What Capable Agents Must Know

Selection Theorems for Robust Decision-Making under Uncertainty

Aran Nayebi • Carnegie Mellon University

Machine Learning Department • Neuroscience Institute • Robotics Institute

uai2026

UAI 2026 Amsterdam

August 17–21, 2026

CLASSICAL VIEW

Belief states and world models *can* implement optimal control.

⟹

THIS PAPER: FROM "CAN" TO "MUST"

Low average-case regret forces the predictive internal structure tested by the task family.

⟹

SELECTION

Competence compresses the space of possible representations.

Three questions this paper resolves

Do capable agents *need* world models?

Yes, for the right long-horizon diagnostics. Even stochastic policies with low *average* regret encode an approximate transition model.

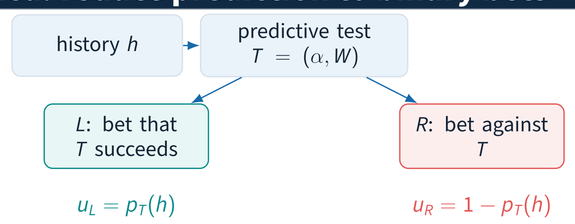
Does necessity survive partial observability?

Yes. Low regret forces predictive state and belief-like memory, resolving the open problem posed by Richens et al. (2025).

What else does the task distribution select?

Block structure selects modularity; shifting mixtures select persistent regime state; shared minimal tasks select representation match.

Main idea: reduce prediction to binary bets



$$\delta = w\frac{4m}{1+2m}, \qquad m = \left|p_T(h) - \tfrac{1}{2}\right|.$$

If the test is not a coin flip ($m \geq \gamma$), then

$$\boxed{w \leq \delta/c(\gamma)}, \qquad c(\gamma) = \frac{4\gamma}{1+2\gamma}.$$

So low regret limits the mass placed on the wrong bet. To keep winning, memory must preserve the predictive distinctions the tests expose.

Theorem 1: long-horizon competence selects a model

Diagnostic goal $G^{(n)}_{s,a,s',k}$. Let $X_i = 1$ when the ith execution of (s,a) transitions to s' and $N_n = \sum_{i=1}^{n} X_i$. At $t=0$, the agent commits to L (predict $N_n \leq k$) or R (predict $N_n > k$); the goal succeeds exactly when the chosen branch is true.

Across these threshold goals, the agent's betting probabilities yield an estimator $\widehat{P}$ with

$$\mathbb{E}\left[\left|\widehat{P}_{ss'}(a) - P_{ss'}(a)\right|\right] \leq \frac{t_\gamma}{\sqrt{n}} + \frac{\bar{\delta}}{c(\gamma)} + O\left(\frac{1}{n}\right).$$

model error $= O(n^{-1/2}) + O(\bar{\delta})$

Longer-horizon competence demands tighter dynamics estimates.

Myopic competence ($n=1$) need not force a world model.

Notation at a glance

- h: interaction history. $T = (\alpha, W)$: take action sequence α and test whether the future observations lie in event W. $p_T(h)$: its true success probability after h.
- π: agent policy; V and V^*: achieved and optimal goal-success probabilities; $\delta_T(\pi; h) = 1 - V/V^*$: normalized regret on test T after h. Also, w: probability mass on the suboptimal bet; $m = |p_T(h) - \frac{1}{2}|$: betting margin; γ: chosen minimum margin.
- $c(\gamma) = 4\gamma/(1+2\gamma)$ and $t_\gamma = \sqrt{(1+2\gamma)/(1-2\gamma)}$: margin-dependent constants.
- n: repeated transition attempts (diagnostic depth); $P_{ss'}(a)$: probability of moving from state s to s' after action a; $\widehat{P}_{ss'}(a)$: estimator read from the agent's bets.
- $\mathcal{H}, D$: evaluation distributions over histories and tests. $\bar{\delta} = \mathbb{E}_{h\sim\mathcal{H},T\sim D}[\delta_T(\pi; h)]$: average normalized regret.
- $\lambda_k = (k - \frac{1}{2})/K$: threshold grid; $q_{T,\lambda_k}(h)$: probability the agent bets on T rather than the λ_k-lottery; $\widehat{p}_T$: resulting predictive-state estimate.
- $\bar{\delta}_K = \mathbb{E}_{h,T}[K^{-1}\sum_k \delta_{T,\lambda_k}(\pi; h)]$: average threshold-bet regret; $\varepsilon_K = 2\bar{\delta}_K + 1/(4K^2)$: recovery error rate.
- $\sigma = (a, o)$: action-observation symbol; B_σ: linear PSR update operator; S: core-history predictive-state matrix; $Y_\sigma = B_\sigma S$; $\|\cdot\|_F$: Frobenius norm; $C(S, Y)$: the explicit conditioning-dependent constant in Theorem 4.
- $M(h) = f(h)$: internal state computed from the full history and used by the policy: $\pi(\cdot \mid h, g_T) = \pi(\cdot \mid M(h), g_T)$. Histories are *aliased* when $M(h) = M(h')$.
- $\mathcal{P}$: distribution over pairs (h, h') with the same last observation. Its pair-averaged regret is $\bar{\delta}_{\mathcal{P}}(\pi) = \mathbb{E}_{(h,h')\sim\mathcal{P}}\frac{\mathbb{E}_{T\sim D}\delta_T(\pi; h) + \mathbb{E}_{T\sim D}\delta_T(\pi; h')}{2}$.
- $S_\gamma(h, h')$: tests demanding opposite confident bets: $p_T(h) \geq \frac{1}{2} + \gamma$ and $p_T(h') \leq \frac{1}{2} - \gamma$. The witnessed aliasing mass is $q^{\text{Alias}}_\gamma(M) = \Pr_{(h,h')\sim\mathcal{P},T\sim D}\left[M(h) = M(h'),\ T \in S_\gamma(h, h')\right]$.

Figure 1: selection depends on the task family

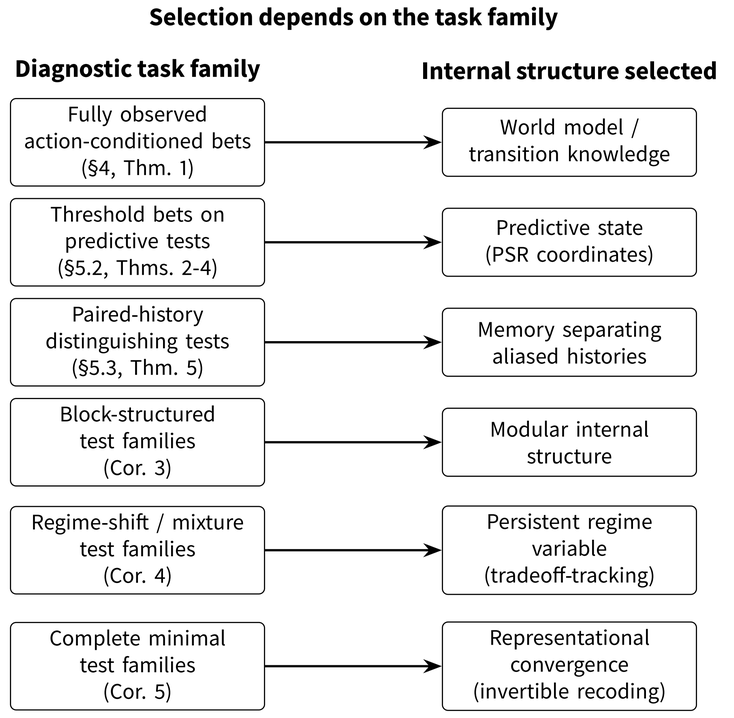
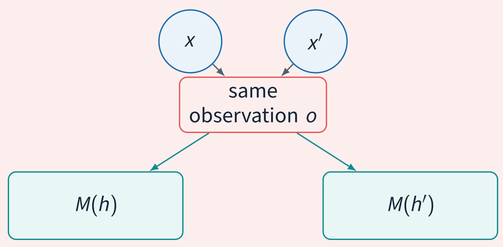


Figure 1. Different diagnostic task families select different internal structure, from world models and predictive state to memory, modularity, persistent regime variables, and representational convergence.

Open problem solved: necessity and recovery in POMDPs

Can world-model necessity survive when observations alias latent states?

Richens et al. (ICML 2025) explicitly left partial observability open. The same observation can mix distinct latent dynamics, so fully observed recovery arguments break.



low regret forces separation whenever confident predictions differ

Answer: yes. Action-conditioned predictive tests bypass physical-state recovery. Low average regret forces:

- **Thm. 2: predictive modeling necessity** on all high-margin tests;
- **Thm. 3: predictive-state recovery** from a curve of threshold bets;
- **Thm. 4: linear PSR-operator recovery** under finite-dimensional linear dynamics;
- **Thm. 5: belief-like memory** via a quantitative no-aliasing lower bound.

THRESHOLD BETS RECOVER PREDICTIVE MAGNITUDES

$$\widehat{p}_T(h) = \frac{1}{K}\sum_{k=1}^{K} q_{T,\lambda_k}(h), \qquad \mathbb{E}(\widehat{p}_T - p_T)^2 \leq 2\bar{\delta}_K + \frac{1}{4K^2}.$$

Repeating the same predictive test across thresholds reveals the full predictive coordinate, not only which side of $1/2$ it lies on.

Thm. 4: with linear PSR dynamics, an invertible core-history matrix, and the stated small-error condition, $\sum_\sigma \|\widehat{B}_\sigma - B_\sigma\|_F^2 \leq C(S, Y)\varepsilon_K$.

Theorem 5: quantitative no-aliasing

If memory collapses histories h, h' that demand opposite γ-margin bets, then every M-based policy pays

$$\bar{\delta}_{\mathcal{P}}(\pi) \geq q^{\text{Alias}}_\gamma(M)\frac{c(\gamma)}{2}$$

Low regret rules out memory states that merge decision-relevant histories.

Corollaries 3–4: organization also emerges

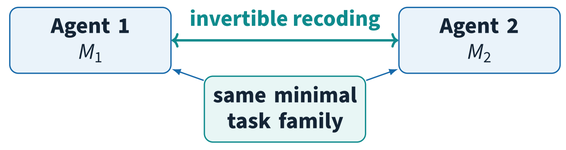
- **Block-structured tasks ⇒ informational modularity.** Each test block forces its own decision-relevant distinctions to be preserved.
- **Shifting task mixtures ⇒ persistent regime tracking.** When regimes flip the correct bet for the same test, memory must track the latent evaluative condition.
- **Functional implication.** Persistent global modulators resemble primitives associated with **emotional affect** or homeostasis—a claim about organization, and *perhaps* phenomenology.

task block ⇒ **module** **regime** ⇒ **persistent state**

Corollary 5: same hard tasks ⇒ convergence

Under γ-minimality and complete diagnostic tests, two vanishing-regret agents induce the same decision partition:

$$M_1 = \varphi(M_2), \qquad M_2 = \psi(M_1).$$



This formalizes a route from shared competence constraints to the representational alignment observed across high-performing models and brains. Our follow-up work on "Contravariance Theory" (**Yamins* & Nayebi* 2026; arXiv:2607.08561**) shows the complexity of this mapping/recoding (φ and ψ) is *at most* linear for modern deep neural networks.

Corollaries 1–2: a sharp causal boundary

RECOVERABLE		NOT IDENTIFIED
L2	≠	L3
Interventional kernel $P(S_{t+1} \mid S_t, \mathrm{do}(A_t))$		Cross-action counterfactual coupling

Even exact interventional recovery does not identify Pearl Level 3 counterfactuals. Those require a structural causal model specifying exogenous noise and its coupling across actions.

What to look for in increasingly agentic AI

- predictive world models that sharpen with horizon;
- belief-like memory that separates aliased histories;
- modules matching task-family structure;
- persistent regime / tradeoff variables *functionally* resembling emotions;
- convergence across agents trained on the same hard tasks.

Robust competence under uncertainty selects what an agent must know.

SCAN FOR PAPER

Paper, contact, and acknowledgements

arXiv:2603.02491 **anayebi@cs.cmu.edu**

Acknowledgements: Lenore and Manuel Blum, Dylan Hadfield-Menell, Daniel Yamins, Santiago Cifuentes, Leo Kozachkov, Reece Keller, Noushin Quazi, and the anonymous reviewers. Funding: Burroughs Wellcome Fund, Foresight Institute, and Protocol Labs.

Key references

- Sondik, E. J. (1971). *The Optimal Control of Partially Observable Markov Processes*. PhD dissertation, Stanford University.
- Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2), 99–134.
- Littman, M. L., Sutton, R. S., & Singh, S. P. (2001). Predictive representations of state. *Advances in Neural Information Processing Systems*, 14.
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.
- Wentworth, J. (2021). Selection theorems: A program for understanding agents. *AI Alignment Forum / LessWrong*.
- Richens, J., & Everitt, T. (2024). Robust agents learn causal world models. *The Twelfth International Conference on Learning Representations*.
- Richens, J., Everitt, T., & Abel, D. (2025). General agents need world models. *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267, 51659–51687.

WHAT CHANGES RELATIVE TO PRIOR WORLD-MODEL RECOVERY?

worst-case optimality	deterministic policies	full observability only
⇓	⇓	⇓
average normalized regret	stochastic policies	necessity + recovery in POMDPs