

When Does an Explicit Self-Model Help? Calibration and Control under Internal Change in a Finite-State Pilot

R4 research manuscript

Source: 14 October 2026
Site reading copy 1: 16 October 2026

Internal AI review; external peer review and publication approval are separate.

Abstract

An actuator that changes during operation creates a latent-state control problem. Estimating the changed actuator can improve behavior, but this does not establish that a representation called a self-model is necessary. We analyze a frozen finite benchmark that separates those questions. A binary actuator polarity is inferred from three noisy probes, may flip once with probability $1/4$, and affects a three-action episode through a noisy XOR channel. Five policies operate under common observation access, feedback, horizon, probe budget, model specification, and permission to retain their declared finite state: a Bayesian tracker designated explicit, a history-matched compiled tracker, the same filter attributed to an external observer, a periodic policy, and a two-contradiction change-point heuristic. Equal primitive-operation or storage costs are not established for all five.

The supplied authoritative artifact reports exact equality of the explicit, compiled, and external policies. Reported to the displayed decimal precision, each has expected task loss 0.2285333333, probe cost 0.09, expected Brier score 0.1678812275, shock-conditional recovery 1.0384, and shock-conditional probability 0.09 that both post-shock actions fail. The periodic policy reports 0.2888, 0.09, 0.2147310769, 1.792, and 0.7268, respectively. The change-point policy reports 0.268992, 0.09, 0.1918559015, 1.71136, and 0.669776. Thus the shared Bayesian recursion outperforms the two encoded weaker rules on task loss, score, recovery, and clustered failure while all policies pay the same fixed probe cost. Exact equality of the strong trio is a controlled null for an advantage caused merely by the estimator's label or attributed location.

Timing is important. For the Bayesian trio, step-zero feedback is incorporated before applying the step-one shock hazard. The change-point policy also applies hazard prediction at step one; unlike the Bayesian trio, it performs no likelihood update after residuals. It counts post-feedback contradictions, swaps its current belief components after two consecutive contradictions, and resets the counter. A retained belief $(4/5, 1/5)$ becomes $(13/20, 7/20)$ at the hazard transition, changing encoded confidence from $37/50$ to $31/50$. A separate task-supplied erroneous value $7/20$ is corrected: applying the hazard to $(9/13, 4/13)$ gives polarity-zero probability $31/52$.

The Brier outcome is a squared-probability score for impending action success. Squared loss is proper for a fixed binary event, but a lower cross-policy aggregate score does not alone prove improved calibration because policies can change event probabilities, resolution, and uncertainty. The benchmark elicits no human confidence report and estimates no type-2 sensitivity.

The result is deliberately narrow. It supports persistent Bayesian use of feedback relative to two

specified restrictions, not superiority over periodic recalibration or change detection as method families. It establishes no biological generality, priority, novelty, phenomenal consciousness, or equal engineering cost. An explicit representation has a genuine functional advantage only when, under predeclared matched constraints, it changes the attainable performance or resource frontier relative to both compiled and equally informed external alternatives.

1. Question, scope, and interpretive framework

Control under internal change poses two distinct questions. First, should a controller infer a latent property of its own action channel? Second, does exposing that inference as an explicitly typed self-model add anything beyond the same computation embedded in a compiled policy or supplied by an external estimator? The frozen pilot answers the first question positively only relative to two restricted policies. It answers the second negatively for three deliberately identical mappings. Separating these questions prevents ordinary state estimation from being treated as evidence for a privileged architectural description.

“Internal” is causal here. The polarity θ_t belongs to the mapping from command to output: changing it changes what a fixed command does. The target g_t is sampled externally. A changed target and a changed actuator can both alter task success, but they occur at different modeled locations. The pilot independently samples targets and permits one actuator flip. It includes no latent environmental mode that can imitate the polarity residual. Consequently, body and world variables are distinguished by construction, while inference of the causal source under ambiguous observations remains untested.

We stipulate a minimal functional definition of explicit self-model: an explicitly identified, updateable probability distribution over a latent parameter assigned to the controller’s actuator channel and used for action or confidence. This definition implies neither introspection nor phenomenal consciousness. The Bayesian tracker qualifies because a named belief variable represents uncertainty about actuator polarity. The same statistic, however, can be described as controller state, a compiled function of history, or an external observer’s estimate. If update, decision, and report functions are identical, terminology alone cannot change behavior.

The comparison therefore has two tiers. The strongest controls preserve the complete benchmark interface and the same finite map. The compiled tracker implements the identical history-to-action and history-to-confidence function without relying on the self-model label. The external observer runs the same filter and supplies the same decisions without delay or communication cost. These controls address attribution and location. Periodic and change-point policies instead omit or compress computations. They address the value of a particular use of evidence, not the value of naming that computation explicit.

Three outcomes should be distinguished. A genuine advantage occurs when, under predeclared matched constraints, an explicit policy class reaches an outcome or Pareto frontier unavailable to compiled and external classes. Controlled equivalence occurs when compared systems realize the same relevant mapping. A measured null occurs when no advantage appears on declared outcomes. Here trajectory-wise controlled equivalence entails a null for architectural labeling, even though the weaker rules lose.

The contribution is analytical and expository around inherited artifacts. The manuscript states the finite model, audits timing, reports the supplied exact results, separates body drift from target variation, develops a matched-control interpretation, and compares the model with eight selected literature excerpts. It does not claim priority. The model retains one latent polarity trajectory throughout an episode rather than drawing independent per-step calibration errors. That persistence matters because a residual changes later beliefs, recovery, and clustered-failure

probabilities.

The answer to the title question is conditional. Hazard propagation, persistent evidence integration, factorized transfer, active probing, modular access, and structured reports are possible mechanisms by which explicit organization might matter. Merely enabling one does not guarantee improvement. A genuine claim requires a measured strict advantage under the same declared objective and resource constraints. In this pilot the measured advantage belongs to the full Bayesian recursion relative to the two specified weaker rules, while the explicit label has advantage zero against mappings preserving the same sufficient statistic.

A finite compiled table can reproduce a deterministic finite tracker if enough storage is allowed, but that fact does not establish equal storage, learning cost, verification effort, or transfer performance. Those dimensions become scientific only when fixed and measured. The present benchmark intentionally removes them, making it a boundary test rather than a general architecture tournament.

2. Frozen evidence and bounded literature comparison

The frozen input inventory for this manuscript consists of fourteen supplied sources: four project artifacts, eight primary-literature excerpt files, one analyst comparison, and one manuscript seed. The recorded discovery scope additionally comprises exactly four capped searches: search-1, Crossref, “internal model control adaptive actuator calibration,” five results; search-2, OpenAlex, “sensorimotor calibration actuator drift adaptive control,” five results; search-3, Crossref, “self-model control metacognitive calibration confidence,” five results; and search-4, OpenAlex, “change point detection actuator fault diagnosis control,” five results. These ranked searches are not exhaustive. Search metadata describes scope only; scientific comparisons rely on supplied passages.

The phrase “internal model” does not identify one invariant object. Francis and Wonham’s supplied abstract passage restricts its setting to “linear, time-invariant, finite-dimensional systems” and refers to “a suitably reduplicated model of the dynamic structure of the disturbance and reference signals.” This supports a bounded statement about exogenous disturbance and reference dynamics in that setting. It does not establish a result about binary actuator polarity.

Wolpert, Ghahramani, and Jordan’s author-hosted abstract states that subjects estimated hand location after movements made in darkness, mentions externally imposed forces, and names “optimal state estimation.” The supplied passage therefore concerns hand-location estimation in a sensorimotor setting. It does not establish plant-parameter adaptation, the control-theoretic internal-model principle, or a necessity theorem for the present tracker.

Ortega, Cisneros, Wang, and van der Schaft state that they estimate plant parameters and then compute controller parameters, and that they apply a certainty-equivalent control law. This supports an estimate-then-control description in their stated context. R4 is analogous only in the local structural sense that it estimates actuator polarity and acts using a MAP value. No convergence, stability, or necessity result is imported from the excerpt.

The cross-source synthesis is that neighboring uses target different quantities: exogenous signal dynamics, motor state, and plant parameters. R4 targets a binary parameter in the command-output channel. Similar vocabulary does not establish mathematical identity; different terminology does not establish computational difference. This comparison is the manuscript’s synthesis, not a conclusion attributed jointly to all three papers.

The sensorimotor passages require the same restraint. Taylor, Krakauer, and Ivry distinguish “explicit learning driven by target error” from “implicit learning of a forward model driven by

prediction error,” and the supplied excerpt mentions verbal reports of aiming direction. Tseng and colleagues define sensory prediction errors as occurring when predicted sensory consequences of an initial motor command do not match observed values. These are separate sourced observations.

The project implementation, by contrast, contains no human participant, verbal report, separately measured implicit process, learned forward-model mechanism, or biological retention test. Its residual is a programmer-defined binary observation. Calling the posterior explicit describes a data structure, not explicit human report. The model’s within-episode retention is an encoded computational fact and cannot be generalized to biological retention.

Confidence concepts are also noninterchangeable. Fleming and Lau state that the probability score averaged across estimates is known as the Brier score; they note that calibration and resolution both contribute and discuss the probability that a previous judgment was correct. Murphy’s supplied abstract separately names event uncertainty, forecast reliability, and forecast resolution. Maniscalco and Lau define meta- d' as a measure of type-2 sensitivity and identify its signal-to-noise scale. Each attribution is limited to its own passage.

R4 instead computes, before observing an action outcome, a number intended to predict whether that impending action succeeds. It elicits no subjective confidence rating and estimates no meta- d' . Its mean squared loss is a proper score when the report equals the relevant conditional probability, but an aggregate Brier number does not isolate reliability, resolution, or metacognitive sensitivity.

As a local audit of the eight supplied primary excerpt files, none directly compares an explicit actuator tracker, a history-matched compiled controller, and an equally informed external observer. This is a statement about the frozen excerpt set, not all scholarship. The direct architectural comparison comes from the project implementation and result artifact.

Discovery results include titles concerning adaptive internal-model control, actuator calibration, sensorimotor learning, confidence, and actuator-fault diagnosis. Titles and abstract-only records establish topic proximity, not detailed findings. Nothing found in four five-result searches proves novelty, and nothing absent proves lack of prior art. The manuscript therefore makes no novelty or publication-priority claim.

3. Finite-state calibration model and exact timing

The horizon is $T = 3$, with actions at $t \in \{0, 1, 2\}$. Initial polarity $\theta_0 \in \{0, 1\}$ is fair. Before action, the controller receives three calibration residuals

$$z_i = \theta_0 \oplus \eta_i, \quad \Pr(\eta_i = 1) = \frac{1}{5},$$

where probe errors are independent. Each probe costs 0.03, so every policy pays 0.09. A single shock bit $D \sim \text{Bernoulli}(1/4)$ is independently drawn. The path is

$$\theta_t = \begin{cases} \theta_0, & t = 0, \\ \theta_0 \oplus D, & t \in \{1, 2\}. \end{cases}$$

The shock, when present, persists; it is not redrawn at each step.

At each step, target g_t is an independent fair bit. The policy issues command a_t , execution noise $e_t \sim \text{Bernoulli}(1/10)$ is independent, and output is

$$r_t = a_t \oplus \theta_t \oplus e_t.$$

Success is $Y_t = \mathbf{1}\{r_t = g_t\}$. Feedback exposes

$$x_t = r_t \oplus a_t = \theta_t \oplus e_t.$$

The residual therefore cancels the target and is a direct noisy polarity observation. This clean separation makes the model useful for isolating temporal belief updating but easier than a system in which body and world causes enter the same observation.

For a probability vector $b = (b(0), b(1))$, an observation x , and $0 < \rho < 1$, define

$$U_\rho(b, x)(j) = \frac{b(j)\ell_\rho(x | j)}{\sum_{k=0}^1 b(k)\ell_\rho(x | k)}, \quad \ell_\rho(x | j) = \begin{cases} 1 - \rho, & x = j, \\ \rho, & x \neq j. \end{cases}$$

The denominator is positive for every probability vector under this restricted domain. The benchmark rates $1/5$ and $1/10$ satisfy it. Starting from $(1/2, 1/2)$, three applications of $U_{1/5}$ yield calibration belief c .

Shock prediction is

$$H(b) = \begin{pmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{pmatrix} b,$$

with beliefs read as columns. Let b_t^- be belief immediately before action and b_t^+ belief after residual feedback. For the Bayesian trio,

$$\begin{aligned} b_0^- &= c, & b_0^+ &= U_{1/10}(b_0^-, x_0), \\ b_1^- &= H(b_0^+), & b_1^+ &= U_{1/10}(b_1^-, x_1), \\ b_2^- &= b_1^+, & b_2^+ &= U_{1/10}(b_2^-, x_2). \end{aligned}$$

Thus step-zero feedback precedes the step-one hazard prediction.

The MAP estimate is $\hat{\theta}(b) = 1$ only if $b(1) > b(0)$; otherwise it is zero. Action is $a_t = g_t \oplus \hat{\theta}(b_t^-)$. The encoded pre-outcome report is

$$q_t = 0.9 \max_j b_t^-(j) + 0.1 \min_j b_t^-(j).$$

If b_t^- is the conditional polarity posterior for the available history and the XOR channel with execution error $1/10$ is correct, this equals the conditional success probability: success has probability 0.9 when the MAP polarity is true and 0.1 otherwise. This condition holds for the strong Bayesian trio. It need not hold merely because a weaker policy stores two weights. Periodic can retain $(4/5, 1/5)$ and report $37/50$ at step one even though hazard propagation would imply success probability $31/50$ conditional only on that calibration evidence, without conditioning on step-zero residual feedback. This differs from the change-point witness below, which specifies its own retained state and counter history.

A worked path audits the recursion. Probes $(0, 0, 1)$ yield $c = (4/5, 1/5)$. If $x_0 = 0$, then

$$b_0^+ = \left(\frac{36}{37}, \frac{1}{37} \right).$$

Applying the hazard gives

$$b_1^- = H(b_0^+) = \left(\frac{109}{148}, \frac{39}{148} \right).$$

If $x_1 = 1$, the post-feedback probability of polarity one is

$$\frac{(39/148)(9/10)}{(39/148)(9/10) + (109/148)(1/10)} = \frac{351}{460}.$$

These values give a present worked-path check of the recursion.

A separate correction concerns the task-supplied erroneous value $7/20$. For pre-hazard belief $(9/13, 4/13)$,

$$H \begin{pmatrix} 9/13 \\ 4/13 \end{pmatrix} = \begin{pmatrix} (3/4)(9/13) + (1/4)(4/13) \\ (1/4)(9/13) + (3/4)(4/13) \end{pmatrix} = \begin{pmatrix} 31/52 \\ 21/52 \end{pmatrix}.$$

The predicted probability of polarity zero is therefore $31/52$, not $7/20$. This corrects a task-supplied value, not a publication-history claim.

The periodic policy uses calibration belief c for all actions. It neither applies H nor updates from residual feedback. The change-point policy begins with c , applies H at step one, and otherwise leaves its current belief unchanged except for component swaps. After each action it compares the residual with the pre-action MAP guess. A contradiction increments the counter; a confirming residual resets it to zero. Once the post-feedback count reaches two, the components are swapped and the counter is reset before the next action. Confidence always uses current component magnitudes.

For a static witness, let the calibration belief be $(4/5, 1/5)$ and let step-zero feedback confirm the MAP guess. The change-point belief remains $(4/5, 1/5)$ through that feedback, then the step-one hazard produces

$$H(4/5, 1/5) = (13/20, 7/20).$$

Confidence changes from

$$0.9(4/5) + 0.1(1/5) = 37/50$$

to

$$0.9(13/20) + 0.1(7/20) = 31/50.$$

A later swap preserves the pair of magnitudes, but hazard prediction does not. The reported change-point row belongs to this hazard-aware heuristic, not to an unexecuted no-hazard variant.

Recovery is evaluated only when $D = 1$, which has probability $1/4$. It checks the pre-action estimate. A match at step one scores zero, a first match at step two scores one, and no observed post-shock match scores two. The clustered-failure event is $Y_1 = Y_2 = 0$, also conditional on shock.

4. Policies, matched resources, and equivalence

All five policies are admitted under one common high-level resource envelope. They receive the same probes, target sequence, observation channels and feedback timing, including each policy’s own commands, outputs and residuals, together with the same horizon, noise specification, shock probability and fixed probe budget. Realized commands and outputs can differ across the five policies; equality of those quantities is asserted only for the strong trio. They may retain the finite history needed by their declared algorithms. No policy receives extra training data, fitted parameters, or an undisclosed optimizer. This establishes common access and admissibility, not equality of primitive operations, bits, latency, or energy.

The explicit Bayesian tracker initializes from probes, updates after step-zero feedback, propagates through the known shock hazard, and continues likelihood updating. Its action is certainty-equivalent in the local descriptive sense that it substitutes the MAP polarity into the XOR control rule. This characterization follows from the implementation and does not import an optimality or stability theorem from adaptive-control literature.

The compiled tracker has the identical finite history-to-belief, action, confidence, and update mapping. “Compiled” is a functional comparator, not evidence that a literal table was separately generated or trained. The implementation invokes the same posterior code under a different

policy name. The pilot therefore tests whether relabeling a shared computation changes outputs; it does not compare code size, compression, training efficiency, or synthesis cost.

The external observer also runs the same recursion and supplies the same action and confidence. It receives every relevant signal at the same time and communicates without delay or cost. This idealization isolates attribution to an internal versus external locus. It does not model distributed control, privacy, packet loss, private proprioception, conflicting incentives, or latency. Adding any such feature changes the comparison and should be named as the operative resource difference.

Periodic is intentionally weaker. It preserves its calibration posterior but ignores within-episode feedback and the hazard transition. Its gap from the strong trio is therefore compound. It cannot be attributed solely to periodicity: another periodic controller might propagate the hazard, schedule a new probe, or use event-triggered recalibration.

Change point is also a narrow label. The encoded policy applies the same step-one hazard operator as the strong trio, but it does not perform likelihood updates after residuals. It compresses feedback into a run counter and changes polarity only after two consecutive contradictions. Its gap therefore concerns graded Bayesian evidence use and threshold timing, not omission of the known hazard. It is not representative of every fault detector, likelihood-ratio method, or Bayesian online change detector.

For the strong trio, equality holds path by path. At initialization their beliefs are equal. Assume their post-feedback beliefs agree at some step. Applying the same deterministic hazard or identity transition gives equal next pre-action beliefs; the same MAP rule, target, and confidence formula then give equal actions and reports. Given common sampled execution noise, outputs and residuals agree, and applying the same update gives equal next states. Induction over the three steps establishes identical beliefs, actions, reports, successes, recovery indicators, and clustered-failure indicators for every admissible realization. Expected metrics are consequently equal before any numerical averaging issue arises.

This theorem is stronger than a statistical failure to reject a difference and narrower than unrestricted architectural equivalence. It covers the encoded interface. It says nothing about omitted outcomes such as memory footprint, transfer, interpretability, robustness outside admitted histories, or communication cost.

If the three fixed functions remain identical, reweighting probabilities of already admitted complete histories cannot break their equality. For every admitted history h , their trajectory values agree. Summing any common unconditional statistic against a new law Q preserves equality. For a shock-conditional statistic Z , the corresponding expression is

$$\mathbb{E}_Q[Z \mid D = 1] = \frac{\sum_{h:D(h)=1} Q(h)Z(h)}{Q(D = 1)}.$$

The shared numerator yields equality only when $Q(D = 1) > 0$. If all mass is assigned to no-shock histories, shock-conditional recovery and clustered failure are undefined, although pointwise and unconditional equality persist. The fixed confidence reports need not remain correct conditional posteriors under an arbitrary new law. This extension does not cover new observations, outputs, state constraints, or policy functions.

Every deterministic policy on a finite history domain induces a finite history-to-output table. Thus an unrestricted compiled policy can reproduce a finite explicit tracker. This theorem does not imply equal storage or learning cost. A tabulation may be much larger than a factorized recursion. The present policies have no private randomization: they are deterministic conditional on sampled environment bits. No broader randomized compilation claim is needed.

A fair architectural comparison must distinguish source-backed matching requirements from methodological recommendations. The project brief requires matching information, memory,

computation, training, and feedback. This manuscript additionally recommends predeclaring policy classes, outcomes, reporting interfaces, and resource accounting before seeing results, because outcome-dependent comparator selection can manufacture a victory. The latter recommendation is local methodology, not a quoted project result.

The fair interpretation is asymmetric. Equality among the strong trio isolates the absence of a label or location effect on measured outputs. Differences from periodic identify the joint value of hazard propagation and feedback updating relative to that rule. Differences from change point identify the full Bayesian likelihood recursion and its immediate graded updating relative to hazard-aware thresholded evidence compression. Neither weaker comparison identifies explicit representation alone.

5. Metrics, scoring, enumeration, and provenance

Task loss is the expected failed-action fraction,

$$L = \mathbb{E} \left[\frac{1}{3} \sum_{t=0}^2 (1 - Y_t) \right].$$

Probe cost is fixed at $3(0.03) = 0.09$. Because no policy chooses how many probes to purchase, this column does not estimate a probe-cost frontier.

The confidence outcome is

$$B = \mathbb{E} \left[\frac{1}{3} \sum_{t=0}^2 (q_t - Y_t)^2 \right].$$

It scores a report made after action selection but before the outcome, for the event that the impending action succeeds. Lower values indicate better squared-probability prediction of that event under the evaluated distribution.

Squared loss is proper for a fixed binary event and information state. Let $Y \in \{0, 1\}$, let $p = \Pr(Y = 1 \mid h, a)$, and let $q \in [0, 1]$. Then

$$\begin{aligned} \mathbb{E}[(q - Y)^2 \mid h, a] &= p(q - 1)^2 + (1 - p)q^2 \\ &= q^2 - 2pq + p \\ &= (q - p)^2 + p(1 - p). \end{aligned}$$

The unique minimizer is $q = p$. This is a local derivation, not a theorem imported from the supplied excerpts.

Propriety does not make every encoded baseline report a true conditional probability. The identity requires that the belief used in the confidence formula be the relevant posterior for the available history. It holds for the strong trio under the declared model. Periodic omits hazard and feedback; change point omits likelihood updating. Their reports are still numbers scored by Brier loss, but their stored weights need not be calibrated posteriors.

Nor does a lower cross-policy average Brier score alone prove better calibration. Policies can alter actions and thereby alter event probabilities and irreducible uncertainty. A calibrated constant forecast with $p = q = 1/2$ has expected score $1/4$, whereas a calibrated forecast with $p = q = 9/10$ has expected score $9/100$. Both are calibrated. Their scores differ because event uncertainty differs. Murphy’s uncertainty, reliability, and resolution terminology and Fleming and Lau’s observation that calibration and resolution both contribute provide literature context, but the counterexample itself is local mathematics.

Recovery is separate from action success. Conditional on shock, it records whether the pre-action MAP estimate matches the realized polarity at step one or step two. Execution noise can make a wrong estimate succeed and a right estimate fail. Censoring nonrecovery at two combines late and never-recovered trajectories. Clustered failure is the indicator that both post-shock actions fail. It preserves temporal dependence and must not be reconstructed by multiplying marginal failure rates.

The implementation enumerates initial polarity, three probe-error bits, one shock bit, three target bits, and three execution-error bits. Each complete realization receives its exact product probability. Policy outcomes accumulate with rational arithmetic; floating-point conversion occurs only in the final JSON. The code asserts that total shock mass is $1/4$ and that the three strong result rows are equal.

Complete-history enumeration retains dependence. One shock bit controls both post-shock polarities, probe evidence persists, and each residual can alter later Bayesian beliefs. Replacing this structure with independent per-step marginal errors would generally give incorrect recovery and clustered-failure calculations.

Exact enumeration eliminates Monte Carlo sampling error for the encoded finite distribution. It does not eliminate specification error, programming error, metric-design error, or external-validity uncertainty. This manuscript did not execute the implementation. The protocol reports two local executions on 11 October 2026 that produced byte-identical result files. The supplied implementation SHA-256 is 55018495c5c5a710710fe2f85d358d758921e2a1b64d80895ad3448d7d846321, and the supplied result SHA-256 is 327246f7e186e033c7bd2e75ab14c5447901bf9557dfd58812e5ea9519a97948. These are source-reported original-file provenance, distinct from hashes of extracted text used for a source inventory.

The authoritative contract has horizon three and shock step one. Project sources describe a separate six-step legacy result as unverified and unattached. No value is inferred from it, cited, or merged. A longer horizon would change recovery censoring and the meaning of clustered failure, so the legacy artifact cannot silently extend the result.

The numerical table is reported as inherited source evidence, not a newly verified computation. That evidence classification is a local methodological statement, separate from what the numerical source itself establishes. Arithmetic deductions shown in the manuscript are proofs from displayed numbers, but they are not assigned invented verification identifiers.

6. Supplied results and bounded interpretation

The authoritative result artifact reports the following values, rounded to the displayed decimal precision. Exact strong-trio equality is a separate pointwise result:

Policy	Task loss	Probe cost	Brier score	Recovery [†]	Both fail [†]
Bayesian explicit tracker	0.2285333333	0.09	0.1678812275	1.0384	0.09
Compiled tracker	0.2285333333	0.09	0.1678812275	1.0384	0.09
External observer	0.2285333333	0.09	0.1678812275	1.0384	0.09
Periodic	0.2888	0.09	0.2147310769	1.792	0.7268
Change point	0.268992	0.09	0.1918559015	1.71136	0.669776

[†] Conditional on shock; both fail refers to the two post-shock actions.

Against periodic, the strong tracker lowers task loss by 0.0602666667, approximately 20.87% of the periodic value. Against change point, it lowers task loss by 0.0404586667, approximately 15.04%.

Absolute Brier-score differences are 0.0468498494 and 0.0239746740. These are arithmetic contrasts within the finite distribution, not sampling intervals or population effect sizes.

Shock-conditional recovery is lower by 0.7536 relative to periodic and 0.67296 relative to change point. Because recovery is capped at two, these means mix immediate match, next-step match, and censored nonmatch. They do not determine a long-run recovery curve.

The conditional probability that both post-shock actions fail is lower by 0.6368 relative to periodic and 0.579776 relative to change point. The contrasts are large within the two-action post-shock window but support no conclusion about repeated shocks or stationary reliability.

All policies pay the same 0.09 probe cost. The strong policy therefore does not buy its gain with extra initial probes. It does, however, use feedback and model structure differently, and computation is not priced. The result is not cost-adjusted dominance.

The central architectural result is exact equality. The explicit tracker has measured advantage zero over compiled and external controls on every metric. Because equality holds for every trajectory, this is not an underpowered statistical null. It refutes within the model the proposition that designating the common posterior as an explicit self-model, or attributing it to an internal rather than external locus, is sufficient to improve the declared outcomes.

This null does not rule out gains under changed constraints. A factorized representation might require fewer bits than a table, support a new interface, learn faster, or transfer differently. An internal estimator might receive lower-latency private signals. None of those quantities is present here, so none can explain the supplied equality or be claimed as a demonstrated advantage.

The positive result also needs precise attribution. Relative to periodic, the strong trio adds both hazard propagation and feedback likelihood updates. Relative to change point, both sides use the hazard, while the strong trio adds graded likelihood updating and immediate posterior changes instead of a two-contradiction threshold. Thus the complete Bayesian recursion beats the two complete encoded rules. The design does not separately estimate the contribution of every operation.

A factorial follow-up could add a hazard-only periodic policy, a feedback updater without hazard propagation, and a likelihood-ratio detector with varied threshold. Such policies are proposals, not reported results. They would distinguish transition knowledge, evidence weighting, and decision timing.

The confidence outcome remains bounded. The strong trio has lower squared-probability loss for action success. Since the same posterior drives action and confidence, improved task loss and score are related consequences of the same polarity belief. The artifact does not isolate a separate self-monitoring faculty. It supplies no reliability diagram, resolution estimate, or meta- d' .

The strong-trio pointwise equality and equality of unconditional expectations survive arbitrary reweighting of admitted histories with fixed policy functions. The two shock-conditional metrics remain equal only when the replacement law assigns positive mass to shock; at zero shock mass they are undefined. Fixed forecasts need not remain correct posteriors under the replacement law. It need not survive an expanded interface. A private sensor, message limit, memory cap, or new report can create previously absent causal differences. A resulting advantage should be attributed first to the changed resource or interface and only then, with suitable controls, to representational organization.

The pilot is therefore both positive and null. It is positive about persistent Bayesian estimation relative to two restricted rules. It is null about the self-model label and idealized estimator location. Those findings are compatible rather than contradictory.

7. Conditions for genuine advantage and proposed experiments

Let $\mathcal{E}$, $\mathcal{C}$, and $\mathcal{O}$ be predeclared policy classes with explicit actuator representations, compiled history policies, and external estimators. Fix observation timing, feedback, memory, operation budget, training data, optimizer, probe opportunities, and reporting interface. For objective J , a genuine explicit advantage requires

$$\inf_{\pi \in \mathcal{E}} J(\pi) < \min \left\{ \inf_{\pi \in \mathcal{C}} J(\pi), \inf_{\pi \in \mathcal{O}} J(\pi) \right\},$$

or a strict Pareto improvement when task loss, scoring, probes, computation, and robustness are jointly evaluated. Policy classes chosen after outcomes are known do not identify an architectural effect.

A first prospective mechanism is memory-limited compositional transfer. Train all classes on matched episodes from several actuator families and evaluate held-out combinations of actuator and task. A factorized actuator variable might permit reuse of an update rule under a fixed bit or parameter budget. A flat policy might also discover an equivalent representation, so improvement is a hypothesis rather than a consequence of explicit naming. The study would require fixed train and test generators, equal training examples, a common optimizer, replication for stochastic training, and a compiled comparator selected before test outcomes. The current benchmark includes no learning or transfer.

A second mechanism is active calibration. Permit a policy to buy a probe at cost c , skip it, or act. A posterior makes value-of-information calculations transparent, but a compiled finite policy can implement the same choice. The relevant comparison is attainable loss at each memory and operation budget. Because the pilot fixes three probes for all policies, it cannot estimate an adaptive probe-cost frontier.

A third mechanism is modular intervention. A planner or maintenance module might query, clamp, or transmit an actuator-state estimate. Explicit organization could reduce communication or simplify counterfactual control. To isolate representation, the compiled comparator must receive an equally sized interface or both systems must share a wiring constraint. Otherwise the explicit system has merely been granted an extra output.

A fourth mechanism is source-specific inference. Introduce a latent environmental mode w_t alongside actuator state θ_t . In the proposed design, body and world shocks would be independently randomized, producing neither, body-only, world-only, and joint conditions. Observations would be partially ambiguous, and evaluation would include separate body, world, and action-success forecasts. Independent interventions are a recommendation for this proposed factorial study, not a theorem that every identifiable observational model requires intervention.

A structured posterior $p(\theta_t, w_t \mid h_t)$ could then compete with a capacity-matched recurrent or compiled controller and an observer with the same evidence and latency. Held-out intervention combinations would test source-specific generalization. Merely recovering task performance after undifferentiated errors would not establish correct attribution.

A fifth mechanism is structured reporting. Require separate probabilities for actuator correctness, environmental mode, and next-action success, each scored against a fixed event. This would distinguish state estimation from success probability, which includes execution noise. Reliability and resolution analyses could complement proper scores. A type-2 sensitivity analysis would require suitable joint correctness/confidence observations and a separately specified measurement procedure; Brier loss alone does not supply such an estimate. Whether the existing episode records would suffice for a particular estimator has not been evaluated, and no necessity for a different task is asserted.

External location can matter for identifiable reasons. Private proprioception, communication

delay, bandwidth, packet loss, or conflicting objectives can favor a local controller. These changes violate the idealized equal-access condition. A positive result would first support lower-latency or privileged access, not an unexplained essence of internality. Conversely, an external observer with broader sensors could gain through information rather than architecture.

The next finite study could cross body shock, world shock, active probing, horizon, and transfer condition. It could compare a structured joint filter, a capacity-matched compiled or recurrent policy, an observer with identical evidence, optimized periodic recalibration, and a principled Bayesian change detector. Predeclared outcomes should include task loss; probe, memory, and operation cost; separate proper scores; source-specific recovery; and clustered failure.

A decisive positive pattern would be a reproducible frontier advantage for the explicit class under fixed constraints. Controlled equivalence would recur if strong methods realize the same sufficient statistic. A null would occur if no measured difference remains. If only deliberately weaker policies lose, the inference should continue to favor the omitted evidence processing rather than explicit selfhood.

These proposals are unexecuted hypotheses. They specify what could change the present conclusion, not that explicit organization will win. No proposed functional test licenses an inference to phenomenal consciousness. Programmer-defined latent states and probability reports can be useful without constituting evidence of experience.

8. Critical audit, limitations, and conclusion

The strongest alternative explanation for the positive comparison is straightforward: the winners implement the correctly specified Bayesian recursion and the weaker rules do not. Periodic discards residuals and omits the hazard. Change point applies the hazard but compresses feedback into a contradiction count and waits for a threshold. The experiment therefore varies inference, not merely explicitness, in the weaker-policy contrasts. Equality of the strong trio confirms that the same inference succeeds regardless of its label.

Baseline strength is a major limitation. Periodic recalibration could mean recalibrating at step one, optimizing a schedule, buying an event-triggered probe, or propagating a hazard between scheduled calibrations. Change detection could use posterior odds, cumulative likelihood ratios, run-length posteriors, or optimized stopping. Some stronger variants could coincide with the Bayesian tracker on this finite binary problem. The artifact ranks only its encoded rules.

The residual is unusually diagnostic. Because $x_t = r_t \oplus a_t$, the target cancels and only execution noise obscures polarity. The model therefore bypasses much causal credit assignment. Delayed effects, continuous gains, saturation, hysteresis, coupled actuators, or environmental disturbances are not represented. Those are possible features of future models, not empirical assertions about a supplied physical system.

The horizon is short. There are two post-shock actions. Recovery is capped at two, and clustered failure concerns exactly those actions. A controller recovering later is indistinguishable under this metric from one never recovering. The excluded six-step artifact cannot repair the limitation.

Task and confidence outcomes are coupled because one belief selects the action and supplies the report. This coherence is valid but does not demonstrate two independent faculties. A stronger self-monitoring test would separately score polarity belief and success belief, or hold action quality fixed while varying reports.

Resource matching is high-level. Posterior updating uses arithmetic and normalization; change point uses a counter and swaps. The implementation does not count operations, states, latency, or energy. Equal access is not equal engineering cost. Future studies should predeclare a machine

model or report a performance-resource frontier.

The compiled control is decisive for the encoded functional claim but modest as an engineering comparison. The implementation routes three policy names through the same branch; it does not synthesize an independent lookup table. This proves that naming has no code-level behavioral effect here. It does not compare ease of learning, verification, compression, transfer, or maintenance.

The external observer is idealized. It sees the residual immediately and influences the same action without communication cost. Location has no causal route to affect output under those assumptions. Adding private signals or delay can create a location effect, but the correct explanation would name those resources.

The fixed probe budget prevents a probe-cost tradeoff. Equal cost rules out extra initial probing as an explanation of the table, but it does not show that three probes are optimal. Active calibration could change policy ordering.

The known model favors Bayesian updating. Noise and hazard parameters are supplied exactly and the environment follows them. Misspecification is untested. Perturbing hazards, noise, persistence, or likelihoods would require new computation and evidence.

Tie handling is part of the policy definition: strict posterior dominance is required for estimate one, so exact ties map to zero. For each estimate, consider polarity-one versus polarity-zero odds. The three probes yield exactly $1/64, 1/4, 4, 64$. Step-zero feedback multiplies these odds by 9 or $1/9$, producing no tie. Hazard propagation sends positive odds r to $(3r + 1)/(r + 3)$, which lies strictly between $1/3$ and 3 and equals one only when $r = 1$. Thus the step-one pre-action odds are nonunit; after step-one feedback, multiplication by 9 or $1/9$ still cannot produce one. Periodic retains its nonunit calibration odds, and change point preserves nonunit odds under hazard prediction and reciprocal swaps. Pre-action ties are therefore unreachable for all five encoded policies. An alternative tie rule could matter under a changed contract and should then be predeclared.

The evidence boundary is equally important. The manuscript did not run code, collect data, or inspect complete papers beyond the frozen passages. Reproduction and hash statements are source-reported. The four searches are capped and nonexhaustive. No priority, novelty, publication-history, or empirical-generality inference follows.

Within scope, the conclusion is clear. The full Bayesian tracker improves task loss, squared-probability score, recovery, and two-step clustered failure relative to the specified periodic and hazard-aware two-contradiction rules, at the same fixed probe cost. The mechanism is persistent use of the declared probabilistic model, not a demonstrated essence of selfhood.

The explicit tracker does not outperform the compiled tracker or equally informed external observer. They implement the same trajectory map. The strongest justified architectural conclusion is negative: labeling or relocating this sufficient statistic creates no advantage on the encoded interface. A stronger positive claim requires a predeclared constraint under which explicit organization changes what can be attained, learned, transferred, communicated, or robustly reported.

The pilot is therefore a boundary result. It demonstrates what one finite Bayesian recursion buys against two restrictions and what an architectural label does not buy against strong controls. It supplies no evidence of phenomenal consciousness and no basis for broad biological or engineering generalization.

9. Critical audit and alternative explanations

The pilot supports a deliberately divided interpretation. On the positive side, the full Bayesian recursion has lower supplied task loss, Brier score, shock-conditional recovery time, and two-action post-shock failure probability than the encoded periodic and change-point rules, with the same fixed probe cost. On the negative side, the explicitly named tracker, history-matched compiled tracker, and equally informed external observer are identical on every reported metric. This is not a tension in the findings. The weaker-policy comparisons concern how evidence is processed, whereas the strong-trio comparison concerns whether naming or locating the same computation changes its functional consequences. Within this finite interface, it does not. The appropriate conclusion is therefore a boundary result: persistent probabilistic tracking helps relative to the specified restrictions, but explicit self-model terminology supplies no additional measured advantage.

The baseline details are essential to that conclusion. Periodic retains the calibration belief while omitting both the step-one hazard propagation and all within-episode likelihood updates. Its deficit is consequently compound and cannot be attributed simply to “periodicity.” The change-point policy is different: it does apply the hazard at step one, but thereafter compresses residual evidence into a contradiction counter instead of applying likelihood updates. Two consecutive contradictions trigger a component swap and counter reset. Its current belief magnitudes determine confidence. For example, hazard propagation maps $(4/5, 1/5)$ to $(13/20, 7/20)$, changing confidence from $37/50$ to $31/50$. Thus the change-point comparison isolates neither hazard knowledge nor internal state in general; it contrasts graded Bayesian updating with a hazard-aware threshold heuristic. Stronger periodic schedules, likelihood-ratio detectors, or Bayesian change detectors were not evaluated and could narrow or eliminate the reported gaps.

Matched information must likewise be distinguished from matched implementation cost. All five policies receive the same probes, targets, observation channels and feedback timing, including each policy’s own commands and outputs, with the same horizon, hazard specification and fixed probe budget; none receives extra training data, fitted parameters or an optimizer. Their declared finite states are admissible under the shared high-level envelope, but their realized commands and outputs can differ. They are therefore admissible under one common high-level resource envelope. Yet equal primitive-operation counts, memory bits, latency, energy, and engineering effort are not established across all five. Exact computational identity is stronger only for the explicit, compiled, and external trio, which share the same update and decision branch. The reported advantage over weaker baselines is accordingly an advantage in use of available evidence, not cost-adjusted dominance. Future architectural comparisons should fix information, memory, computation, training, feedback, and reporting interfaces before outcomes are known; otherwise either an unlimited compiled table or an artificially weakened comparator can predetermine the answer.

The strong-trio equality is more informative than a failure to reject a statistical difference. Their states, actions, confidence reports, residuals, and updates coincide for every admitted realization, so metric equality follows before averaging. It remains narrower than unrestricted architectural equivalence. The benchmark does not compare compression, learnability, transfer, modular access, verification effort, or communication. A deterministic finite tracker can always be represented by a finite history table if sufficient storage is allowed, but this observation does not make tabulation equally economical. Similarly, the external observer has immediate access to all relevant signals and cost-free influence over action. An internal estimator could benefit from private or lower-latency observations in another design, but such a result would first establish an access or communication advantage rather than an unexplained benefit of internality.

The proper-score result also requires care. For a fixed binary event with true conditional

probability p , squared loss satisfies

$$\mathbb{E}[(q - Y)^2] = (q - p)^2 + p(1 - p),$$

so the truthful report $q = p$ uniquely minimizes expected loss. In the strong trio, the encoded confidence equals conditional action-success probability when the maintained belief is the relevant posterior under the declared XOR channel. This does not imply that every baseline’s retained weights remain a correct posterior after ignored evidence or transitions. Nor does a lower aggregate Brier score by itself prove better calibration across policies. A calibrated forecast at $p = 1/2$ has expected score $1/4$, while a calibrated forecast at $p = 9/10$ has expected score $9/100$; event uncertainty differs although both reports are calibrated. Policy changes can also alter success probabilities and resolution. The artifact therefore establishes lower squared prediction loss for impending action success, not a complete reliability analysis, subjective confidence measure, or type-2 sensitivity estimate.

The selected literature sharpens these distinctions without resolving the architectural question. Francis and Wonham’s excerpt concerns disturbance and reference dynamics in linear finite-dimensional systems; Wolpert, Ghahramani, and Jordan concern hand-location estimation and optimal state estimation; Ortega and colleagues describe plant-parameter estimation followed by certainty-equivalent control. Comparing these targets with binary actuator polarity is manuscript synthesis, not transfer of a stability, convergence, or necessity theorem. Taylor and colleagues distinguish explicit target-error learning from implicit forward-model learning and mention aiming reports, while Tseng and colleagues define sensory prediction error through mismatch between predicted and observed consequences. R4 contains no participant, verbal report, separately measured implicit process, or biological retention test. Fleming and Lau discuss Brier scoring, contributions from calibration and resolution, and the probability of a previous judgment being correct; Murphy names uncertainty, reliability, and resolution; Maniscalco and Lau define meta- d' as type-2 sensitivity. These concepts are relevant but noninterchangeable with R4’s pre-outcome success forecast.

Literature scope is consequently bounded. The frozen input inventory comprises fourteen sources: four project artifacts, eight primary-literature excerpt files, one analyst comparison and one prior manuscript seed. Recorded discovery is separately limited to four five-result queries: Crossref for “internal model control adaptive actuator calibration,” OpenAlex for “sensorimotor calibration actuator drift adaptive control,” Crossref for “self-model control metacognitive calibration confidence,” and OpenAlex for “change point detection actuator fault diagnosis control.” Ranked metadata establishes topic proximity, not detailed findings. As local corpus-audit synthesis, none of the eight supplied primary excerpts directly compares the explicit, compiled, and external trio, but absence from this small excerpt set is not evidence that no such work exists. The searches cannot establish exhaustiveness, novelty, priority, or a universal architectural distinction.

Model scope imposes further limits. The residual $x_t = r_t \oplus a_t = \theta_t \oplus e_t$ directly exposes noisy polarity and cancels the external target. This cleanly distinguishes actuator drift from target variation by construction but does not test source attribution when body and environmental changes produce ambiguous observations. The horizon contains only two post-shock actions; recovery is censored at two, and clustered failure covers exactly those actions. Noise and hazard are known, probe number is fixed, and misspecification, active calibration, continuous dynamics, repeated shocks, and environmental regime changes are absent. Persistent latent uncertainty is retained correctly across each episode, but no biological persistence follows from that computational property.

Finally, the numerical artifact is source-supported rather than newly verified here. The frozen packet reports two byte-identical reproductions and supplies implementation and result hashes, and this manuscript claims no new computational execution or independent implementation verification. Exact enumeration removes Monte Carlo error for the encoded distribution, not

specification, implementation, metric, or external-validity error. The unattached six-step legacy result remains excluded. These limitations leave the central inference intact but narrow: the supplied three-step model favors full Bayesian evidence use over two encoded restrictions, while providing no evidence that a self-model label, phenomenal consciousness, or explicit representation as such is necessary.

References

- R4 project authors unspecified. “R4 finite-state calibration study protocol — reconciled 2026-10-11.” Frozen source b215242141184c2b8679f99628764e0f, version unspecified.
- R4 project authors unspecified. “R4 evidence and manuscript packet — reconciled 2026-10-11.” Frozen source 4b7ab498a3ff4afcad2bad74aa7aab4, version unspecified.
- R4 project authors unspecified. “R4 exact finite-state benchmark implementation.” Frozen source 895fc79fbb35483ea99f5c1a1e5a97ae, version unspecified; supplied SHA-256 55018495c5c5a710710fe2f85d358d758921e2a1b64d80895ad3448d7d846321.
- R4 project authors unspecified. “R4 deterministic three-step results — reproduced 2026-10-11.” Frozen source a9766c358f02447f987e8ee4918fe5a7, version unspecified; supplied SHA-256 327246f7e186e033c7bd2e75ab14c5447901bf9557dfd58812e5ea9519a97948.
- Wolpert, Daniel M.; Zoubin Ghahramani; Michael I. Jordan. 1995. “An Internal Model for Sensorimotor Integration.” *Science* 269(5232), 1880–1882. DOI 10.1126/science.7569931. Author-hosted abstract excerpts in frozen source 9308e46e377247cc8929300d15181e0b.
- Francis, B. A.; W. M. Wonham. 1976. “The internal model principle of control theory.” *Automatica* 12(5), 457–465. DOI 10.1016/0005-1098(76)90006-6. Publisher-abstract excerpts in frozen source 774960676b4943879ec71d8f542008f0.
- Ortega, Romeo; Rafael Cisneros; Lei Wang; Arjan van der Schaft. 2022. “Indirect Adaptive Control of Nonlinearly Parameterized Nonlinear Dissipative Systems.” arXiv:2201.05749v1. DOI 10.48550/arXiv.2201.05749. Primary-preprint excerpts in frozen source aa1f7c9c73044605a0a7694d6f55cde9.
- Taylor, Jordan A.; John W. Krakauer; Richard B. Ivry. 2014. “Explicit and Implicit Contributions to Learning in a Sensorimotor Adaptation Task.” *Journal of Neuroscience* 34(8), 3023–3032. DOI 10.1523/JNEUROSCI.3619-13.2014. Published-paper excerpts in frozen source 38cb5ad6f7de494793889a8bbb0f7c46.
- Tseng, Ya-Weng; Jörn Diedrichsen; John W. Krakauer; Reza Shadmehr; Amy J. Bastian. 2007. “Sensory prediction errors drive cerebellum-dependent adaptation of reaching.” *Journal of Neurophysiology* 98(1), 54–62. DOI 10.1152/jn.00266.2007. Abstract excerpt in frozen source 67464ba02c76430aaf02a2f89013aae5.
- Fleming, Stephen M.; Hakwan C. Lau. 2014. “How to measure metacognition.” *Frontiers in Human Neuroscience* 8, article 443. DOI 10.3389/fnhum.2014.00443. Methodological excerpts in frozen source cfacd9c49aea483d99293a4799c48f56.
- Maniscalco, Brian; Hakwan Lau. 2012. “A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings.” *Consciousness and Cognition* 21, 422–430. DOI 10.1016/j.concog.2011.09.021. Published-paper excerpts in frozen source a4810c59713a4c528db691d29a451c20.
- Murphy, Allan H. 1973. “A New Vector Partition of the Probability Score.” *Journal of Applied Meteorology* 12(4), 595–600. DOI 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2. Publisher-abstract excerpts in frozen source 8b6abf485c86437b9d0026d2be956cb3.