When Information Fails to Help Control: An Exact Audit of Probing, Beliefs, and Decisions

Research working draft | 3 October 2026

Human authorship and submission metadata pending

Separately rendered reading copy. This PDF reproduces the substantive manuscript from paper.tex using a separate document renderer. Native LaTeX compilation was unavailable in this environment. Equations are rendered as mathematical images; the source remains editable. Figure data are redrawn from the source. This is a research working draft, not a submitted or peer-reviewed publication.

Status: reproducible diagnostic study; not submitted or peer reviewed.

Abstract

Prediction error, parameter information, and decision quality measure different properties of a world model. We build a small, executable audit that separates these properties with exact reference answers. In a two-stage environment, hidden actuator signs remain fixed while an agent chooses at most one noisy probe and then a continuous terminal control. Classical Bayesian decision theory gives the optimal control, the task value of a probe, and a decomposition of excess loss into probe-selection regret and posterior-mean error. We instantiate Gaussian and correlated binary observation channels, including a fitted parametric belief model. Across 30 Monte Carlo seeds of 5,000 episodes, a deliberately misaligned Gaussian instance has mean total loss 0.50937 for known-model task-value probing versus 1.01000 for parameter-information probing with task-aware abstention; the paired difference is -0.50063 with a 95% seed-bootstrap interval $[-0.50459, -0.49684]$. In a correlated binary instance, a fitted joint belief improves over a fitted independent belief, but can lose near the probe-selection boundary. Overconfident posterior estimates can make probing worse than doing nothing. These are controlled diagnostics of established principles, not evidence of a new foundation model, a new exploration principle, or general superiority to existing world models.

1. Motivation and scope

An action-conditioned predictor can be accurate on its training distribution while remaining unreliable for decisions outside that distribution. Conversely, a model need not reproduce every environmental detail to support a particular decision. These distinctions are central to controlled identifiability, value equivalence, and task-oriented exploration [5, 6, 2]. They motivate a practical question: when a learned controller fails to use an observation profitably, did it choose the wrong experiment, infer the wrong belief, or turn uncertainty into the wrong control?

The answer is difficult to identify from end-to-end return alone. An experiment can reduce uncertainty about an irrelevant variable. A nominally irrelevant variable can become useful through correlation. A model can report a large expected benefit while using an overconfident posterior. A hard latent-state estimate can produce worse squared control loss than a posterior mean even with the same observation. A transparent environment with known optimal answers can isolate these possibilities before expensive visual training.

This study supplies three concrete artifacts: an exactly characterized environment family; executable oracle and fitted-belief comparisons with frozen internal protocols, raw seed results, and numerical checks; and tests in which the apparent advantage disappears or reverses. The propositions below are specializations of classical Bayesian decision theory. We do not claim to introduce task-oriented experiment design, dual control, belief states, or the distinction between predictive and interventional accuracy. Nor do we claim that these small environments are sufficient evidence for any modern visual world model's performance.

2. Relationship to existing work

Decision-oriented experiment design.

Huang et al. optimize experimental design for downstream decision utility and learn amortized experimental and decision policies [1]. Wagenmaker et al. provide task-dependent exploration guarantees for linear and nonlinear control [2, 3]. Go et al. directly optimize a decision objective under model uncertainty [4]. These works are immediate antecedents, rather than algorithms surpassed by our toy calculation. We do not run their implementations and do not claim a performance comparison with them.

Beliefs, sufficient models, and world models.

Predictive state representations summarize action-observation tests [7]; learned belief representations address task-relevant hidden state under partial observability [8]. Value equivalence formalizes modeling what matters for planning [6]. Recent controlled-world-model theory separates representation identification from identification of action effects under explicit assumptions [5]. The discrete counterexamples here are consistent with that agenda but do not replace its guarantees.

Exploration and control.

Dual control considers the coupled informational and physical effects of actions [9]. Our probes have no residual effect on terminal physical state, so the present environment is more specifically an experiment-then-control problem. Plan2Explore learns exploration using world models before downstream tasks are given [10]; our task is known before probing. Comparing the two as if their information budgets and objectives were identical would be inappropriate. A future learned-control study also needs carefully tuned recurrent baselines [11].

3. Environment and information available to the agent

An episode samples a persistent sign vector $\theta \in \{-1, +1\}^d$. Given a goal $g \in \mathbb{R}^d$, terminal control $u \in \mathbb{R}^d$ produces the deterministic response $x_j = \theta_j u_j$ and loss

$$L(\theta, u; g) = \sum_{j=1}^{d} (\theta_j u_j - g_j)^2. \tag{1}$$

Before terminal control, the agent may choose one probe a with known nonnegative cost c_a, observe $Y \sim p_a(\cdot \mid \theta)$, and update its belief. It may instead stop, incurring no probe cost. A probe leaves no unmodeled residual physical effect. Signs are redrawn between episodes, not between probing and control. The objective is expected terminal loss plus the actual probe cost.

Controls are unconstrained; the results also hold in a box containing all $g_j m_j$ for $m_j \in [-1, 1]$. The primary Gaussian experiments grant every nonprivileged policy the same true prior, observation likelihood, goal, costs, and one-probe budget. They differ only in probe selection or the stated terminal rule. A privileged reference receives θ itself, uses $u = \theta \odot g$, and has zero loss. It is an unattainable reference, not a fair learning baseline.

The binary experiments distinguish known-model references from fitted models. Fitted models receive independent calibration trajectories. Those trajectories reveal both signs through noiseless terminal unit actions and reveal both probe outputs. This strong training access is explicit; at test time only the selected scalar probe output reaches the belief/controller interface. Test signs and unselected outputs are used only by the evaluator.

4. Exact reference identities

4.1 Passive prediction does not certify action identification

Proposition 1: Passive ambiguity

Suppose passive histories H are independent of a uniform sign vector θ, as occurs when all behavior actions are zero and sensor noise is sign-independent. Then the best passive Bayes terminal loss is $\|g\|_2^2$. The same value is the passive minimax loss, while the sign-informed optimum is zero.

Proof.

For any randomized U based only on H and independent private randomness, the cross terms vanish:

$$\mathbb{E}L(\theta, U; g) = \mathbb{E}\|U\|_2^2 + \|g\|_2^2 \geq \|g\|_2^2.$$

The constant rule $U = 0$ attains equality for every sign vector. Worst-case loss is at least uniform-prior average loss, so the same rule attains the minimax lower bound. If signs are known, $U_j = \theta_j g_j$ attains zero.

Noiseless zero-action responses are nonetheless perfectly predictable. Appending nuisance state z with covariance I_k and dynamics $z_{t+1} = z_t$ gives a representation $h = z$ that predicts perfectly on the passive observation distribution, with full covariance rank and effective rank k, but no sign information. Here effective rank means the exponential entropy of normalized covariance eigenvalues. This is an existence counterexample; it does not say that high rank is harmful or that passive data with informative action variation cannot identify dynamics.

For a quantitative coverage diagnostic, compare $f(a) = a$ with $\tilde{f}(a) = -a$ under behavior $a = \varepsilon Z, Z \sim N(0, 1)$. Their prediction mismatch is $4\varepsilon^2$, while their squared disagreement at unit action is 4. For $\varepsilon > 0$, the models are distinguishable in principle. The statement is about weak excitation and error amplification, not impossibility at every nonzero excitation.

4.2 Control and probe value

Let $m_j(H) = \mathbb{E}[\theta_j \mid H]$ under the true model, allowing correlations.

Proposition 2: Bayes control and one-probe value

The optimal terminal control and risk are

$$u_j^*(H) = g_j m_j(H), \qquad \mathcal{R}(H) = \sum_j g_j^2 (1 - m_j(H)^2). \tag{2}$$

For $m'_j = \mathbb{E}[\theta_j \mid H, a, Y]$, the expected reduction in terminal Bayes risk is

$$\Delta(a \mid H) = \sum_j g_j^2 \mathbb{E}[(m'_j - m_j)^2 \mid H, a]. \tag{3}$$

The optimal single-probe rule chooses the largest positive $\Delta(a \mid H) - c_a$, and stops if none is positive.

Proof.

Complete the square in the conditional loss:

$$\mathbb{E}[L \mid H, u] = \sum_j \{(u_j - g_j m_j)^2 + g_j^2 (1 - m_j^2)\}.$$

The tower property gives $\mathbb{E}[m'_j \mid H, a] = m_j$. Substituting into the difference of pre- and post-observation Bayes risks proves (3). Comparing the resulting expected total costs, including stopping, proves single-probe optimality.

Equation (3) is the law of total variance with task weights. If signs are correlated, a probe can change several posterior means; reducing the sum to the probed coordinate is generally wrong. Expected raw information value is nonnegative for a correctly specified optimal decision maker, but it can be smaller than acquisition cost. A particular realized observation need not lower conditional risk. The proposition does not establish greedy optimality for multiple probes.

4.3 Separating selection and belief error

Fix a pre-probe history, goal, and true test distribution. Let $J_B(a)$ be true expected loss plus cost when probe a is followed by the true Bayes controller, including a no-probe option, and let $J^* = \min_a J_B(a)$. A fitted model, conditioned on its independent training data D, deterministically chooses $\hat{a} = \hat{a}(D)$ and uses $\hat{u} = g \odot \widehat{m}$. A randomized selector obeys the corresponding identity after also averaging over its independent selection randomness.

Corollary 1: Exact diagnostic decomposition

For each fixed training dataset,

$$J(D) - J^* = [J_B(\hat{a}) - J^*] + \mathbb{E}\left[\sum_j g_j^2 (\widehat{m}_j - m_j)^2 \mid D, \hat{a}\right]. \tag{4}$$

First term: probe-selection regret. Second term: posterior-mean error penalty.

Proof.

Apply the completed-square identity after the selected observation and average over fresh test episodes conditional on D. Add and subtract $J_B(\hat{a})$.

The posterior penalty includes approximation or misspecification error; ordinary reliability calibration alone does not characterize it. If an arbitrary controller differs from $g \odot \widehat{m}$, its gap from the true Bayes action can still be measured, but separately adding its squared error and the posterior penalty introduces a cross term. We make no three-way additive claim without that term.

5. Channels, fitted models, and counterexamples

5.1 Gaussian probes

Independent uniform signs observed through $Y_j = \theta_j + \varepsilon_j$, with $\varepsilon_j \sim N(0, \sigma_j^2)$, have posterior

$$m'_j = \tanh(Y_j/\sigma_j^2), \qquad A(\sigma_j) = \mathbb{E}_{Z \sim N(0,1)} \tanh^2(\sigma_j^{-2} + Z/\sigma_j). \tag{5}$$

The task gain is $g_j^2 A(\sigma_j)$. We compute it and mutual information by adaptive one-dimensional quadrature, truncated to $Z \in [-12, 12]$ with absolute and relative tolerance 10^{-11}. The omitted standard-normal mass is below 4×10^{-33}; quadrature values are numerical references, not symbolic exact numbers. A separate integral formulation and Monte Carlo simulations check the implementation.

The primary configuration has $g = (1, 0.1)$, $\sigma = (1, 0.25)$ and both costs 0.05. Posterior second moments are approximately $(0.55040049, 0.99990178)$; parameter mutual informations are $(0.33683082, 0.69305365)$ nats. Thus the more informative second channel has net task value -0.04000098, while the first has net value 0.50040049. The cost-aware information baseline first selects the channel of maximal mutual information and abstains when that selected channel has nonpositive true task value. It does not re-rank channels by task value, which would reproduce the task rule in this environment.

5.2 Binary channels and correlation

A binary symmetric probe observes $Y_j = \theta_j \eta_j$, with reliability $\rho_j = \mathbb{E}\eta_j = 1 - 2\alpha_j$. With independent uniform signs its task gain is $g_j^2 \rho_j^2$ and parameter information is $1 - h_2(\alpha_j)$ bits. For $g = (1, 0.1)$ and $\rho = (0.6, 0.9)$, the two no-cost terminal risks are exactly 0.65 and 1.0019; information maximization selects the latter.

For correlated signs, use

$$P(\theta_1, \theta_2) = \frac{1 + \gamma\theta_1\theta_2}{4}, \qquad -1 \le \gamma \le 1. \tag{6}$$

After probing coordinate 1, posterior means are $(\rho_1 Y, \gamma\rho_1 Y)$; probing coordinate 2 yields $(\gamma\rho_2 Y, \rho_2 Y)$. Consequently

$$\Delta_1 = \rho_1^2(g_1^2 + \gamma^2 g_2^2), \qquad \Delta_2 = \rho_2^2(\gamma^2 g_1^2 + g_2^2). \tag{7}$$

With $g = (1, 0)$, $\rho = (0.6, 0.9)$ and equal costs, probe 2 is optimal when $|\gamma| > 2/3$. It has zero direct goal weight yet reveals the relevant sign through correlation. Parameter information still favors channel 2, and now agrees with task value at sufficiently strong correlation. This rules out a claim that parameter information must always fail.

The fitted joint model estimates $\widehat{\gamma} = n^{-1}\sum_i \theta_{i1}\theta_{i2}$ and $\widehat{\rho}_j = n^{-1}\sum_i \theta_{ij} Y_{ij}$ from calibration trajectories. The fitted independent model uses the same reliability estimates and fixes $\widehat{\gamma} = 0$. Both choose by their estimated task value and apply their estimated posterior mean. These are moment estimators in a tiny, correctly specified family (except for the independent model under correlation), not neural or visual world models.

5.3 Overconfidence can reverse the benefit

For one uniform sign with true reliability ρ and assumed reliability $\widehat{\rho}$, the plug-in controller has terminal risk

$$\mathbb{E}L = g^2(1 + \widehat{\rho}^2 - 2\rho\widehat{\rho}) = g^2(1 - \rho^2) + g^2(\widehat{\rho} - \rho)^2. \tag{8}$$

At $g = 1$, $\rho = 0.2$ and $\widehat{\rho} = 0.9$, it incurs 1.45, versus 1 for no probe and 0.96 for the correctly specified probe controller, all before cost. Thus maximizing an estimated value does not guarantee real benefit. This stress case is an analytic misspecification example, not an observed deployment failure.

6. Experimental protocol

Gaussian diagnostic.

The internal protocol was written before the first outcomes. Fourteen fixed configurations vary cost, relative noise, and goal weights. Seven policies include no probe, uniform random probe, full-parameter information (always probe), information with task-aware abstention, task value with posterior-mean control, task value with hard MAP control, and the privileged reference. The hard-MAP variant uses the same probe as the task-value selector but replaces the observed posterior mean by its sign; unobserved coordinates retain zero controls. There is no trained model in this experiment.

Each configuration uses 30 independent seeds and 5,000 episodes per seed. Hidden signs, potential probe noises, and random-probe indices are paired across methods. The same random draws also support configuration sweeps, so configurations are not independent replications. There are 2,940 method-configuration-seed records; seven policy evaluations on an episode are not seven independent samples.

Fitted binary diagnostic.

We use $\gamma \in \{0, 0.5, 0.65, 0.8, 1\}$ and $n \in \{64, 256, 1024\}$ calibration trajectories, 30 independent calibration/test seeds per cell, and 5,000 test episodes per seed. Here $g = (1, 0)$, $\rho = (0.6, 0.9)$ and costs are 0.05. Training and test random streams are distinct, and test data do not fit any parameters. The primary fitted-model comparison is joint minus independent belief at $\gamma = 0.8, n = 256$. Known-joint Bayes and information policies know true parameters and are labeled as oracle references. They do not have the same model knowledge as fitted models.

Statistics and reproducibility.

Intervals use 10,000 percentile-bootstrap resamples of the 30 seed means, with paired resampling for differences. Gaussian intervals quantify Monte Carlo uncertainty; binary fitted-model intervals additionally reflect random calibration samples. They do not measure uncertainty about generalization to other environments or scientific novelty. Secondary sweeps are descriptive without multiplicity correction. There is no seed exclusion or result-dependent hyperparameter selection. Internal protocol freezes are not external preregistration. The selection/posterior decomposition was added after the initial outcomes and is marked as a descriptive post-run analysis.

The release includes the scripts, frozen JSON protocols, per-seed CSVs, JSON summaries, pinned Python dependencies, generated figures, and source/protocol hashes. Calculations ran on CPU using Python 3.12 and NumPy/SciPy. No pretrained visual backbone, robot dataset, GPU training, or paid compute was used in these experiments.

7. Results and failure cases

7.1 Objective mismatch under matched model knowledge

Policy	Mean total loss	95% seed interval	Numerical reference
No probe	1.01000	[1.01000, 1.01000]	1.01000
Uniform random probe	0.78031	[0.77698, 0.78366]	0.77980
Information, always probe	1.05000	[1.05000, 1.05000]	1.05000
Information + task abstention	1.01000	[1.01000, 1.01000]	1.01000
Task value + posterior mean	0.50937	[0.50541, 0.51316]	0.50960
Task value + hard MAP	0.69003	[0.68256, 0.69701]	0.69462
Privileged sign reference	0	[0, 0]	0

Table 1. Gaussian primary configuration. The cost of a taken probe is 0.05. Information and task selectors have identical true model access. The privileged reference has extra information. The very narrow information-always interval collapses under displayed rounding; full precision and separate cost/terminal-loss columns are in the machine-readable release.

The primary paired difference, task-value minus information-with-abstention, is -0.50063 with interval $[-0.50459, -0.49684]$. This is the intended distinction between objectives in a constructed instance. It is not a claim that a learned acquisition algorithm has beaten another learned algorithm. The hard-MAP comparison isolates a second issue: retaining posterior uncertainty matters for the quadratic decision, even after the probe choice is fixed.

Equal task weights, reversed noise, and several equal-noise settings produce agreement between information and task selectors. At high costs or very weak channels, both cost-aware methods abstain. For tied information scores, the implementation uses the first coordinate; agreement in the equal-noise case therefore partly depends on the

documented tie rule. The passive coverage diagnostic reaches prediction mismatch approximately 4×10^{-8} at $\varepsilon = 10^{-4}$ while retaining unit-action disagreement 4; this compares specified candidate functions and is not a fitted-model result.

7.2 Correlation helps only if the belief model uses it

γ	Known joint Bayes	Fitted joint	Fitted independent	Information + abstention
0.00	0.69170	0.69397	0.69397	1.00000
0.50	0.68920	0.69079	0.69079	0.84918
0.65	0.69643	0.70058	0.69821	0.70487
0.80	0.53481	0.54110	0.69255	0.53481
1.00	0.24026	0.24102	0.69510	0.24026

Table 2. Mean total test loss with 256 independent calibration trajectories per seed. Known joint Bayes and information references receive true channel/prior parameters; fitted models receive matched calibration data. Monte Carlo differences from exact population references are expected.

At $\gamma = 0.8$, the fitted joint and independent means are 0.54110 and 0.69255. Their paired difference is -0.15145 with 95% interval $[-0.16067, -0.13841]$. The joint model selects the population-optimal second probe in 29 of 30 calibration draws. The known-model information baseline is optimal here and attains the same episode outcomes as known-joint Bayes. Hence the study does not support universal dominance over information acquisition.

The fitted joint model can also lose to the simpler fitted independent model. At $\gamma = 0.65, n = 256$, its mean loss is 0.70058 versus 0.69821; ten calibration draws select the wrong second probe. These descriptive means do not establish a statistically significant reversal. More directly, their exact fresh-test regrets averaged over those 30 calibration draws are 0.00877 and 0.00205. At $\gamma = 0.5, n = 64$, the corresponding regrets are 0.03420 and 0.00746. Estimating a correlation can introduce selection errors when the task-value ranking is close or sample size is small.

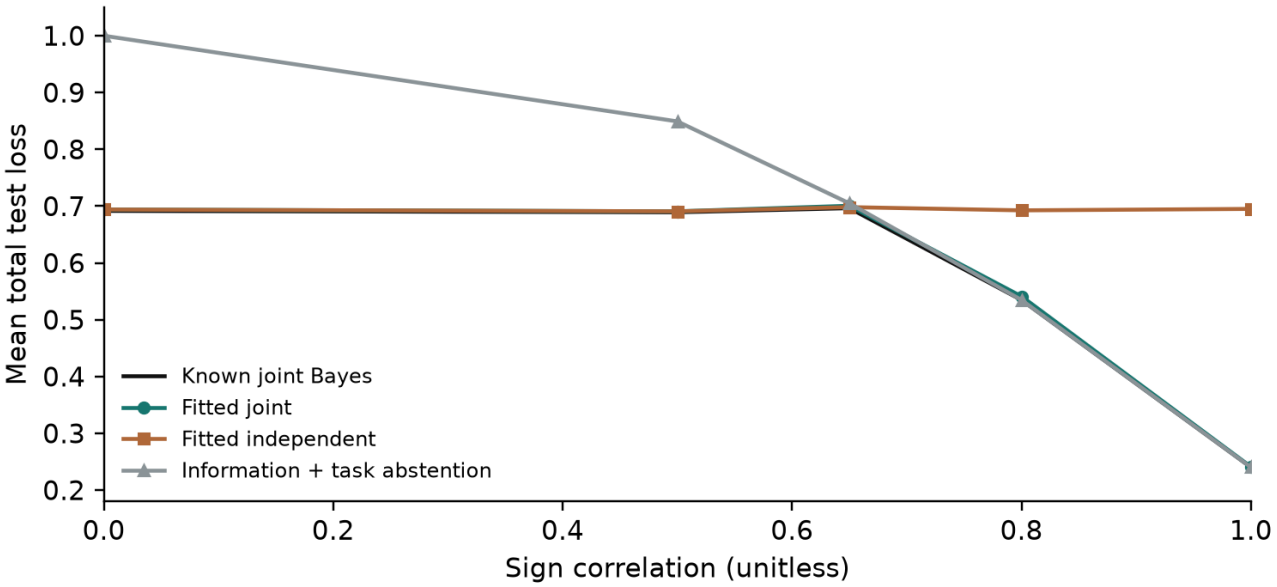


Figure 1. Synthetic binary audit, 30 seeds × 5,000 test episodes, 256 calibration trajectories per seed; generated 3 October 2026. Points are sample means; connecting lines guide the eye and do not assert continuous measured performance. Full confidence intervals are in the release.

7.3 Where fitted-model regret comes from

Calibration n	Fitted belief	Selection regret	Posterior penalty	Total expected regret
64	Joint	0.015840	0.009808	0.025648
64	Independent	0.158400	0.010378	0.168778
256	Joint	0.005280	0.001850	0.007130
256	Independent	0.158400	0.002874	0.161274
1024	Joint	0	0.000373	0.000373
1024	Independent	0.158400	0.000706	0.159106

Table 3. Post-run decomposition at $\gamma = 0.8$. Entries are exact fresh-test quantities averaged over the 30 realized calibration datasets, not an exact expectation over every possible training dataset. The population known-model optimum is 0.5316. Components sum before rounding.

The independent model's main limitation is structural probe selection, while its estimated channel reliability improves with more calibration data. The joint model reduces both errors in this configuration. This conclusion is narrower and more useful than interpreting aggregate return alone as evidence that a representation has learned general physics.

7.4 A stronger control: information about the same task variable

The primary Gaussian reversal could be dismissed as measuring irrelevant parameters. We therefore added an exploratory channel-ordering experiment after reviewing the initial diagnostic, and froze its protocol before collecting its own outcomes. There is only one task-relevant sign, with $g = 1$, and both candidate channels observe that same sign. Channel A is a BSC with $\rho = 0.6$. Channel B is an erasure channel: it reveals the sign perfectly with probability p and outputs zero otherwise. Each costs 0.05 and both use exact posterior-mean terminal control.

Channel A has information $1 - h_2(0.2) = 0.2780719051$ bits, posterior second moment 0.36, and total risk 0.69. Channel B has information p bits, posterior second moment p, and total risk $1.05 - p$. Thus, whenever

$$1 - h_2(0.2) < p < 0.36,$$

information and quadratic task value rank the channels oppositely, even though the uncertain variable is identical. This is a classical difference between scoring objectives, not a newly discovered information principle. Outside the disagreement interval the selectors can agree; at $p = 0.36$ the two expected control risks tie.

We evaluate $p \in \{0.2, 0.2780719051, 0.3, 0.36, 0.5\}$ with 30 paired seeds of 5,000 episodes each. At $p = 0.3$, the information selector chooses erasure and has mean total loss 0.74992 (95% interval $[0.74789, 0.75200]$), while the task selector chooses the BSC and has 0.690672 ($[0.68595, 0.69546]$). The paired difference is -0.059248 with interval $[-0.063759, -0.054446]$, consistent with the exact difference -0.06. These intervals are descriptive evidence for the exploratory extension, not a new confirmatory test. This control removes nuisance-variable selection as the sole explanation, while preserving the study's narrow known-model scope.

8. Limitations and the next falsifiable experiment

The environments contain at most two hidden bits, a known observation family, non-state-changing probes, a known quadratic objective, and one terminal control. Their references exploit these assumptions. There are no image observations, contact dynamics, long horizons, multi-agent interactions, open-world events, or actual robots. Calibration trajectories expose hidden signs; a realistic model may not obtain such supervision. The fitted independent comparison tests a deliberately restrictive model family. It cannot establish superiority to flexible learned recurrent beliefs or existing world-model architectures.

The reported benefit in the Gaussian case is constructed by task/noise misalignment. Sweeps and correlated counterexamples expose this dependence but do not turn a toy demonstration into a broad empirical result. Source review found close antecedents; neither the equations nor the general acquisition principle establish novel theory. No foundation-model benchmark or claim of historical impact is supported.

A meaningful next study should retain separate selector, belief, and controller interfaces, then use a learned recurrent belief model in a state-changing partially observed control environment. A minimal design would fix a recurrent architecture and training budget, swap known versus learned beliefs into identical selectors, compare decision utility to full-parameter and task-variable information, and include a strong recurrent model-free baseline. Task, noise, dynamics, and goal shifts must be held out by complete trajectories. Reporting should include acquisition cost, failure severity, posterior-mean error where identifiable, held-out task loss, and all failed regimes. Exact small-environment oracles should remain unit-level diagnostics rather than substitutes for those tests.

The immediate scientific question is therefore not whether this study replaces a successful architecture, but whether its diagnostics predict which trained beliefs will select useful real actions. That question remains unanswered here.

Reproducibility and assistance disclosure

The accompanying artifact uses `benchmark.py`, `belief_audit.py`, `channel_ordering.py`, and `analyze_evidence.py`, with their JSON protocols and per-seed data. The LaTeX source embeds its figure data and bibliography and is intended as a standalone document; local compilation was not verified because the built-in compiler reported an environment error. The study was drafted, implemented, executed, and checked with AI assistance. Multi-agent cross-checks are not independent human peer review. Human authorship, responsibility for the manuscript, and submission metadata remain to be established. No manuscript has been submitted and no publication or acceptance is represented.

A. Nonuniform binary beliefs and repeated probes

For one sign with prior mean m and BSC reliability ρ, supported observations obey

$$P(Y=y)=\frac{1+y\rho m}{2}, \qquad m'(y)=\frac{m+y\rho}{1+y\rho m}.$$

Direct substitution yields

$$\Delta = g^2\frac{\rho^2(1-m^2)^2}{1-\rho^2m^2}.$$

At a degenerate endpoint $|m|=|\rho|=1$, the supported-observation value is zero; the displayed fraction should not be evaluated as $0/0$. Repeated independent informative probes can identify a persistent sign. For n BSC probes with $0<\rho\le 1$, a majority-sign controller has error at most $\exp(-n\rho^2/2)$ by Hoeffding's inequality, giving terminal loss at most $4g^2\exp(-n\rho^2/2)$. Bayes posterior-mean control is no worse. Known negative reliability can be handled by flipping each observation before majority voting. This bound ignores acquisition cost and

assumes the latent sign persists. It does not justify probing indefinitely in a costly environment.

B. Multiple probes require a different oracle

For a belief b, terminal Bayes risk $R(b)$, and at most k probes, the exact finite-horizon recursion is

$$V_0(b) = R(b), \qquad V_k(b) = \min\left\{R(b), \min_a \left[c_a + \mathbb{E}_Y V_{k-1}(b^{a,Y})\right]\right\}.$$

The one-step value used in this study is the $k = 1$ case. A multi-step extension should compare to this recursion when tractable; reusing a greedy single-step selector and calling it globally optimal would be unsupported.

References

[1] D. Huang, Y. Guo, L. Acerbi, and S. Kaski. Amortized Bayesian Experimental Design for Decision-Making. *NeurIPS*, 2024. arXiv:2411.02064.

[2] A. Wagenmaker, M. Simchowitz, and K. Jamieson. Task-Optimal Exploration in Linear Dynamical Systems. *ICML*, PMLR 139:10641-10652, 2021. Primary proceedings.

[3] A. Wagenmaker, G. Shi, and K. Jamieson. Optimal Exploration for Model-Based RL in Nonlinear Systems. *NeurIPS*, 2023. arXiv:2306.09210.

[4] J. Go, X. Qian, and B.-J. Yoon. Goal-driven Bayesian Optimal Experimental Design for Robust Decision-Making Under Model Uncertainty. Preprint, 2026. arXiv:2605.26093.

[5] X. Zhang, Y. Guan, B. Zhang, H. Li, Y.-Q. Zhang, and S. E. Li. On the Identifiability of Controlled World Models. Preprint, version 2, 27 July 2026. arXiv:2607.22430v2.

[6] C. Grimm, A. Barreto, S. Singh, and D. Silver. The Value Equivalence Principle for Model-Based Reinforcement Learning. *NeurIPS*, 2020. arXiv:2011.03506.

[7] M. L. Littman, R. S. Sutton, and S. Singh. Predictive Representations of State. *NIPS*, volume 14, 2001. Primary paper.

[8] A. Wang, A. C. Li, T. Q. Klassen, R. Toro Icarte, and S. A. McIlraith. Learning Belief Representations for Partially Observable Deep RL. *ICML*, PMLR 202:35970-35988, 2023. Primary proceedings.

[9] E. D. Klenske and P. Hennig. Dual Control for Approximate Bayesian Reinforcement Learning. *JMLR*, 17(127):1-30, 2016. Primary journal article.

[10] R. Sekar, O. Rybkin, K. Daniilidis, P. Abbeel, D. Hafner, and D. Pathak. Planning to Explore via Self-Supervised World Models. *ICML*, PMLR 119:8583-8592, 2020. Primary proceedings.

[11] T. Ni, B. Eysenbach, and R. Salakhutdinov. Recurrent Model-Free RL Can Be a Strong Baseline for Many POMDPs. *ICML*, PMLR 162:16691-16723, 2022. Primary proceedings.