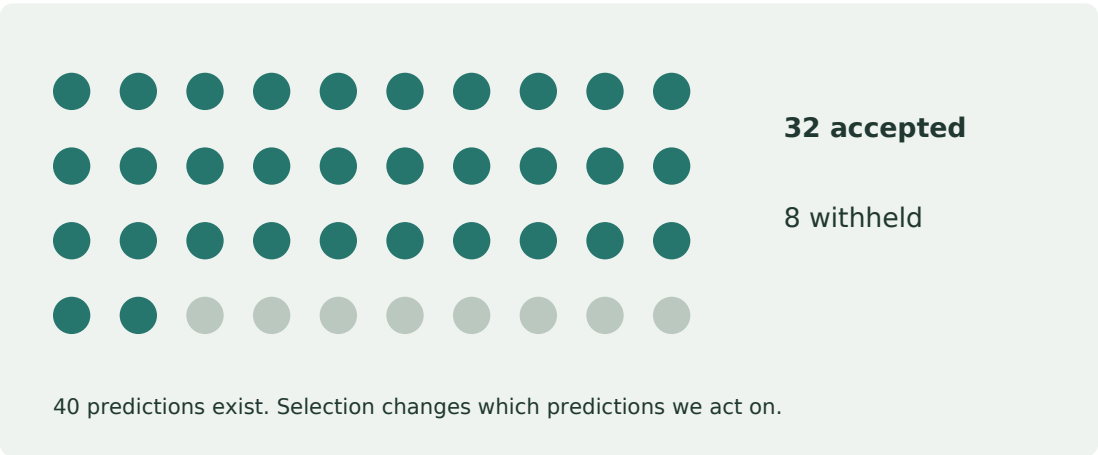


Figure 11. Schematic forty-case cohort at 80% coverage: thirty-two accepted, eight withheld. Lower selective error is not full-coverage improvement.

After a common model update, mean accepted trajectory RMSE was 0.004732855 for the monitor versus 0.005014883 for random selection within range. Under shifted field/coupling it was 0.01088819 versus 0.01140392. Paired twenty-seed intervals favored the monitor over random selection in those comparisons.

The stronger baseline changed the interpretation: monitor-versus-ensemble-spread intervals crossed zero in every reported contrast. Incremental benefit beyond that cheaper uncertainty heuristic was

not established. An inconclusive interval is not proof of equality, but it cannot support the claimed superiority either.

An index makes a calculation reproducible The implementation retained 680 stable keys linking model versions, fitted values, support, inputs, predictions, decisions and evaluator errors. Replaying eighty cohort predictions and monitor scores reproduced the stored values exactly. Recorded selected IDs agreed. Deliberately corrupted or stale references were rejected.

Those are useful engineering properties. An index can say which model produced a decision and reconstruct the calculation. It cannot recover observations that were never retained, authenticate every source declaration or establish that the prediction was true. Replay fidelity and predictive validity are independent claims.

An operational self-model is not consciousness The modest meaning of self-representation here is a stored, executable description of the learner's predictive reliability that affects acceptance decisions. That is already a useful computational object. Nothing in the study measures a subjective point of view or recursive general self-awareness.

The later event-monitor implementation had guarded windows, recorded action states and fresh-window processing. Its warnings remained heuristics. It did not introduce a continuous autonomous daemon, guaranteed drift detection or automatic safe retraining. Engineering completion did not resolve the earlier monitor-versus-spread uncertainty. [P4, P9]

What would earn the stronger claim A protected new regime, a fixed total budget and a predeclared useful improvement over spread-only acceptance would make a meaningful test. Refreshed, stale and shuffled monitors should share the same predictor and coverage. If the uncertainty interval does not clear the required improvement, the claim of incremental value should be rejected or left unresolved.

A system can make its uncertainty visible without claiming that uncertainty is calibrated. An ensemble's disagreement is a signal to test, not automatically a probability of failure. This is the bridge from keeping an evidence log to choosing when more information might be worth acquiring. [R18]

11 | The price of observing more

Information arrives on a schedule A compact online learner senses a few coordinates. After residual error rises, it scans more coordinates and refits. The attractive story is that it uses little information until the world changes. The less attractive part is the delay while it notices and learns.

In the first retained sensing study, a change detector needed thirty-two examples and the diagnostic scan needed sixty-four more. A ninety-six-example delay occupies half of a regime lasting 192 examples. Excellent late predictions cannot erase the earlier mistakes. [P10]

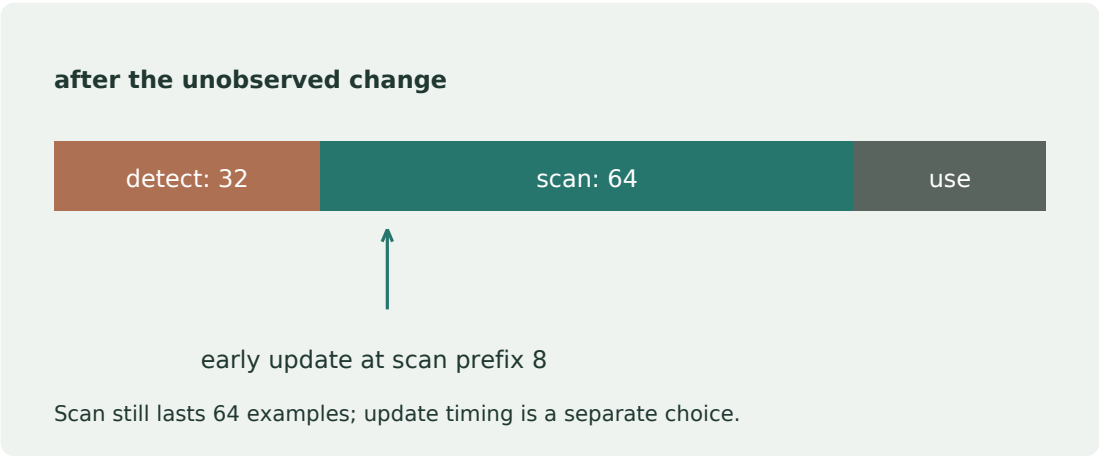


Figure 12. Protocol timing schematic. A 32-example detection delay plus a 64-example scan consumes 96 examples; early updates can start before the scan ends.

At sixteen dimensions with sparse changes and output noise 0.1, the compact adaptive policy used about 26.2% of the full sensor scalars, but postchange MSE was about 0.430 versus 0.277 for the tested full-observation policy. Less sensing bought more error. The study did not establish readiness to replace full observation.

Use information before the scan ends The follow-up fitted useful prefixes after eight or sixteen diagnostic examples instead of waiting for sixty-four. This attacked latency rather than inventing faster access to the entire scan. All scans still acquired sixty-four examples. Updating earlier is not the same as sensing fewer.

The incremental study retained 400 streams, 2,400 learner runs and 1,520,640 scored-before-feedback predictions. Early updating improved some sparse cases. Dense changes still defeated the narrow selected representation. The preserved reports separate native detectors and shared-controller analyses. [P11]

Match recovery before crediting compression In the matched-recovery study, full-state learners received the same trigger events and early refit opportunities. In the sixteen-dimensional, low-noise sparse cell, selective update8 achieved postchange MSE 0.2225 and full reset8 achieved 0.2298. Their difference was -0.0073 , with paired interval $[-0.0164, +0.0018]$. This did not establish postchange superiority, equivalence or noninferiority.

For the specified first-sixty-four-after-change comparison, the reported interval lay below zero: -0.0360 $[-0.0627, -0.0093]$. Full reset8 improved on full reset64 by 0.1087 $[0.0946, 0.1227]$. Earlier useful recovery explained substantial gain. These are ordinary paired Student-t 95% intervals over twenty independent seeds per cell, unadjusted for multiple comparisons. Prediction rows are workload, not independent replicates. [P12]

At sixty-four dimensions, a selective advantage remained against the tested small-prefix full ridge fit. A full-window control improved that comparator, and high-dimensional fits on tiny prefixes were weak. The finding does not overturn the optimum obtainable with full information. All four dense-change cells favored richer methods over selective recovery.

More information cannot worsen the free optimum Let H be available information and $Z = s(H)$ a measurable summary, so $\sigma(Z)$ is contained in $\sigma(H)$. Define $m(H) = E[Y|H]$ and $mZ = E[Y|Z]$. For square-integrable Y and unrestricted squared-loss prediction:

$$R^*(Z) - R^*(H) = E[(m(H) - mZ)^2] \geq 0$$

The unrestricted full-information predictor can imitate any summary-based predictor. A finite trained full model can nevertheless lose through variance, regularization, model mismatch, poor updates or compute limits. Those practical losses are important; they are not proof that removal creates information. [P5]

Equal ceilings are not equal actual spending. Shared trigger access is a mechanism-analysis privilege, not automatically a deployable independent algorithm. The next claim must name both the accuracy frontier and the complete bill.

12 | Refit the rule, or ask another question?

Two causes can look like one residual A predictor used to work and now fails. If both boxes used to double and now both triple, the total remains sufficient. The mapping is stale: refitting can replace 2z with 3z. If the action changes to doubling only the left, the total lacks a necessary distinction. Refitting a deterministic function of total alone cannot solve exact prediction for every arrangement.

Noise, optimization and finite samples create other explanations. Stale fit and missing information are useful controlled cases, not a complete taxonomy of real failure.

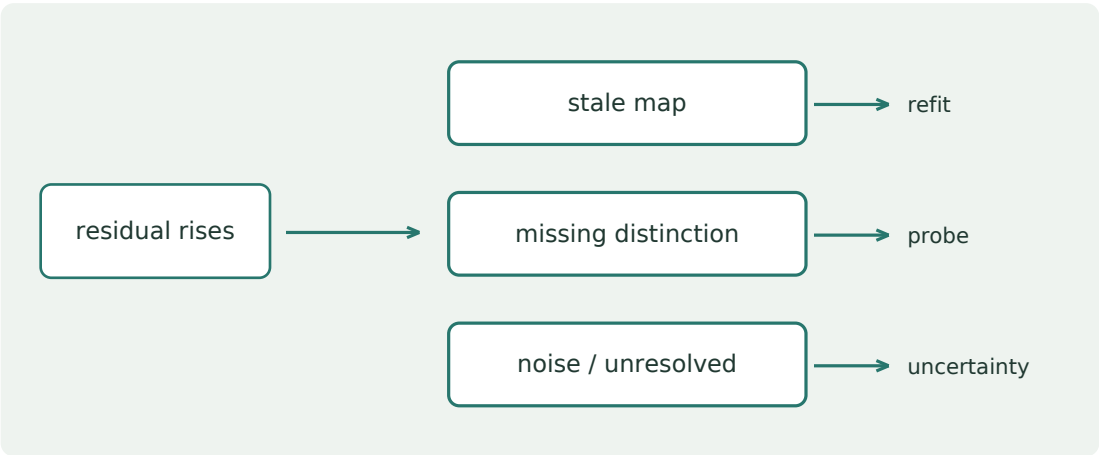


Figure 13. Different failure explanations require different responses. A residual alone need not distinguish stale mapping, missing information and irreducible noise.

Diagnosis needs access, not a confident label If two hidden causes produce the same distribution of visible histories under every permitted adaptive probing policy, no diagnostic function of those histories can distinguish them. With equal priors, expected binary accuracy is at most one half. At a fixed policy and horizon, optimal equal-prior accuracy is $(1 + TV)/2$, where TV is total variation distance between the history laws.

The statement concerns distributional equality under declared access. A new sensor or intervention can break that equality. A prior can change the decision. Neither should be smuggled into an apparently information-matched comparison. [P2, P5]

For independent standard Gaussian x_2 and u , the targets $x_1 + x_2 + \text{small noise}$ and $x_1 + u + \text{small noise}$ have the same compact-view conditional law when only x_1 is observed and the hidden variables are identically distributed. Observing x_2 can distinguish their predictive relation. A residual alarm alone cannot tell us which explanation generated the unseen term.

What the retained diagnosis experiment found The bounded diagnosis study compared eight policies over two hundred paired streams: fifty seeds in each of four regimes. It retained 1,600 policy runs and 819,200 scored-before-feedback predictions. The selector refitted on all fifty stale-fit streams and acquired the additional sensor on all fifty planted missing-information streams. Those observations describe the constructed conditions, not a universal diagnosis guarantee. [P13]

In the independent-noise null, it acquired unnecessarily on nine of fifty streams. Its online MSE exceeded the fixed compact predictor by 0.01113, with paired interval [0.00381, 0.01845]. Rolling compact performed better in the stale-fit condition; rolling full performed better in the missing-information condition. Waiting for a diagnosis had a price.

The later calibration study continued this experiment's generator, not the separate sixteen/sixty-four-coordinate latency study. That provenance distinction prevents an apparently repeated percentage from being treated as the same experiment.

What remains a proposal The broad working note describes a benchmark contract: observations, actions, change distributions, train/test split, horizon, costs and baselines fixed before outcomes. It is a scaffold rather than an empirical paper by itself. The later executed pilots test narrower procedures. They do not complete every proposed benchmark field or establish exclusive causal diagnosis.

A useful selector must beat strong refit-only, periodic, random and uncertainty-triggered acquisition methods under matched probe access and charged overhead. If it does not, the right result is a negative finding, not a new name for the same failed advantage. [R7-R10]

13 | Calibration and the price of choosing

A winner can be a lucky winner Compare two candidate models on a small validation set and choose the lower loss. The chosen difference is no longer an ordinary unselected estimate. Noise helped some candidates win. Repeated choice can make apparent gains more optimistic than fresh performance. This statistical mechanism is established; our pilot illustrates it rather than inventing it. [R17]

The retained calibration study allowed one residual alarm and one forty-eight-example paid probe. It fitted compact and full ordinary least-squares candidates on thirty-two probe examples and selected on the next sixteen. A chosen candidate was refitted on all forty-eight and affected predictions only afterward. [P14]

Freeze the margin before evaluation Calibration used two hundred seeds in each of two null conditions. It compared a forced fixed window with alarm-conditioned probe windows. For each mode, the margin was the maximum of two empirical 95th percentiles, bounded below by the inherited 0.02 margin. The frozen margins were 0.1144 and 0.1193. These empirical quantiles set thresholds; they do not certify a population error bound. Evaluation used one hundred fresh paired seeds in each of three conditions and five policies: 1,500 runs, 768,000 scored-before-feedback predictions.

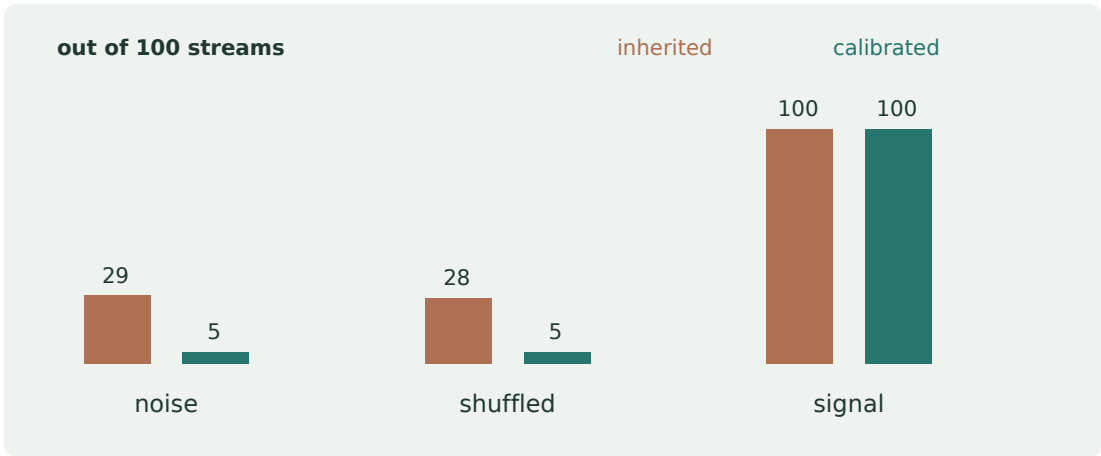


Figure 14. Retained deployment counts, 100 fresh seeds per condition. Both calibrated selectors had the same counts. Bars are descriptive; intervals are in the text.

Independent-noise full-model deployments fell from 29/100 to 5/100; shuffled-null deployments fell from 28/100 to 5/100. Every selector deployed the full model on all one hundred strong planted-signal streams. Fixed-window and alarm-conditioned methods were nearly indistinguishable here. There was no demonstrated extra benefit of conditioning on the alarm.

Five percent is a count, not a guarantee A five-in-one-hundred estimate has a 95% Wilson interval of approximately [2.15%,11.18%]. It is not proof of a universal 5% false-acquisition bound. One hundred successes have a lower Wilson bound near 96.3%, not proof of perfect detection outside these draws. The signal amplitude was strong and no weak-signal sweep was performed.

Among the five alarm-calibrated noise selections, average compact-minus-full validation gain was +0.1732. On the subsequent fresh evaluator-only diagnostic it was -0.2143. These tiny selected subsets are descriptive and unstable. The shadow diagnostic scored the original thirty-two-row candidates, not the forty-eight-row deployed refits; it never influenced policy decisions. Actual online MSE separately scored the deployed models.

The shuffled null preserved targets and the sensor multiset while permuting the sensor column across the finite stream. It created dependence and was an approximate additional null, not an exact conditional-randomization guarantee.

Keep the losing result and the invoice Signal-condition online MSE was 0.14988 for the selectors versus 0.08996 for rolling full. The difference, +0.05992 [0.05578,0.06406], favored rolling full. Discovery latency remained even after calibration improved false deployments.

Every run had a 24,000-unit abstract-work ceiling. Both calibrated policies paid 7,014.11 units per run for conservative amortization of the combined calibration work over three hundred planned streams. Maximum realized work was 23,261. Rolling full had eleven denied refits per stream. The amortized calibration charge and maximum realized work use the same declared accounting basis, including the charged calibration where applicable. These are proxies and reuse assumptions, not dollars, instructions, energy or an optimized economic allocation.

Nineteen recorded audit checks passed, including frozen hashes, disjoint seeds, noninterference, accounting and deterministic replay. They are internal engineering evidence. Ordinary paired intervals are not familywise or time-uniform guarantees. Calibration reduced a specified failure rate while leaving selection optimism and the information-rich baseline's advantage visible. [R17, R25, R26]

14 | From representations to useful agents

Prediction becomes a decision A model predicts what happens. A policy chooses what to do. A reward describes the preference used by a specified task. A value estimates expected accumulated reward under a policy; it is not a complete account of human value.

$$V\pi(x) = E\pi[\sum_t \gamma^t r_{t+1} \mid x_0 = x], 0 \leq \gamma < 1$$

The expectation must be well-defined. Bounded rewards are one sufficient condition for finite value when $\gamma < 1$; finite horizons still require appropriate reward integrability. Additional assumptions can support other discount choices. A temporally extended action, or option, has an initiation set, a policy and a termination condition. Its planning model needs consequences, accumulated reward and relevant timing or discount information, not merely the final picture. [P3, R13]

A skill can preserve an endpoint and lose its cost Our hand-written option transfers one counter at a time from left to right until the left is empty. It ends at $(0, u + v)$, takes u steps and returns $-u$ with reward -1 per step and no discount over its finite execution. All retained starting states passed these exact checks.

States $(0,10)$ and $(5,5)$ have the same total and final total, but durations zero and five. Total is sufficient for the endpoint total, not for planning the time or cost. Storing $(total, left)$ resolves the ambiguity, but no longer reduces the number of coordinates of this toy.

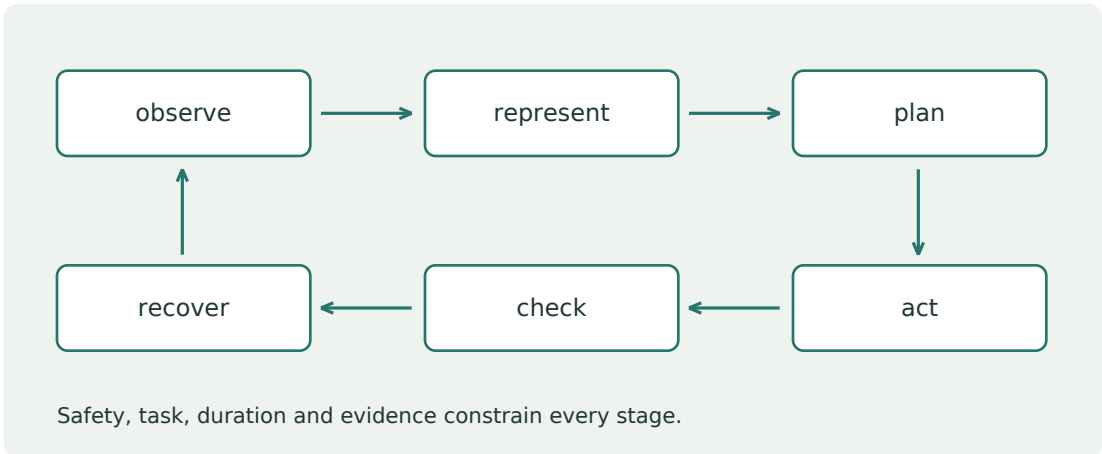


Figure 15. A proposed agent loop. The option must model timing and cost as well as endpoint; feedback and safeguards are part of the contract. Not an executed robot system.

The OaK direction links feature construction, subtasks, options, option models and planning. Our foundation supplied the feature, subtask and option. It did not execute a continually learning complete architecture, learned option discovery or a meta-learned step-size system. A research direction should remain a direction until its components and tests exist. [P3, R13]

Physical agents sharpen the contract Shubham reports that Phi9 has robotic arms and has trained some compression-based policies. This edition has not inspected checkpoints, datasets, controller code, force limits or task-performance logs. This is an attributed, unverified project report. The supplied evidence does not verify the hardware configuration, trained policies or task capability.

A proposed first demonstration should use supervised, nonclinical tabletop tasks with safe props: sort objects into marked trays, move a blunt training tool, and operate a zipper pouch fixed to a jig. These are proposals, not tasks executed for this book. A human supervisor, emergency stop, limited workspace, conservative speed/force limits and a defined fallback belong in any actual protocol.

The medical-workflow motivation is assistance with routine manipulation around a clinician's work. This is not permission to operate near a patient, handle contaminated sharps or claim surgical autonomy, sterility or clinical validity. A safe-prop bench test cannot establish those properties. Patient

data, infection control and clinical deployment would require separate qualified processes.

A demonstration should reveal failures A future protocol could score object-placement success, zipper completion, contact or drop events, completion time, human intervention and recovery. Hold out objects and layouts; distinguish taught tasks from generalization; compare compression variants with a strong full-observation policy under the same permissible information and safety conditions. Record failures as prominently as successes.

We have not frozen or run that protocol here. The current book ends where a credible next experiment begins: choose a useful distinction, define how to check it and admit when the evidence is insufficient. The goal is not a representation that sounds universal. It is a system whose promises survive contact with the task.

Appendix A | The technical contract

A1. Exact deterministic closure Let $F_a: X \rightarrow X$ be total for each admitted action a , Q a family of readouts and $\varphi: X \rightarrow Z$. Readout and represented transition maps on $\varphi(X)$ exist if and only if, for every collision $\varphi(x) = \varphi(x')$, both $q(x) = q(x')$ for each q in Q and $\varphi(F_a(x)) = \varphi(F_a(x'))$ for each a . Necessity applies a function to equal inputs. Sufficiency chooses any representative, defines the maps using it and invokes the collision conditions. Induction preserves any finite action word. This proof is elementary and classical; no new Lean run certified the converse construction.

Define $x \approx x'$ when all specified future outputs agree after every finite action word, including the empty word. Prefixing a word by an action establishes successor stability. Any representation that closes under every admitted action and preserves every specified readout can merge only $\approx$ -equivalent states. The quotient by $\approx$ is coarsest in this behavioral sense, not necessarily cheapest to compute, fewest bits under every encoding or easiest to learn. Finite-horizon equivalence at H gives successor equivalence at $H-1$, not necessarily at H ; a remaining-horizon index may be needed.

A2. Linear and affine special case On finite-dimensional vector spaces, $\varphi(x)=Px$, $F_a(x)=A_ax+b_a$ and $q(x)=Qx$ close precisely when $A_a \ker(P)$ is contained in $\ker(P)$ for every a and $\ker(P)$ is contained in $\ker(Q)$. Differences of colliding states lie in $\ker(P)$. Their transformed differences must remain invisible, and their task differences must vanish. Define $G_a(Px)=PA_ax+Pb_a$ and $h(Px)=Qx$; these conditions make them well-defined on the image. An arbitrary extension outside the image is not determined by the contract. Restricted physical domains require their own collision analysis rather than assuming the unrestricted vector-space necessity.

A3. Continuous dynamics With a specified differentiable projection and full vector field, the chain rule gives $\dot{z}=D\varphi(x)b(x)$. Fiber consistency makes a single-valued projected field possible. A well-posed autonomous reduced initial-value problem additionally needs appropriate regularity and solution conditions. The cubic projection in chapter eight separates consistent velocity from unique flow. Numerical integration error is a third issue, independent of both.

A4. Stochastic and predictive promises For a finite controlled Markov system, equality of the joint conditional law of the next represented state and required emitted output or reward, for colliding states under each action, is sufficient to preserve their joint trace law. General spaces need measurability and appropriate conditional kernels. Expected additive-return tasks may require expected rewards rather than their full distributions, under the specified Markov model. State-space bisimulation, output-trace equality and prediction under one behavior policy are not interchangeable promises.

Let $Z = s(H)$ be a measurable summary, so $\sigma(Z)$ is contained in $\sigma(H)$. For square-integrable Y , conditional expectation minimizes squared loss. Write $Y-mZ=(Y-m)+(m-mZ)$; the cross term has expectation zero because $m-mZ$ is H -measurable and $E[Y-m|H] = 0$. The difference of optimal risks is $E[(m-mZ)^2]$. Zero defect implies conditional-mean sufficiency under that law, not full-distributional sufficiency or exact dynamic closure.

A5. Distinguishability Equal history laws under every permitted policy imply equal output laws for any history-based diagnostic. Under equal binary priors its expected accuracy is at most one half. The fixed-experiment optimal accuracy $(1+TV)/2$ is a standard testing identity. Different priors, longer horizons or revealing probes change the experiment. Finite-sample difficulty is not the same as exact indistinguishability.

Appendix B | Evidence ledger, not a combined benchmark

Each row below belongs to its own protocol. Different outputs, losses, horizons, sensors, randomization units and proxies prevent pooling these studies into one average “Phi9 improvement”. Counts are retained execution reports, not new runs. Mechanical audits are internal checks, not external replication.

Study	Retained scale	Main boundary
Exact foundation	121 states; 31 tests; 10,285 comparisons	Supplied deterministic integer/rational toy; no universal learning
Interaction fitting	160 fits; 20 seeds; 10,240 trajectory evaluations	Unequal tuning budgets; strong physics prior; degradation reversal
Scalar identification	360 fits; 7,200 rollouts; 9 cells	Failure penalties dominate one highlighted contrast
Self-monitor	200 ensemble fits; 3,200 predictions; 12,800 decisions	80% coverage; spread-only superiority unresolved
Initial sensing	320 streams; 960 runs; 552,960 predictions	Detection and scan latency; sparse/dense difference
Incremental sensing	400 streams; 2,400 runs; 1,520,640 predictions	Early use of same 64-example scan
Matched recovery	240 streams; 1,680 learner runs; 967,680 predictions	Shared controller; update opportunities not equal cost
Diagnosis	200 streams; 1,600 runs; 819,200 predictions	9/50 noise acquisitions; rolling baselines can win
Calibration	1,500 runs; 768,000 predictions; 19 audit checks	5/100 is a pilot estimate; no global error guarantee
Oscillator demo	6 trajectories; 3 assertions	No model trained; numerical rather than robotics evidence

The calibration comparison in full The following table reproduces the retained online MSE means. Lower is better. One hundred paired seeds per condition; predictors are scored before their own feedback. The names refer to the frozen policies, not hypothetical optimally tuned families.

Condition	Policy	Online MSE	Postchange MSE
Noise	Inherited	0.58790	1.04408
Noise	Fixed-window	0.57834	1.02670
Noise	Alarm-conditioned	0.57809	1.02623
Noise	Rolling compact	0.58044	1.03032
Noise	Rolling full	0.58450	1.03732
Signal	Inherited	0.14988	0.26118
Signal	Fixed-window	0.14988	0.26118
Signal	Alarm-conditioned	0.14988	0.26118
Signal	Rolling compact	0.57176	1.01483
Signal	Rolling full	0.08996	0.15375
Shuffled	Inherited	0.57796	1.02568
Shuffled	Fixed-window	0.56964	1.01083
Shuffled	Alarm-conditioned	0.56964	1.01083
Shuffled	Rolling compact	0.57176	1.01483
Shuffled	Rolling full	0.57765	1.02556

What this edition does not rerun No neural model, robot policy, C-MAPSS model or new empirical campaign was trained or evaluated. The oscillator curves are analytic explanatory drawings here; six earlier numerical trajectories remain in their original archive. The new gradient example and Koopman lift are worked algebra, not execution receipts. Existing study wall times and checks are cited from retained reports. No formal binaries or compiler chain were independently recertified.

The consolidated v2 source bundle has known historical raw-part gaps documented in its archive. This book does not manufacture missing rows or imply that textual reports replace every original raw array. Diagnosis and calibration retain separate raw-data packaging. Reproduction of the book requires only its own source and rendering dependencies; reproduction of each experiment requires that experiment's preserved code, settings and data companions.

Appendix C | Terms that should stay distinct

State: a description of the system at one instant, with a stated domain.

Observation: what the sensing process provides, which may not determine the state. Representation: retained coordinates or features of state/history.

Readout: a map from the representation to a requested output.

Transition: a next-state rule under an action. Vector field: an instantaneous direction and rate. Flow: the time-evolution map, with existence/uniqueness conditions. Trajectory: the resulting path. Basis change: an invertible re-expression when all coordinates are retained. Compression: a reduction that may merge distinctions.

Closure: self-consistent evolution of the specified representation. Task sufficiency: enough information for the specified output promise; it may be weaker than whole-representation closure. Reconstruction: recovery of an input/state, a different objective. Conditional mean: the squared-loss optimal average under a specified law, not the whole predictive distribution.

Learning: changing a fitted rule, representation or other specified component from experience. In-context adaptation: a computation depending on supplied examples, possibly with fixed weights. Calibration: a procedure relating scores or decisions to outcomes under a protocol, not automatic immunity to future shift. Selection optimism: apparent performance inflated by choosing on finite noisy data.

Policy: an action-selection rule. Option: a temporally extended action with start, continuation and stop conditions. Coverage: the fraction of cases on which a selective system acts. Risk: the declared loss on the declared population or selected cases. Cost: a stated resource measure, not an interchangeable synonym for milliseconds, sensing, energy and money.

A useful habit is to ask four questions before accepting any claim: which variable is represented, which operation changes it, which output is evaluated, and which receipt supports the result? If the words slide between meanings, the diagram should slow us down rather than conceal the slide.

Appendix D | Sources, coverage and editorial method

Preserved Phi9 source families [P1] The world behind the number, candidate chapter 01, edition 1, 2 October 2026, nine pages in the actual attached PDF. This supplies the core box narrative, elementary closure proof and proposed experiment. It is the one explicitly numbered pre-consolidation book chapter.

[P2] Representation closure and diagnosis after change, working-note.pdf, four-page paper scaffold. Its analysis plan and proposed benchmark are not results. [P3] From addition to OaK, ten-unit foundation notebook, preserved in consolidated v2. [P4] Task-relative sufficiency technical study, revision 2, and What a model is allowed to forget, essay revision 2, cover scheduler, formal/runtime and monitor boundaries.

[P5] Sufficiency is a contract, not a representation size, fifteen-page literature/thesis revision, 1 October 2026. It replaces overbroad claims and retains a primary-text source ledger. [P6] The state, the motion and the readout, eleven-section companion and source/demo archive, 2 October 2026. Its earlier Codex critique and six numerical trajectories remain separate from this book critique.

[P7] An executable core, and what its proofs guarantee, verified-core supplement. [P8] Technical-paper-measured and Research-article-measured: interaction learning and execution comparisons. [P9] Self-monitor research note and technical supplement: fixed-coverage error selection and indexed replay.

[P10] Sensing latency and representation cost. [P11] Incremental diagnostics results. [P12] Matched recovery results. These are separate one-step sensing studies, not a single dynamical world-model trial. [P13] Budgeted diagnosis code/results and retained protocol, 2 October 2026. [P14] Calibrating selective sensor acquisition after a residual alarm, seven-page actual PDF and original study source, 2 October 2026.

Coverage is thematic consolidation of sixteen inspected/inventoried PDFs plus diagnosis protocol/results and relevant retained source documents. Redundant versions, full coefficient dumps, every exploratory contrast, raw prediction rows and all original notebooks are not reproduced in the reading text. The input inventory and attachment mapping in the editable archive locate them. Originals were not edited. Companion sections and notebook units were reorganized, not miscounted as completed chapters.

Established references and reading trail The earlier thesis review inspected selected primary-text passages; this edition uses that retained ledger rather than claiming a fresh full-paper review of every citation. Links locate the primary sources. A reference supplies context, not proof that our different implementation inherits its guarantee.

[R1] Hopcroft (1971), An $n \log n$ Algorithm for Minimizing States in a Finite Automaton. <https://www.cs.cmu.edu/~cdm/resources/Hopcroft1971.pdf>

[R2] Fisher (1922), On the Mathematical Foundations of Theoretical Statistics. <https://errorstatistics.com/wp-content/uploads/2021/04/fisher-1922.pdf>

[R3] Shalizi and Crutchfield (2001), Computational Mechanics: Pattern and Prediction, Structure and Simplicity. <https://bactra.org/research/cmpps.pdf>

[R4] Li, Walsh and Littman (2006), Towards a Unified Theory of State Abstraction for MDPs. <https://thomasjwalsh.net/pub/aima06Towards.pdf>

[R5] Ferns, Panangaden and Precup (2004), Metrics for Finite Markov Decision Processes. <https://arxiv.org/pdf/1207.4114>

[R6] Littman, Sutton and Singh (2001/2002), Predictive Representations of State. https://proceedings.neurips.cc/paper_files/paper/2001/file/1e4d36177d71bbb3558e43af9577d70e-Paper.pdf

[R7] Shim, Hwang and Yang (2018), Joint Active Feature Acquisition and Classification with Variable-Size Set Encoding. https://proceedings.neurips.cc/paper_files/paper/2018/file/e5841df2166dd424a57127423d276bbe-Paper.pdf

[R8] Bifet and Gavaldà (2007), Learning from Time-Changing Data with Adaptive Windowing. Author manuscript:

<https://www.cs.upc.edu/~Gavalda/papers/adwin06.pdf>

[R9] Rahbar et al. (2023), Efficient Online Decision Tree Learning with Active Feature Acquisition.

<https://www.ijcai.org/proceedings/2023/0463.pdf>

[R10] Wang and Stadler (2021 preprint), Online Feature Selection for Efficient Learning in Networked Systems. <https://arxiv.org/pdf/2112.08253>

[R11] Tishby, Pereira and Bialek (1999), The Information Bottleneck Method. https://www.princeton.edu/~wbialek/our_papers/tishby+al_99.pdf

[R12] Zhang et al. (2021), Learning Invariant Representations for Reinforcement Learning without Reconstruction.

<https://arxiv.org/pdf/2006.10742v2>

[R13] Sutton, Bowling and Pilarski (2022), The Alberta Plan for AI Research.

<https://arxiv.org/abs/2208.11173> ; Sutton, Oak architecture invited talk:

<https://neurips.cc/virtual/2025/loc/san-diego/invited-talk/109601>

[R14] Rice (1953), Classes of Recursively Enumerable Sets and Their Decision Problems. <https://www.ams.org/journals/tran/1953-074-02/S0002-9947-1953-0053041-6/S0002-9947-1953-0053041-6.pdf>

[R15] Shao, Mao and Zhang (2022), Learning Task-Oriented Communication for Edge Inference: An Information Bottleneck Approach.

<https://arxiv.org/pdf/2102.04170v3>

[R16] Kemertas and Aumentado-Armstrong (2021), Towards Robust Bisimulation Metric Learning. <https://proceedings.neurips.cc/paper/2021/file/e/256bf8e6923a52fda8ddf7dc050a1148-Paper.pdf>

[R17] Cawley and Talbot (2010), On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation.

<https://jmlr.org/papers/v11/cawley10a.html>

[R18] Geifman and El-Yaniv (2019), SelectiveNet; Lakshminarayanan, Pritzel and Blundell (2017), Deep Ensembles.

<https://arxiv.org/abs/1901.09192> ; <https://arxiv.org/abs/1612.01474>

[R19] Kingma and Welling (2013), Auto-Encoding Variational Bayes.

<https://arxiv.org/abs/1312.6114>

[R20] Vaswani et al. (2017), Attention Is All You Need.

<https://arxiv.org/abs/1706.03762>

[R21] Hafner et al. (2019), Dream to Control: Learning Behaviors by Latent Imagination. <https://arxiv.org/abs/1912.01603>

[R22] Garg et al. (2022), What Can Transformers Learn In-Context?

<https://arxiv.org/abs/2208.01066>

[R23] Chen et al. (2018), Neural Ordinary Differential Equations.

<https://arxiv.org/abs/1806.07366>

[R24] Brunton, Brunton, Proctor and Kutz (2016), Koopman Invariant Subspaces and Finite Linear Representations of Nonlinear Dynamical Systems for Control. New primary PDF check for this edition: introductory invariant-subspace discussion and operator/generator definitions, printed pp. 2-5; the polynomial lifting construction on printed pp. 8-9 was also checked against the example in chapter six. Not a claim to have reviewed all nineteen pages.

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0150171>

[R25] Dwork et al. (2015), Generalization in Adaptive Data Analysis and Holdout Reuse. <https://arxiv.org/abs/1506.02629>

[R26] Howard et al. (2021), Time-uniform, nonparametric, nonasymptotic confidence sequences. <https://doi.org/10.1214/20-AOS1991>

What was learned from Welch Labs The official title located is The Welch Labs Illustrated Guide to AI. Its official page describes 376 pages, nine chapters, full-color visuals and exercises, spanning perceptrons through AI video generation. One current book was established, not a verified collection of additional volumes. <https://www.welchlabs.com/ai-book>

Access was the official public product page, the freely linked fifty-seven-page exercises PDF and the author's linked public code directory. The complete paid book was not acquired or read. Selected rendered sample pages inspected were PDF pages 3, 24, 25, 50 and 51, corresponding to printed pages 54, 139, 140, 292 and 293. These show concrete perceptron cases, forward/backward calculation tables and attention matrices annotated with dimensions. The exercises document was text-extracted to identify its coverage; not every exercise was solved.

The useful principles were to keep one concrete object in view, draw what each operation changes, expose shapes and dimensions, let a worked calculation bridge intuition and equations, and pair explanations with an activity that can reveal misunderstanding. This book applies those general principles to original Phi9 box, flow, error and acquisition diagrams. Welch prose, exercises and artwork are not reproduced. Public code was identified and its directory inspected, not executed. No purchase or unauthorized copy was made.

Official sample:

https://www.welchlabs.com/s/exercises_rev_12_nov_14_1.pdf Author code:
https://github.com/stephencwelch/ai_book

Editorial review and reproducibility A bounded text-only Codex review of the chapter map, mathematical statements, quantitative claims and selected teaching passages accompanies this edition. Its exact prompt, candidate answer, effective-isolation check and execution receipt are included in the editable archive. The receipt records provider completion, a submitted model turn and simulation false. Zero tool-call or approval-request events appeared in the returned ledger. The review covered complete chapters 3, 6, 8 and 11-14, Appendix A, the full chapter map and a truncated teaching excerpt from chapter 4. It did not inspect unseen files or the final PDF. It is not an independent replication, formal proof or a review of every rendered page. The independently assessed edits make nested information explicit in the Bayes-risk identity, require both closure and readout preservation in behavioral minimality, preserve joint stochastic laws where joint traces are promised, distinguish Koopman operators from generators, and state value-integrability and statistical sampling assumptions. The separate disposition records adopted and unimplemented suggestions. Codex output is critique, not mathematical proof or replication.

The book renderer, editable Markdown, original diagram instructions, result tables, source coverage, attachment mapping and archive manifests are supplied together. The book can be rebuilt without Codex credentials, robot access or rerunning any experiment. This edition's visual/layout checks concern the delivered pages; its scientific boundaries concern the evidence actually retained.