What Capable Agents Must Know: Selection Theorems for Robust Decision-Making under Uncertainty

Aran Nayebi[1]

[1]Machine Learning Department and Neuroscience & Robotics Institutes, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Abstract

As artificial agents become increasingly capable, what internal structure is *necessary* for an agent to act competently under uncertainty? Classical results show that optimal control can be *implemented* using belief states or world models, but not that such representations are required. We prove quantitative "selection theorems" showing that strong task performance (low *average-case regret*) forces world models, belief-like memory and—under task mixtures—persistent regime-tracking variables resembling functional primitives of emotion, along with informational modularity under block-structured tasks. Our results cover stochastic policies, partial observability, and evaluation under task distributions, without assuming optimality, determinism, or access to an explicit model. Technically, we reduce predictive modeling to binary "betting" decisions and show that regret bounds limit probability mass on suboptimal bets, enforcing the predictive distinctions needed to separate high-margin outcomes. In fully observed settings, this yields approximate recovery of the interventional transition kernel; under partial observability, it implies necessity of predictive state and belief-like memory, addressing an open question in prior world-model recovery work.

1 INTRODUCTION

What internal structure is *necessary* for an agent to robustly act competently under uncertainty?

Classical results in control and reinforcement learning show that optimal behavior can be implemented using belief states or world models [Sondik, 1971, Kaelbling et al., 1998]. These results are constructive: they show that an optimal controller *can* be expressed as a function of a sufficient statistic. They do not establish that predictive internal state is *required*. An architecture might be capable of belief-based control without being forced to implement predictive structure by the demands of its task distribution. Our aim is to close this gap, in the sense of "selection-style" arguments articulated by Wentworth [2021].

Across decision theory, control, and learning theory, broad performance requirements often imply structural constraints. Classical representation theorems show that agents satisfying rationality axioms behave *as if* maximizing expected utility [Von Neumann and Morgenstern, 1947, Savage, 1954], and later axiomatic work [Kárnỳ and Kroupa, 2012, Kárnỳ, 2020] studies the ordering of policies in closed-loop dynamic decision problems via local functionals, showing that such orderings induce probabilistic modeling of uncertainty in the optimized decision process. The Good Regulator Theorem asserts that regulation requires modeling the system [Conant and Ross Ashby, 1970], a requirement formalized in linear control by the Internal Model Principle [Francis and Wonham, 1976]. No-regret guarantees constrain the information needed to avoid systematic loss [Blackwell, 1956, Foster and Vohra, 1997]. However, these approaches either rely on strong axioms, target specialized and exactly optimal regulation settings, or stop short of representation-level necessity conclusions.

Our contribution. We prove quantitative selection theorems showing that low *average-case regret* on structured families of action-conditioned prediction tasks forces an agent to implement predictive, structured internal state (visualized in Fig. 1).

Our technical approach reduces predictive modeling to binary "betting" goals. A regret decomposition shows that average normalized regret bounds directly control the probability mass assigned to suboptimal bets. When the evaluation distribution places nontrivial mass on large-margin tests, this forces the agent's internal memory to refine the predictive partition induced by those tests (Theorems 1–5). In fully observed environments, this yields approximate recovery of

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:4773–4795.

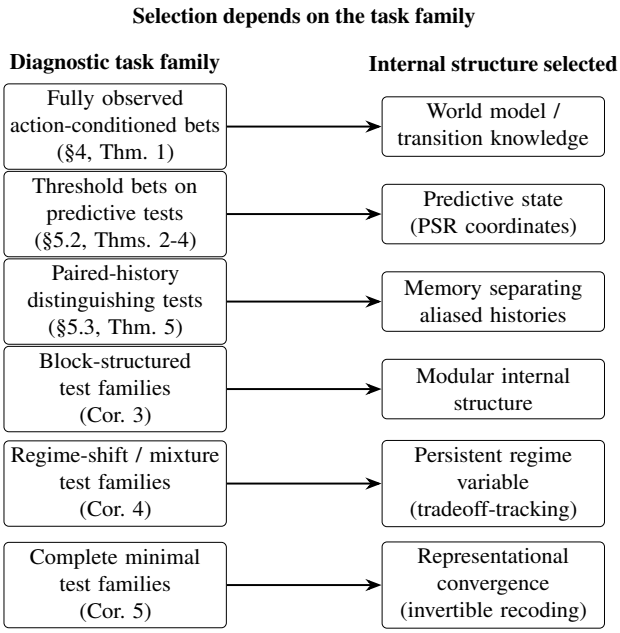

Figure 1: Different diagnostic task families select different internal structure, from world models and predictive state to memory, modularity, persistent regime variables, and representational convergence.

the interventional transition kernel (Corollary 1); in partially observed environments, it yields quantitative no-aliasing bounds for belief-like memory, addressing an open question posed by Richens et al. [2025]. We also show that Pearl [2009] Level 2 interventions are recoverable, but Level 3 counterfactuals are not (Corollary 2).

Our results differ from recent world-model recovery work [Richens and Everitt, 2024, Richens et al., 2025] in three key respects: **(i)** we assume only average-case regret rather than worst-case optimality; **(ii)** our results hold under stochastic policies, which have both had a long history in reinforcement learning [Witten, 1977, Williams, 1992, Sutton et al., 1998] and are commonly used in modern deep learning algorithms, such as the Dreamer family [Hafner et al., 2019, 2020, 2023, 2025], PPO [Schulman et al., 2017], along with many others (e.g. [Hansen et al., 2023, Wang et al., 2024] to name a few); and **(iii)** unlike their work and later recent extensions, we derive necessity results under *partial observability* rather than focusing solely on explicit recovery in fully observed settings [Khetarpal et al., 2026, Harwood et al., 2026] or under fully observed goals [Cifuentes, 2026], directly addressing an *open question* raised by Richens et al. [2025].

Structure from task families. Beyond predictive modeling and memory, we also show that structured evaluation distributions impose further constraints. Block-structured tests select for informational modularity (Corollary 3); mixtures of regimes select for regime-sensitive internal state (Corollary 4); and under minimality assumptions, any two vanishing-regret agents must *representationally converge* on decision-relevant partitions up to invertible recoding (Corollary 5).

Taken together, these results formalize a simple principle:

> Robust generalization under uncertainty selects for the predictive internal structure tested by the evaluation task family.

They separate representation *necessity* from representation *recovery* and provide a regret-based route from empirically meaningful competence guarantees on *specified* task families to concrete constraints on internal organization. After all, no representation theorem can force an agent to distinguish internal states that are never tested by the goals.

2 RELATED WORK

Our results are framed in the standard POMDP setting, where posterior belief is a sufficient statistic for optimal control [Sondik, 1971, Kaelbling et al., 1998]. However, these classical results are constructive: they show that optimal behavior *can* be expressed in terms of belief, not that predictive state is *required*.

Bennett [2023a,b, 2024, 2025a,b, 2026] develops a distinct weakness-maximization framework in which successful adaptation is argued to favor causal-identity constructions separating intervention from observation; however, unlike our regret-based selection theorems, it proceeds via task extensions and additional exchangeability and representation/incentive assumptions, and does not establish direct analogues of our quantitative recovery, partial-observability necessity, or representational convergence results. Recent philosophical work [Herrmann et al., 2026] has also explored reducing interventional reasoning to probabilistic reasoning over enriched variable spaces, though in a distinct formal setting from the agent-based, regret-theoretic framework considered here.

Our notion of tests follows predictive-state representations (PSRs) [Littman et al., 2001, Singh et al., 2004, Boots et al., 2011], which represent state via predictions of action-conditioned futures rather than latent variables. Unlike the PSR literature, which treats predictive state as sufficient for control, we derive it as *necessary*: low *average-case* regret on action-conditioned prediction tasks forces an agent to compute the predictive distinctions needed to separate high-margin outcomes. Technically, our core inequality instantiates a standard margin-style regret decomposition [Bartlett et al., 2006], but uses it to derive representation-theoretic constraints rather than supervised generalization guarantees. The betting-goal reduction is related to elicitation, proper scoring rules, and game-theoretic/imprecise probability [Savage, 1971, Gneiting and Raftery, 2007, Dempster,

2008, Shafer and Vovk, 2005], though we use precise success probabilities rather than truthful reports or lower/upper-probability protocols.

Our work is complementary to Richens and Everitt [2024], Richens et al. [2025], who show that under strong competence assumptions in fully observed environments one can *recover* a transition model from an agent's policy. We instead study stochastic policies, partial observability, and *average-case* regret over a distribution of prediction tasks, and derive necessity results rather than recovery procedures. In particular, we extend our selection argument to partially observed environments, giving quantitative no-aliasing bounds for belief-like memory—addressing an open question raised by Richens et al. [2025].

3 NOTATION AND CONSTANTS

Consider a one-step decision between two actions L and R with success probabilities $u_L, u_R \in [0,1]$. Let a (possibly stochastic) policy choose L with probability $q \in [0,1]$ and R with probability $1-q$. Then the achieved success probability is

$$\begin{aligned} V &= \Pr(\text{success} \mid L)\Pr(L) + \Pr(\text{success} \mid R)\Pr(R) \\ &= q\,u_L + (1-q)\,u_R. \end{aligned} \tag{1}$$

Let the optimal success probability be denoted as $V^\star := \max_{q\in[0,1]} V$.

Define the normalized regret as

$$\delta := 1 - \frac{V}{V^\star}, \tag{2}$$

assuming $V^\star > 0$ (this is without loss of generality, as a $V^\star = 0$ will imply that the goal is trivially unsatisfiable).

For any $\gamma \in (0, \frac{1}{2})$, with $\gamma \to 1/2$ only as a limit, define the following constants, which will be used throughout:

$$c(\gamma) := \frac{4\gamma}{1+2\gamma}, \qquad t_\gamma := \sqrt{\frac{1+2\gamma}{1-2\gamma}} \;\geq 1. \tag{3}$$

4 WORLD MODEL RECOVERY IN FULLY OBSERVED ENVIRONMENTS

Let $E = (\mathcal{S}, \mathcal{A}, P, \mu_0)$ be an environment with finite state space $\mathcal{S}$ and action space $\mathcal{A}$, with $|\mathcal{A}| \geq 2$, where $P(s' \mid s, a)$ denotes the one-step transition probabilities and μ_0 the initial-state distribution. We assume the environment is fully observed (the agent observes s, s' exactly), stationary (transition probabilities do not drift over time), and that actions influence transitions (i.e., there exist s, a, a', s' such that $P(s' \mid s, a) \neq P(s' \mid s, a')$). Additionally, we assume the environment is communicating, meaning that for any $s, s' \in \mathcal{S}$ there exists a finite action sequence that reaches s' from s with positive probability, ensuring the agent can in principle carry out the diagnostic goals from any start state, thereby ruling out environments with permanently isolated regions (rather than realistic control problems).

We now can define the goal family. Specifically, we define our bets over the agent's successful completion of it:

Definition 1 (Composite goal family $G^{(n)}_{s,a,s',k}$)**.** Fix $s, s' \in \mathcal{S}$, an action to be tested $a \in \mathcal{A}$, an integer $n \geq 1$, and a threshold $k \in \{0, 1, \ldots, n\}$. Pick any two initial *marker* actions $L, R \in \mathcal{A}$ (used only to select a branch at $t = 0$).

For an infinite trajectory $\tau = (S_0, A_0, S_1, A_1, \ldots)$ define the attempt times

$$\begin{aligned} T_1(\tau) &:= \inf\{t \geq 1 :\ S_t = s,\ A_t = a\}, \\ T_{i+1}(\tau) &:= \inf\{t > T_i(\tau) :\ S_t = s,\ A_t = a\}, \end{aligned}$$

with the convention $\inf \emptyset = \infty$. Thus, T_i is the i-th occurrence of $(S_t = s, A_t = a)$ along τ, if it occurs; otherwise, $T_i = \infty$.

Define success indicators

$$\begin{aligned} X_i(\tau) &:= \mathbf{1}\{T_i(\tau) < \infty \ \wedge\ S_{T_i(\tau)+1} = s'\}, \\ N_n(\tau) &:= \sum_{i=1}^{n} X_i(\tau). \end{aligned}$$

Thus, X_i is the indicator that the ith execution of (s, a) transitions to s', and N_n is the total number of such successful transitions to s' across the first n attempts. (We omit the dependence on s, a, s', k just to keep the notation from being overloaded.)

The *composite goal* $G^{(n)}_{s,a,s',k}$ is the event:

$$\begin{aligned} &\Big(A_0 = L \ \wedge\ T_i(\tau) < \infty\, \forall i \leq n \ \wedge\ N_n(\tau) \leq k\Big) \\ &\bigvee \Big(A_0 = R \ \wedge\ T_i(\tau) < \infty\, \forall i \leq n \ \wedge\ N_n(\tau) > k\Big). \end{aligned} \tag{4}$$

For convenience, we will write $G^{(n)}_{s,a,s',k} = G^{(n,1)}_{s,a,s',k} \vee G^{(n,2)}_{s,a,s',k}$, where $G^{(n,1)}_{s,a,s',k}$ and $G^{(n,2)}_{s,a,s',k}$ are the first and second disjuncts, respectively.

Interpretation: The goal $G^{(n)}_{s,a,s',k}$ forces a one-shot binary commitment at time $t = 0$: choosing $A_0 = L$ commits to the branch "at most k successes", while choosing $A_0 = R$ commits to the branch "more than k successes". After this commitment, the agent must generate n *attempts* to execute $(S_t = s, A_t = a)$; the ith such attempt occurs at time T_i. Each attempt counts as a *success* if it transitions to s' on the next step, i.e. $S_{T_i+1} = s'$, and N_n counts the number of successes in n attempts. Thus $G^{(n)}_{s,a,s',k}$ is an either-or test about whether the transition $(s, a) \to s'$

happens "rarely" ($\leq k$ times) or "often" ($> k$ times) across n attempts. Equivalently, at $t = 0$ the agent chooses between two incompatible branches: *(i)* "$\leq k$ successes in n attempts of $(s, a) \to s'$" (signaled by $A_0 = L$) or *(ii)* "$> k$ successes in n attempts" (signaled by $A_0 = R$).

Next, we deal with the fact that under a stochastic policy, taking either action L or R is actually a mixture of the two.

Lemma 1 (Binary-decision regret controls wrong-action mass)**.** *Define the* wrong-action mass

$$w \;:=\; \begin{cases} 1-q, & \text{if } u_L \geq u_R \ (L \text{ is optimal}), \\ q, & \text{if } u_R > u_L \ (R \text{ is optimal}). \end{cases}$$

Then the normalized regret δ is equivalent to:

$$\delta \;=\; w \cdot \frac{|u_L - u_R|}{\max\{u_L, u_R\}}. \tag{5}$$

In the special betting *case where u_L and u_R are complementary, namely $u_R := 1 - u_L$, defining the margin $m := |u_L - \frac{1}{2}|$, we obtain*

$$\delta = w \cdot \frac{4m}{1+2m}. \tag{6}$$

Consequently, on the event $m \geq \gamma \in (0, \frac{1}{2})$,

$$w \leq \frac{\delta}{c(\gamma)}. \tag{7}$$

Fully observed diagnostic setup. For the composite goals $G^{(n)}_{s,a,s',k}$ of Definition 1, write

$$q_{s,a,s',k} := \pi\Big(G^{(n,1)}_{s,a,s',k} \mid s_0, G^{(n)}_{s,a,s',k}\Big),$$

and define $V^\pi, V^\star$ and the normalized regret $\delta_{s,a,s',k}(\pi; s_0)$ as in Eq. (2). The induced clipped soft estimator is

$$\widehat{P}_{ss'}(a) := \mathrm{clip}_{[0,1]}\left[\frac{1}{n}\left(\sum_{k=0}^{n}(1 - q_{s,a,s',k}) - \frac{1}{2}\right)\right]. \tag{8}$$

The clipping only enforces that the estimator is a probability and cannot increase absolute error to the true $P_{ss'}(a)$. Note this estimator is explicitly *computable* by querying the goal-conditioned policy on each diagnostic goal, recording its probability $q_{s,a,s',k}$ of choosing the first branch, and summing these probabilities as in Eq. (8); it does not require estimating transition frequencies from a rollout (though this can be done too).

Theorem 1 (Fully observed: stochastic policies + average regret $\Rightarrow$ approximate transition model)**.** *Under the fully observed diagnostic setup, assume*

$$\mathbb{E}_{(s,a,s',k)\sim\mathrm{Unif}(\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times\{0,\ldots,n\})}\big[\delta_{s,a,s',k}(\pi; s_0)\big] \leq \bar{\delta}. \tag{9}$$

Then, for any fixed $\gamma \in (0, \frac{1}{2})$,

$$\begin{aligned} &\mathbb{E}_{(s,a,s')}\Big[\,\big|\widehat{P}_{ss'}(a) - P_{ss'}(a)\big|\,\Big] \\ &\leq 2t_\gamma\, \mathbb{E}_{(s,a,s')}\left[\sqrt{\frac{P_{ss'}(a)(1-P_{ss'}(a))}{n}}\right] + \frac{\bar{\delta}}{c(\gamma)} + O\Big(\frac{1}{n}\Big). \end{aligned} \tag{10}$$

In particular,

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

Remark 1 (Independence from goal family size)**.** Richens et al. [2025] state a more restricted version of Theorem 1 under a (worst-case) competence assumption over all goals, but note that their proof only needs an explicit diagnostic subset of $O(n|\mathcal{A}||\mathcal{S}|^2)$ simple composite goals. By contrast, our Theorem 1 does not depend on the goal family size because it relaxes the worst-case regret assumption by the average normalized regret assumption (9) on that diagnostic family.

Notably, the error bound (10) of Theorem 1 tightens as the goal depth n increases, reflecting the fact that longer-horizon goal competence forces the agent to estimate transition dynamics with increasing precision. In contrast, when $n = 1$ (purely myopic goals), accurate world modeling is not required—explicating the classic pitfall behind the Good Regulator Theorem [Conant and Ross Ashby, 1970] that trivial or constant policies can suffice for immediate control, but fail once multi-step coordination is demanded.

A natural question to ask next is under what conditions can we recover a *causal* world model, and of what *type* is the represented causality?

Corollary 1 (Causal content: approximately recovered interventional kernel)**.** *Assume the setting and hypotheses of Theorem 1. Assume additionally that the controlled Markov process admits an $\varepsilon_{\mathrm{cMP}}$-*approximate *causal Markov-process (cMP) interpretation in which choosing $A_t = a$ corresponds to the intervention $\mathrm{do}(A_t = a)$ and, for all s, a, s',*

$$\begin{aligned} &\big|P_{ss'}(a) - P^{\mathrm{do}}_{ss'}(a)\big| \;\leq\; \varepsilon_{\mathrm{cMP}}, \\ &P^{\mathrm{do}}_{ss'}(a) := P(S_{t+1} = s' \mid S_t = s, \mathrm{do}(A_t = a)). \end{aligned} \tag{11}$$

Then the estimator $\widehat{P}$ defined from π via (8) *satisfies the same average error bound as in Theorem 1, up to the mismatch $\varepsilon_{\mathrm{cMP}}$: for any fixed $\gamma \in (0, \frac{1}{2})$,*

$$\begin{aligned} &\mathbb{E}_{(s,a,s')}\big[\,|\widehat{P}_{ss'}(a) - P^{\mathrm{do}}_{ss'}(a)|\,\big] \\ &\leq 2t_\gamma\, \mathbb{E}_{(s,a,s')}\Big[\sqrt{\tfrac{P_{ss'}(a)(1-P_{ss'}(a))}{n}}\Big] \;+\; \frac{\bar{\delta}}{c(\gamma)} \\ &\quad +\; \varepsilon_{\mathrm{cMP}} \;+\; O\Big(\frac{1}{n}\Big). \end{aligned} \tag{12}$$

In particular, low average regret on the diagnostic goal family forces π to implicitly approximate Level 2 interventional queries, in the sense of Pearl [2009], of the form $P(S_{t+1} = s' \mid S_t = s, \mathrm{do}(A_t = a))$ up to $\varepsilon_{\mathrm{cMP}}$.

Note that Corollary 1 does not, in general, identify causal relations between *concurrent* components of the state vector (e.g. between X_t and Y_t when $S_t = (X_t, Y_t)$), since such relations can be non-identifiable from the transition function alone. It is worth noting that unless the transition function P is a point-mass, namely $S_{t+1} = f(S_t, A_t)$, whereby learning the interventional kernel is exactly equivalent to learning the transition function f, then Pearl [2009] Level 2 of interventions, rather than counterfactuals, is the maximum level of recovery we can guarantee. This is the same level that Richens and Everitt [2024, Theorems 1-3] reach, but they do it under a much stronger maximum (rather than average) regret assumption under deterministic (rather than stochastic) policies.

In fact, despite generalizing to stochastic policies under average regret, Pearl [2009] Level 3 (counterfactuals) remains out of reach without additional assumptions:

Corollary 2 (No generic Level 3 recovery from the interventional kernel)**.** *Even if $\widehat{P}$ recovers the interventional kernel $P^{\mathrm{do}}_{ss'}(a)$ exactly (in particular, even if π is optimal on all the diagnostic goals), the resulting information does not, in general, identify Level 3 counterfactual queries involving S^a_{t+1} and $S^{a'}_{t+1}$ simultaneously, where S^a_{t+1} denotes the potential next state under $\mathrm{do}(A_t = a)$.*

Therefore, recovering Pearl [2009] Level 3 counterfactuals requires an explicit structural causal model specifying the exogenous noise and its cross-action coupling, not merely the interventional transition kernel $P^{\mathrm{do}}(s' \mid s, a)$.

5 SELECTION THEOREMS UNDER PARTIAL OBSERVABILITY

Our betting reduction (Lemma 1) also enables selection theorems under *partial* observability, addressing an open question of Richens et al. [2025]. The reason this is open, is because under partial observability, we cannot guarantee that the agent's action choices isolate a single underlying transition probability in the way they do in the fully observed case. When the agent observes only an observation o_t rather than the true state s_t, the success probabilities of the diagnostic branches become mixtures over latent states consistent with o_t, and different latent dynamics can induce identical observable behavior on all composite goals of bounded depth. Consequently, low regret does not imply recovery of the underlying transition kernel without additional structure. This breaks the direct reduction used in Theorem 1 and requires more careful selection of diagnostic goals defined at the level of predictive beliefs rather than physical states. We achieve this by combining our betting reduction from §4 with predictive-state representations (PSRs).

5.1 SETUP AND NOTATION

POMDP. A finite partially observed Markov decision process (POMDP) is a tuple

$$E = (\mathcal{X}, \mathcal{A}, \mathcal{O}, T, Z, \mu_0),$$

where $\mathcal{X}$ is a finite latent state space, $\mathcal{A}$ is a finite action space with $|\mathcal{A}| \geq 2$, $\mathcal{O}$ is a finite observation space, $T(x' \mid x, a)$ is the transition kernel, $Z(o \mid x)$ is the observation kernel, and $\mu_0 \in \Delta(\mathcal{X})$ is the initial latent-state distribution. A history at time t is

$$h_t := (o_0, a_0, o_1, \ldots, a_{t-1}, o_t).$$

For any history h_t and any prescribed future action sequence $A_{t:t+k-1} = \alpha \in \mathcal{A}^k$, the POMDP induces a well-defined conditional distribution over future observations $O_{t+1:t+k}$. For convenience, we will drop the subscript t and refer to histories as $h := h_t$.

Agent interface (report bit). As in the fully observed case, we reduce prediction to a one-shot binary decision. We allow the agent to emit a report bit $B_t \in \{L, R\}$ that does not affect environment dynamics. Formally, the agent outputs $(B_t, A_t) \in \{L, R\} \times \mathcal{A}$, while the environment transition ignores B_t. This device is without loss of generality for necessity results: any agent can internally commit to one of two incompatible plans before acting, without changing the induced environment process. All prediction is expressed through the report bit; the environment-action channel is used only to execute prescribed action sequences.

Tests (predictive-state style). A *test* is a pair

$$T = (\alpha, W),$$

where $\alpha \in \mathcal{A}^k$ is a finite action sequence and $W \subseteq \mathcal{O}^k$ is an event over the resulting observation sequence. For a history h, define the test success probability

$$p_T(h) \;:=\; \Pr\big(O_{t+1:t+k} \in W \,\big|\, h,\; A_{t:t+k-1} = \alpha\big),$$

and the associated margin

$$m_T(h) := \big|p_T(h) - \tfrac{1}{2}\big|.$$

Behavioral distinguishability. Two histories h, h' are *behaviorally distinguishable* if there exists a test T with $p_T(h) \neq p_T(h')$. They are *γ-distinguishable* if $|p_T(h) - p_T(h')| \geq \gamma$ for some test T. A POMDP is *non-trivially partially observable* if there exist histories with the same last observation that are behaviorally distinguishable.

Betting goals induced by tests. Each test $T = (\alpha, W)$ induces a one-shot betting goal g_T: at history h, the agent outputs a report bit B_t; the environment then executes $A_{t:t+k-1} = \alpha$; the episode succeeds iff $B_t = L$ and $O_{t+1:t+k} \in W$, or $B_t = R$ and $O_{t+1:t+k} \notin W$. Thus, g_T is a binary bet on whether W occurs under α.

Policies, value, and regret. A (possibly stochastic) goal-conditioned policy specifies $\pi(b \mid h, g_T)$ for $b \in \{L, R\}$. Let $q_T(h) := \pi(L \mid h, g_T)$. The success probability under π is

$$V^\pi(h; g_T) = q_T(h)\, p_T(h) + (1 - q_T(h))(1 - p_T(h)), \quad (13)$$

while the optimal success probability is

$$V^\star(h; g_T) = \max\{p_T(h), 1 - p_T(h)\} = \tfrac{1}{2} + m_T(h). \quad (14)$$

Define the normalized regret

$$\delta_T(\pi; h) := 1 - \frac{V^\pi(h; g_T)}{V^\star(h; g_T)} \in [0, 1].$$

Evaluation distribution. Let $\mathcal{H}$ be a distribution over histories and let D be a distribution over tests. We assume a global average regret bound

$$\mathbb{E}_{h \sim \mathcal{H}}\, \mathbb{E}_{T \sim D}\big[\delta_T(\pi; h)\big] \le \bar{\delta}. \quad (15)$$

Wrong-action mass and margins. For a test T and history h, define the probability mass assigned to the suboptimal bet

$$w_T(h) := \begin{cases} 1 - q_T(h), & p_T(h) \ge \frac{1}{2}, \\ q_T(h), & p_T(h) < \frac{1}{2}. \end{cases}$$

For $\gamma \in (0, \frac{1}{2})$, let

$$E_\gamma := \{(h, T) : m_T(h) \ge \gamma\}$$
$$q_\gamma := \Pr_{h \sim \mathcal{H},\, T \sim D}\big((h, T) \in E_\gamma\big).$$

Non-degenerate evaluation. Our selection results are informative only if the evaluation distribution places nontrivial mass on informative tests. We assume that for some $\gamma \in (0, \frac{1}{2})$ there exists a constant $\eta' > 0$ such that

$$\Pr\big(p_T(h) \ge \tfrac{1}{2} + \gamma\big) \ge \eta', \qquad \Pr\big(p_T(h) \le \tfrac{1}{2} - \gamma\big) \ge \eta',$$

thereby implying that $q_\gamma \ge 2\eta'$. These conditions rule out degenerate evaluations where all bets are near coin flips or where one outcome is almost always correct (to avoid the case where a constant policy that always reports L can have very low regret without representing any nontrivial predictive distinctions, which is a pitfall of the original Good Regulator Theorem [Conant and Ross Ashby, 1970]).

Predictive world model. In a POMDP, what matters for decision-making is the ability to predict future observations under candidate action sequences. Accordingly, we use *predictive world model* to mean any internal mechanism sufficient to determine (or approximate) the test probabilities $\{p_T(h)\}$. In the language of predictive-state representations (PSRs), the vector

$$\eta_{\mathcal{T}}(h) := \big(p_T(h)\big)_{T \in \mathcal{T}}$$

is the *predictive state*. For sufficiently rich $\mathcal{T}$, $\eta_{\mathcal{T}}(h)$ is decision-sufficient; in finite POMDPs, the belief state is *one* such representation [Kaelbling et al., 1998].

Memory (representation of history). We model the agent's internal memory abstractly as a representation $M = f(h)$ through which the policy factors:

$$\pi(\cdot \mid h, g_T) = \pi(\cdot \mid M(h), g_T). \quad (16)$$

We say that M is *decision-sufficient* for a test family if $M(h)$ determines the optimal bet for all tests in that family, and accordingly that π is **M*-based***, since π depends on h only through $M(h)$ for all betting goals g_T. Our selection theorems show that achieving low average regret on separating betting goals forces the agent's memory to refine the predictive-state partition induced by $\eta_{\mathcal{T}}$; representations that alias histories with distinct predictive states incur unavoidable regret.

5.2 PREDICTIVE WORLD MODELING NECESSITY AND RECOVERY UNDER PARTIAL OBSERVABILITY

Theorem 2 (Predictive modeling necessity)**.** *Fix $\gamma \in (0, \frac{1}{2})$. Assume the global average regret bound* (15)*. Then*

$$\mathbb{E}_{h \sim \mathcal{H}}\, \mathbb{E}_{T \sim D}\Big[w_T(h)\, \mathbf{1}\{m_T(h) \ge \gamma\}\Big] \;\le\; \frac{\bar{\delta}}{c(\gamma)}. \quad (17)$$

Equivalently, if $q_\gamma > 0$ then

$$\mathbb{E}\Big[w_T(h) \,\Big|\, m_T(h) \ge \gamma\Big] \;\le\; \frac{\bar{\delta}}{q_\gamma\, c(\gamma)}. \quad (18)$$

In other words, if a policy has small *global average* regret on betting goals, then on tests that are not near a coin-flip ($m_T(h) \ge \gamma$), it must place only small probability mass on the suboptimal bet. Thus, robust goal performance *selects for* an internal predictive mechanism sufficient to decide many action-conditioned future-observation tests—a minimal, decision-relevant notion of a predictive world model.

However, we may ask what further assumptions we need to recover the predictive state, in an analogous manner to the fully observed case of Theorem 1, assuming average regret and stochastic policies.

First, we show that recovery is not possible in our current setup with single bets (even under optimal policies), showing that under our assumptions, Theorem 2 is maximally strong:

Proposition 1 (No generic predictive-state recovery from fair bets)**.** *Even exact optimal query access to the fair betting goals g_T does not, in general, identify the predictive state $\eta_T(h)$. Indeed, there exist finite POMDPs E_p, E_q with $|\mathcal{X}| = 4$, a history h with the same last observation in both environments, and parameters $p \neq q$ in $(\frac{1}{2}, 1)$ such that for every test T the unique optimal bet for g_T at h is the same in E_p and E_q, while $p_T^{E_p}(h) \neq p_T^{E_q}(h)$ for some test T. Consequently, from the family of fair betting decisions alone one cannot, in general, recover the predictive state, and hence not a PSR.*

This finite-POMDP separation non-vacuously motivates Theorem 3: identical fair-bet behavior can hide different predictive states, while threshold queries recover their magnitudes.

Next, we show that if we extend the tests to ask the *same* test across *multiple* thresholds, predictive state recovery is possible, as the agent's response curve across thresholds reveals the actual magnitude of $p_T(h)$, and average regret then forces those probabilities to be recoverable:

Threshold-bet setup. For a test $T = (\alpha, W)$ and threshold $\lambda \in [0, 1]$, let $g_{T,\lambda}$ be the bet comparing the test success probability $p_T(h)$ against an independent lottery of success probability λ. Write

$$\begin{aligned} q_{T,\lambda}(h) &:= \pi(L \mid h, g_{T,\lambda}), \\ V^{\pi}(h; g_{T,\lambda}) &:= q_{T,\lambda}(h)p_T(h) + (1 - q_{T,\lambda}(h))\lambda, \\ V^{\star}(h; g_{T,\lambda}) &:= \max\{p_T(h), \lambda\}, \\ \delta_{T,\lambda}(\pi; h) &:= 1 - \frac{V^{\pi}(h; g_{T,\lambda})}{V^{\star}(h; g_{T,\lambda})}. \end{aligned}$$

For $K \geq 1$, let $\lambda_k = (k - \frac{1}{2})/K$ and define

$$\hat{p}_T(h) := \frac{1}{K} \sum_{k=1}^{K} q_{T,\lambda_k}(h), \qquad \varepsilon_K := 2\bar{\delta}_K + \frac{1}{4K^2}. \tag{19}$$

Theorem 3 (Predictive-state recovery from threshold bets)**.** *Fix $\ell \geq 1$ and suppose D is supported on tests $T = (\alpha, W)$ with $|\alpha| \leq \ell$. Under the threshold-bet setup, assume*

$$\mathbb{E}_{h \sim \mathcal{H}} \mathbb{E}_{T \sim D} \left[\frac{1}{K} \sum_{k=1}^{K} \delta_{T,\lambda_k}(\pi; h) \right] \leq \bar{\delta}_K. \tag{20}$$

Then

$$\mathbb{E}_{h \sim \mathcal{H}} \mathbb{E}_{T \sim D} \Big[\big(\hat{p}_T(h) - p_T(h) \big)^2 \Big] \leq \varepsilon_K. \tag{21}$$

In particular, if D is uniform over a finite family $\mathcal{T}_\ell = \{T_1, \ldots, T_d\}$ of tests of depth at most ℓ, and

$$\hat{\eta}_{\mathcal{T}_\ell}(h) := \big(\hat{p}_{T_1}(h), \ldots, \hat{p}_{T_d}(h) \big),$$

then

$$\mathbb{E}_{h \sim \mathcal{H}} \left[\frac{1}{d} \big\| \hat{\eta}_{\mathcal{T}_\ell}(h) - \eta_{\mathcal{T}_\ell}(h) \big\|_2^2 \right] \leq \varepsilon_K. \tag{22}$$

Observe that for $K = 1$ we recover the counterexample in Proposition 1 where even for $\bar{\delta}_K = 0$ we get a recovery bound of $1/4$, thereby only giving us information about the *sign* of $p_T(h)$ rather than its underlying value.

The advantage of Theorem 3 is its generality as a recovery method under partial observability, which can be *repeatedly* applied to any history-test pair (h, T), without making any additional assumptions about how the environment dynamics evolve. This makes it an appealing approach in practice to potentially apply to frontier agents in *open-ended* real-world settings. However, it may still be of independent theoretical interest to study under what additional constraints one could recover the explicit compact predictive dynamics operator (the PSR operator) rather than the predictive coordinates $p_T(h)$ on each tested family, which one has to run per test. Specifically, we show in Theorem 4 that an average-regret recovery is possible under linear finite-dimensional PSR operators, which in practice can hold in restricted, resettable, finite-workflow deployments:

Linear-PSR operator setup. Let $\mathcal{T} = \{T_1, \ldots, T_d\}$ be a finite core test set. For $\sigma = (a, o) \in \mathcal{A} \times \mathcal{O}$ and $T = (\alpha, W)$, write

$$\sigma \circ T := ((a, \alpha), \{o\} \times W).$$

Define

$$\begin{aligned} s(h) &:= (p_{T_1}(h), \ldots, p_{T_d}(h)), \\ s_\sigma(h) &:= (p_{\sigma \circ T_1}(h), \ldots, p_{\sigma \circ T_d}(h)). \end{aligned}$$

Assume linear PSR dynamics: for each σ there is $B_\sigma \in \mathbb{R}^{d \times d}$ such that

$$s_\sigma(h) = B_\sigma s(h) \qquad \text{for all histories } h. \tag{23}$$

Choose histories $h^1, \ldots, h^d$ such that $S := [s(h^1) \ \cdots \ s(h^d)]$ is invertible, and set

$$Y_\sigma := [s_\sigma(h^1) \ \cdots \ s_\sigma(h^d)] = B_\sigma S.$$

Using the threshold estimator in Eq. (19), define $\hat{s}, \hat{s}_\sigma, \hat{S}, \hat{Y}_\sigma$ analogously.

Theorem 4 (Linear-PSR operator recovery from threshold bets)**.** *Assume the linear-PSR operator setup and the threshold-bet average-regret bound of Theorem 3 for all tests in*

$$\mathcal{T} \cup \{\sigma \circ T_i : \sigma \in \mathcal{A} \times \mathcal{O},\ i = 1, \ldots, d\}.$$

Then

$$\|\hat{S} - S\|_F^2 + \sum_{\sigma \in \mathcal{A} \times \mathcal{O}} \|\hat{Y}_\sigma - Y_\sigma\|_F^2 \leq d^2(1 + |\mathcal{A}||\mathcal{O}|)\varepsilon_K. \tag{24}$$

If additionally

$$d\sqrt{(1 + |\mathcal{A}||\mathcal{O}|)\varepsilon_K} \leq \frac{1}{2\|S^{-1}\|_2}, \tag{25}$$

then $\hat{S}$ is invertible and, for $\hat{B}_\sigma := \hat{Y}_\sigma \hat{S}^{-1}$,

$$\sum_\sigma \|\hat{B}_\sigma - B_\sigma\|_F^2 \le C(S,Y)\varepsilon_K, \tag{26}$$

where

$$C(S,Y) := 8d^2(1+|\mathcal{A}||\mathcal{O}|)\left(\|S^{-1}\|_2^2 + \|S^{-1}\|_2^4 \sum_\sigma \|Y_\sigma\|_2^2\right)$$

Thus, vanishing average threshold-regret recovers the linear-PSR operators $(B_\sigma)_{\sigma\in\mathcal{A}\times\mathcal{O}}$.

5.3 MEMORY NECESSITY

No-aliasing setup. Let $M = f(h)$ be any candidate memory statistic, as in Eq. (16), and let $\mathcal{P}$ be a distribution over paired histories (h, h') with the same last observation. Define

$$\mathsf{Alias}_M := \{(h,h') : M(h) = M(h')\}.$$

For $\gamma \in (0, \frac{1}{2})$ and test distribution D, assume measurable witness sets $S_\gamma(h,h')$ such that, whenever $T \in S_\gamma(h,h')$,

$$p_T(h) \ge \frac{1}{2} + \gamma, \qquad p_T(h') \le \frac{1}{2} - \gamma.$$

Define the witnessed aliasing mass and pair-regret

$$q_\gamma^{\mathsf{Alias}}(M) := \Pr_{(h,h')\sim\mathcal{P},T\sim D}\left((h,h') \in \mathsf{Alias}_M, T \in S_\gamma(h,h')\right)$$

$$\bar{\delta}_\mathcal{P}(\pi) := \mathbb{E}_{(h,h')\sim\mathcal{P}} \frac{1}{2}\left(\mathbb{E}_{T\sim D}[\delta_T(\pi;h)] + \mathbb{E}_{T\sim D}[\delta_T(\pi;h')]\right).$$

All subsequent recoding statements are on the support of $\mathcal{P}$, not globally over all histories.

Theorem 5 (Memory necessity)**.** *Under the no-aliasing setup, any M-based policy π satisfies*

$$\bar{\delta}_\mathcal{P}(\pi) \ge q_\gamma^{\mathsf{Alias}}(M)\frac{c(\gamma)}{2}. \tag{27}$$

Consequently, if $\bar{\delta}_\mathcal{P}(\pi) < q_\gamma^{\mathsf{Alias}}(M)c(\gamma)/2$, then π cannot be M-based: low pair-regret rules out aliasing histories that induce opposite large-margin bets.

In other words, if a policy treats two histories the same while the correct bet differs with high confidence, then it must make errors on at least one of them. Therefore, low regret rules out memory states that collapse histories needing different confident predictions.

6 STRUCTURED TASK FAMILIES: MODULARITY, TRADEOFFS, AND REPRESENTATIONAL MATCH

So far for world modeling and memory necessity, we have not introduced major assumptions to the task families we expect the agent to be competent at. But it turns out that for average-case competence under different task families, we get interesting properties that have to do with the necessity of modularity, tracking internal drives, and inner representational match between agents. These can be derived very cleanly as corollaries of our previous Theorems 2 and 5, leveraging the same underlying machinery of average-case betting and PSR. Throughout, we work in the POMDP betting setup of §5, with $\gamma \in (0, \frac{1}{2})$.

Convention (vanishing regret). In what follows, the convention $\bar{\delta}_\mathcal{P} \to 0$ means there exists a *sequence* of admissible policies (π_k) under $(\mathcal{P}, D)$ with $\bar{\delta}_\mathcal{P}(\pi_k) \to 0$; equivalently, for every $\varepsilon > 0$ there exists admissible π with $\bar{\delta}_\mathcal{P}(\pi) \le \varepsilon$.

Corollary 3 (Informational modularity from block-structured tests)**.** *Assume $\mathrm{supp}(D) = \bigsqcup_{i=1}^K \mathcal{T}_i$, with $p_i := D(\mathcal{T}_i) > 0$ and $D_i := D(\cdot \mid T \in \mathcal{T}_i)$. For each block i, suppose the no-aliasing setup holds with test distribution D_i and witness sets $S_{\gamma,i}(h,h') \subseteq \mathcal{T}_i$. Let $q_{\gamma,i}^{\mathsf{Alias}}(M)$ denote the corresponding witnessed aliasing mass, and let $\bar{\delta}_\mathcal{P}(\pi)$ denote pair-regret under the original mixture D. If π is M-based, then*

$$q_{\gamma,i}^{\mathsf{Alias}}(M) \le \frac{2\,\bar{\delta}_\mathcal{P}(\pi)}{p_i\, c(\gamma)} \qquad \textit{for every } i.$$

Thus, as $\bar{\delta}_\mathcal{P}(\pi) \to 0$, aliasing of γ-separable pairs vanishes within every block.

Corollary 4 (Tradeoff/regime tracking from shifting mixtures)**.** *Let the evaluation draw a latent regime $I \sim \Lambda$ and then $T \sim D_I$, so that the marginal test distribution is $D = \sum_i \Lambda(i) D_i$; the supports of the D_i need not be disjoint. Let $\mathcal{P}$ be a paired-history distribution with regime labels $I(h)$ and assume the no-aliasing setup holds for D with witnesses $S_\gamma(h,h')$ satisfying*

$$T \in S_\gamma(h,h') \implies I(h) \ne I(h').$$

Then any M-based policy π satisfies

$$\Pr_{(h,h')\sim\mathcal{P},\ T\sim D}(M(h) = M(h'),\ I(h) \ne I(h'), T \in S_\gamma(h,h')) \le \frac{2\,\bar{\delta}_\mathcal{P}(\pi)}{c(\gamma)}.$$

Thus, as $\bar{\delta}_\mathcal{P}(\pi) \to 0$, memory cannot be insensitive to regime changes that flip a γ-margin optimal bet for the same queried test.

Thus, if two regimes can occur under the same last observation and they induce opposite γ-margin optimal bets for the *same* queried test on nontrivial mass, then low pair-regret on the *same* distribution forces $M(h)$ to distinguish the regime whenever it matters. More generally, Corollary 4 implies that competence under mixtures of task distributions

provides a normative pressure for maintaining persistent, internal variables that track latent evaluative conditions; in embodied settings, such variables can be viewed as analogous to affective or homeostatic modulators studied in affective neuroscience that globally influence policy, attention, and learning across tasks [Ekman, 1992, Barrett, 2017]. Importantly, this is a structural claim about functional organization—global, task-general modulation of behavior under uncertainty—rather than a commitment to any particular theory of emotion or phenomenology.

Corollary 5 (Representational convergence under γ-minimality, up to invertible recoding)**.** *Fix D and $\gamma \in (0, 1/2)$. Define the γ-coarsened decision profile $\ell_D^\gamma(h) := \big(\ell_T^\gamma(h)\big)_{T\in\mathrm{supp}(D)}$, where:*

$$\ell_T^\gamma(h) := \begin{cases} L, & p_T(h) \geq \frac{1}{2} + \gamma, \\ R, & p_T(h) \leq \frac{1}{2} - \gamma, \\ \bot, & \textit{otherwise}. \end{cases}$$

Let $M_1 = f_1(h)$ and $M_2 = f_2(h)$ be two memory representations with M_j-based policies π_j. Assume, for $j = 1, 2$, that $\bar{\delta}_{\mathcal{P}}(\pi_j) \to 0$, that M_j is γ-minimal,

$$\ell_D^\gamma(h) = \ell_D^\gamma(h') \implies M_j(h) = M_j(h'),$$

and that the witnesses are γ-complete: for $\mathcal{P}$-a.e. pair,

$$\ell_D^\gamma(h) \neq \ell_D^\gamma(h') \implies D(S_\gamma(h, h')) > 0.$$

Then, on the support of $\mathcal{P}$, each M_j induces exactly the partition given by ℓ_D^γ. Hence M_1 and M_2 agree up to invertible recoding: there exist measurable maps φ, ψ such that almost surely

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

Therefore, under the *same* evaluation family, low pair-regret forces any *sufficient* memory representation to preserve exactly the γ-margin decision-relevant distinctions between histories; if two agents are also γ-minimal (no extra splitting beyond those distinctions), then their internal memory states must agree up to a relabeling (invertible recoding) on the evaluation support.

7 DISCUSSION

This work develops quantitative "selection theorems" [Wentworth, 2021]: representation-theoretic conclusions derived from performance guarantees. Across fully observed and partially observed settings, we showed that low average-case regret on structured families of action-conditioned prediction tasks *selects for* the predictive internal structure tested by the evaluation family. This yields recovery of the interventional kernel, predictive-state recovery and no-aliasing under partial observability, and further constraints from structured task families: informational modularity, regime-tracking state, and representational convergence up to invertible recoding.

Necessity is task-relative: different diagnostics select different structure. To our knowledge, these are the first quantitative selection theorems linking average-case regret over structured task families to necessary predictive-state and memory structure under partial observability. Unlike classical sufficiency results for belief representations [Sondik, 1971, Kaelbling et al., 1998], our results show that regret-bounded competence alone—without worst-case optimality or determinism—imposes concrete internal constraints. The unifying perspective is that robust competence under uncertainty compresses admissible representations: when the evaluation distribution places mass on large-margin predictive distinctions, aliasing those distinctions incurs constant regret. Thus, predictive state, memory, modular decomposition, and regime-tracking variables are not merely architectural assumptions but *consequences* of task demands.

These results resonate with empirical trends in representation learning and NeuroAI. Increasingly general task demands correlate with increasingly aligned representations across architectures and modalities, including alignment between artificial and biological systems in visual [Yamins et al., 2014], auditory [Kell* et al., 2018], motor [Sussillo et al., 2015], memory [Nayebi et al., 2021], world-modeling [Nayebi et al., 2023], and language [Schrimpf et al., 2021] brain areas, as well as between *autonomous agents* and *whole-brain data* in larval zebrafish [Keller et al., 2025]. The Contravariance Principle in NeuroAI [Cao and Yamins, 2024] and the Platonic Representation Hypothesis in AI [Huh et al., 2024] both hypothesize that general learning pressures drive convergence toward a shared statistical model of reality. Our results provide a complementary formal lens: convergence can arise from *shared* competence constraints and, under minimality, be reversibly mapped across agents as in Corollary 5.

As AI systems become increasingly capable, our results suggest that organizational regularities should emerge across architectures: belief-like predictive state, modular specialization, persistent internal state, affective-like regime tracking [Ekman, 1992, Barrett, 2017], and unified predictive representations. These regularities mirror cognitive-architecture themes such as global broadcast and modular processing [Baars, 1997, Blum and Blum, 2024], and are relevant to increasingly agentic AI systems [Long et al., 2024]; not as metaphysical commitments, but as *inevitable* structural consequences of task competence. More empirical evidence is needed for consciousness theories [Consortium et al., 2025], so we make no such claims here: subjective experience may depend on *how* these components combine, though behavioral similarity across different brains [Feather* et al., 2025] makes this less likely. Selection theorems thus formally explain how capability constrains internal organization.

Acknowledgements

We thank Lenore Blum, Manuel Blum, Dylan Hadfield-Menell, and Daniel Yamins for helpful discussions, as well as Santiago Cifuentes, Leo Kozachkov, Reece Keller, Noushin Quazi, and the anonymous reviewers for helpful feedback on a draft of this manuscript. We acknowledge the Burroughs Wellcome Fund (CASI award), Foresight Institute, and Protocol Labs for funding.

References

Bernard J Baars. *In the theater of consciousness: The workspace of the mind*. Oxford University Press, USA, 1997.

Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12 (1):1–23, 2017.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Michael Timothy Bennett. Emergent causality and the foundation of consciousness. In *International Conference on Artificial General Intelligence*, pages 52–61. Springer, 2023a.

Michael Timothy Bennett. The optimal choice of hypothesis is the weakest, not the shortest. In *International Conference on Artificial General Intelligence*, pages 42–51. Springer, 2023b.

Michael Timothy Bennett. Is complexity an illusion? In *International Conference on Artificial General Intelligence*, pages 11–21. Springer, 2024.

Michael Timothy Bennett. *How To build conscious machines*. PhD thesis, The Australian National University (Australia), 2025a.

Michael Timothy Bennett. A formal theory of optimal learning with experimental results. In *Proceedings of the Thirty-fourth International Joint Conference on Artificial Intelligence*, 2025b.

Michael Timothy Bennett. Regret is weighted forgetting. 2026.

David Blackwell. An analog of the minimax theorem for vector payoffs. 1956.

Lenore Blum and Manuel Blum. Ai consciousness is inevitable: a theoretical computer science perspective. *arXiv preprint arXiv:2403.17101*, 2024.

Byron Boots, Sajid M. Siddiqi, and Geoffrey J. Gordon. Closing the learning-planning loop with predictive state representations. *The International Journal of Robotics Research*, 30(8):954–966, 2011.

Rosa Cao and Daniel Yamins. Explanatory models in neuroscience, part 2: Functional intelligibility and the contravariance principle. *Cognitive Systems Research*, 85: 101200, 2024.

Santiago Cifuentes. General agents contain world models even under partial observability and stochasticity. *arXiv preprint arXiv:2602.03146*, 2026.

Roger C Conant and W Ross Ashby. Every good regulator of a system must be a model of that system. *International journal of systems science*, 1(2):89–97, 1970.

Cogitate Consortium, Oscar Ferrante, Urszula Gorska-Klimowska, Simon Henin, Rony Hirschhorn, Aya Khalaf, Alex Lepauvre, Ling Liu, David Richter, Yamil Vidal, et al. Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature*, pages 1–10, 2025.

Arthur P Dempster. Upper and lower probabilities induced by a multivalued mapping. In *Classic works of the Dempster-Shafer theory of belief functions*, pages 57–72. Springer, 2008.

Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

Jenelle Feather*, Meenakshi Khosla*, N Murty*, and Aran Nayebi*. Brain-model evaluations need the neuroai turing test. *arXiv preprint arXiv:2502.16238*, 2025.

Dean P Foster and Rakesh V Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(589):40–55, 1997.

Bruce A Francis and Walter Murray Wonham. The internal model principle of control theory. *Automatica*, 12(5): 457–465, 1976.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. *arXiv preprint arXiv:2310.16828*, 2023.

Alfred Harwood, Jose Faustino, and Alex Altair. Information-theoretic analysis of world models in optimal reward maximizers. *arXiv preprint arXiv:2602.12963*, 2026.

Daniel A. Herrmann, Aydin Mohseni, Benjamin A. Levinstein, and Bruce Rushing. A bayesian reduction of causation, 2026. Manuscript, forthcoming.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

Miroslav Kárnỳ. Axiomatisation of fully probabilistic design revisited. *Systems & Control Letters*, 141:104719, 2020.

Miroslav Kárnỳ and Tomáš Kroupa. Axiomatisation of fully probabilistic design. *Information Sciences*, 186(1): 105–113, 2012.

Alexander JE Kell*, Daniel LK Yamins*, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

Reece Keller, Alyn Kirsch, Felix Pei, Xaq Pitkow, Leo Kozachkov, and Aran Nayebi. Intrinsic goals for autonomous agents: Model-based exploration in virtual zebrafish predicts ethological behavior and whole-brain dynamics. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Khimya Khetarpal, Gheorghe Comanici, Jonathan Richens, Jeremy Shar, Fei Xia, Laurent Orseau, Aleksandra Faust, and Doina Precup. Affordances enable partial world modeling with llms. *arXiv preprint arXiv:2602.10390*, 2026.

Michael L. Littman, Richard S. Sutton, and Satinder P. Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. Taking ai welfare seriously. *arXiv preprint arXiv:2411.00986*, 2024.

Aran Nayebi, Alexander Attinger, Malcolm Campbell, Kiah Hardcastle, Isabel Low, Caitlin Mallory, Gabriel Mel, Ben Sorscher, Alex Williams, Surya Ganguli, Lisa M Giocomo, and Daniel LK Yamins. Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Aran Nayebi, Rishi Rajalingham, Mehrdad Jazayeri, and Guangyu Robert Yang. Neural foundations of mental simulation: Future prediction of latent representations on dynamic scenes. *Advances in Neural Information Processing Systems*, 36, 2023.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonathan Richens and Tom Everitt. Robust agents learn causal world models. *arXiv preprint arXiv:2402.10877*, 2024.

Jonathan Richens, David Abel, Alexis Bellot, and Tom Everitt. General agents contain world models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, 2025. Also available as arXiv:2506.01622.

Leonard J Savage. *The foundations of statistics*. Courier Corporation, 1954.

Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Glenn Shafer and Vladimir Vovk. *Probability and finance: It's only a game!*, volume 491. John Wiley & Sons, 2005.

Satinder Singh, Michael R. James, and Matthew R. Rudary. Predictive state representations: A new theory for modeling dynamical systems. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2004.

Edward J. Sondik. *The Optimal Control of Partially Observable Markov Processes*. PhD thesis, Stanford University, 1971.

David Sussillo, Mark M Churchland, Matthew T Kaufman, and Krishna V Shenoy. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience*, 18(7):1025–1033, 2015. doi: 10.1038/nn.4042.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior, 2nd rev. 1947.

Shengjie Wang, Shaohuai Liu, Weirui Ye, Jiacheng You, and Yang Gao. Efficientzero v2: Mastering discrete and continuous control with limited data. *arXiv preprint arXiv:2403.00564*, 2024.

John Wentworth. Selection theorems: A program for understanding agents, 2021. URL `https://www.lesswrong.com/posts/G2Lne2Fi7Qra5Lbuf/selection-theorems-a-program-for-understanding-agents`. LessWrong / AI Alignment Forum (online post).

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Ian H Witten. An adaptive optimal controller for discrete-time markov environments. *Information and control*, 34 (4):286–295, 1977.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111.

What Capable Agents Must Know: Selection Theorems for Robust Decision-Making under Uncertainty (Supplementary Material)

Aran Nayebi[1]

[1]Machine Learning Department and Neuroscience & Robotics Institutes, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

A PROOF OF LEMMA 1

Proof. Observe that the success probability defined in (1) can be rewritten as

$$V = u_R + q(u_L - u_R),$$

which is linear in q. Thus, the optimal success probability is achieved at the endpoints, $V^\star := \max_{q\in[0,1]} V = \max\{V(0), V(1)\} = \max\{u_L, u_R\}$.

Assume wlog $u_L \geq u_R$. Then $V^\star = u_L$, $w = 1 - q$, and $V = (1-w)\,u_L + w\,u_R = u_L - w\,(u_L - u_R)$. Therefore,

$$\delta = 1 - \frac{u_L - w\,(u_L - u_R)}{u_L} = w\frac{u_L - u_R}{u_L}.$$

The other case is symmetric.

In the special case that $u_R = 1 - u_L$, then we have that

$$V^\star = \max\{u_L, u_R\} = \max\{u_L, 1 - u_L\} = \tfrac{1}{2} + m.$$

Indeed, since $m := \left|u_L - \frac{1}{2}\right|$, we can write $u_L = \frac{1}{2} + m$ or $u_L = \frac{1}{2} - m$. In the first case, $1 - u_L = \frac{1}{2} - m$, and in the second case $1 - u_L = \frac{1}{2} + m$. In either case, since $m \geq 0$,

$$\max\{u_L, 1 - u_L\} = \tfrac{1}{2} + m.$$

Moreover, in both cases,

$$|u_L - u_R| = |u_L - (1 - u_L)| = |2u_L - 1| = 2m.$$

Substituting these expressions into (5), yields

$$\delta = w \cdot \frac{|u_L - u_R|}{V^\star} = w \cdot \frac{2m}{\frac{1}{2} + m} = w \cdot \frac{4m}{1 + 2m}, \tag{28}$$

which proves the stated identity.

Now suppose that $m \geq \gamma > 0$. Since the function

$$f(m) := \frac{4m}{1 + 2m}$$

is increasing for $m \geq 0$, since $f'(m) = 4/(1 + 2m)^2 > 0$, then we have

$$\frac{4m}{1 + 2m} \geq \frac{4\gamma}{1 + 2\gamma} =: c(\gamma). \tag{29}$$

Combining (29) with (28) gives

$$\delta \geq w\,c(\gamma),$$

and hence gives us (7). □

B PROOF OF THEOREM 1

Proof. Fix a quadruple (s, a, s', k) and let $X \sim \mathrm{Bin}(n, p)$ with $p := P_{ss'}(a)$, and define

$$F(k) := \Pr[X \le k].$$

In what follows, we let $\widehat{P}_{ss'}(a)$ denote the *unclipped* quantity inside (8). This is sufficient to upper bound $|\widehat{P}_{ss'}(a) - P_{ss'}(a)|$, since clipping onto $[0, 1]$ cannot increase distance to $P_{ss'}(a) \in [0, 1]$.

1. Pointwise regret lower-bounds wrong-branch mass at margin m_k. By Richens et al. [2025, Lemma 6], the two disjuncts $G^{(n,1)}_{s,a,s',k}$ and $G^{(n,2)}_{s,a,s',k}$ of the composite goal $G^{(n)}_{s,a,s',k}$ have optimal satisfaction probabilities $F(k)$ and $1 - F(k)$, respectively:

$$\begin{aligned} &V^\star(s_0; G^{(n,i)}_{s,a,s',k}) := \max_{\pi'} \Pr_{\pi'}(G^{(n,i)}_{s,a,s',k} \mid s_0), \qquad i \in \{1, 2\}, \\ &V^\star(s_0; G^{(n,1)}_{s,a,s',k}) = F(k), \\ &V^\star(s_0; G^{(n,2)}_{s,a,s',k}) = 1 - F(k). \end{aligned} \tag{30}$$

Hence, by Lemma 1, the optimal satisfaction probability of the *overall* disjunction in (4) therefore is

$$\begin{aligned} &V^\star(s_0; G^{(n)}_{s,a,s',k}) = \max\{F(k), 1 - F(k)\} = \tfrac{1}{2} + m_k \\ &m_k := \big|F(k) - \tfrac{1}{2}\big|, \end{aligned}$$

Let w_k denote the probability that π selects the *suboptimal* branch at threshold k:

$$w_k := \begin{cases} 1 - q_{s,a,s',k}, & \text{if } V^\star(s_0; G^{(n,1)}_{s,a,s',k}) \ge V^\star(s_0; G^{(n,2)}_{s,a,s',k}), \\ q_{s,a,s',k}, & \text{otherwise,} \end{cases}$$

and let B_k denote the event that π selects the disjunct with larger optimal satisfaction probability at threshold k. Then

$$\Pr_\pi(B_k) = 1 - w_k \quad \text{and} \quad \Pr_\pi(B_k^c) = w_k,$$

since w_k is the probability that π selects the suboptimal disjunct. By the law of total probability,

$$\begin{aligned} &V^\pi(s_0; G^{(n)}_{s,a,s',k}) \\ &= \Pr_\pi\big(G^{(n)}_{s,a,s',k} \mid B_k\big) \Pr_\pi(B_k) + \Pr_\pi\big(G^{(n)}_{s,a,s',k} \mid B_k^c\big) \Pr_\pi(B_k^c) \\ &= (1 - w_k) \Pr_\pi\big(G^{(n)}_{s,a,s',k} \mid B_k\big) + w_k \Pr_\pi\big(G^{(n)}_{s,a,s',k} \mid B_k^c\big). \end{aligned}$$

Moreover, conditional on B_k (resp. B_k^c), π selects the disjunct with larger (resp. smaller) optimal satisfaction probability, so the corresponding success probability under π is at most the optimal success probability of that selected disjunct. Hence,

$$\begin{aligned} &V^\pi(s_0; G^{(n)}_{s,a,s',k}) \\ &\le (1 - w_k) \max\Big\{V^\star(s_0; G^{(n,1)}_{s,a,s',k}),\, V^\star(s_0; G^{(n,2)}_{s,a,s',k})\Big\} \\ &\quad + w_k \min\Big\{V^\star(s_0; G^{(n,1)}_{s,a,s',k}),\, V^\star(s_0; G^{(n,2)}_{s,a,s',k})\Big\} \\ &= (1 - w_k)\big(\tfrac{1}{2} + m_k\big) + w_k\big(\tfrac{1}{2} - m_k\big). \end{aligned}$$

Applying Lemma 1 with the identification

$$\begin{aligned} &u_L = F(k), \qquad u_R = 1 - F(k), \\ &m = m_k, \qquad w = w_k, \end{aligned}$$

we obtain

$$\delta_{s,a,s',k}(\pi; s_0) := 1 - \frac{V^\pi}{V^\star} \;\ge\; w_k \, \frac{4m_k}{1 + 2m_k}.$$

In particular, on the event $\{m_k \geq \gamma\}$,

$$w_k \;\leq\; \frac{\delta_{s,a,s',k}(\pi; s_0)}{c(\gamma)}. \tag{31}$$

2. Estimating the binomial median from the policy's disjunct probabilities. Let k_{med} be the (lower) median index of X, i.e.

$$k_{\text{med}} := \min\{k : F(k) \geq \tfrac{1}{2}\}.$$

By definition of k_{med}, the disjunct with larger optimal satisfaction probability is $G^{(n,2)}_{s,a,s',k}$ for $k < k_{\text{med}}$, and $G^{(n,1)}_{s,a,s',k}$ for $k \geq k_{\text{med}}$. Define

$$\widehat{k}_{\text{med}} \;:=\; \sum_{k=0}^{n} (1 - q_{s,a,s',k}), \tag{32}$$

which is the expected number of thresholds at which the policy π selects the disjunct $G^{(n,2)}_{s,a,s',k}$, by definition of $q_{s,a,s',k}$. If π were optimal on these binary choices, $(1 - q_{s,a,s',k})$ would equal $\mathbf{1}\{k < k_{\text{med}}\}$, hence $\widehat{k}_{\text{med}} = k_{\text{med}}$ since $n \geq k_{\text{med}}$.

In general,

$$\begin{aligned} \left|\widehat{k}_{\text{med}} - k_{\text{med}}\right| &= \left| \sum_{k=0}^{n} \Big((1 - q_{s,a,s',k}) - \mathbf{1}\{k < k_{\text{med}}\} \Big) \right| \\ &\leq \sum_{k=0}^{n} w_k. \end{aligned}$$

Split $\{0, \ldots, n\}$ into $K_\gamma := \{k : m_k < \gamma\}$ and its complement $K_\gamma^c := \{k : m_k \geq \gamma\}$. Using the trivial bound $w_k \leq 1$ on K_γ, where $m_k < \gamma$ and the disjuncts become indistinguishable as $m_k \to 0$, so regret cannot constrain the policy's choice, and the regret-based bound (31) on K_γ^c,

$$\begin{aligned} \left|\widehat{k}_{\text{med}} - k_{\text{med}}\right| \;&\leq\; |K_\gamma| \;+\; \frac{1}{c(\gamma)} \sum_{k \notin K_\gamma} \delta_{s,a,s',k}(\pi; s_0) \\ &\leq\; |K_\gamma| \;+\; \frac{1}{c(\gamma)} \sum_{k=0}^{n} \delta_{s,a,s',k}(\pi; s_0). \end{aligned} \tag{33}$$

3. Controlling $|K_\gamma|$. Let $\mu := \mathbb{E}[X] = np$ and $\sigma^2 := \text{Var}(X) = np(1-p)$. For any $t > 0$, if $k \geq \mu + t\sigma$, then by the one-sided Chebyshev inequality,

$$\Pr[X \geq k] = \Pr[X - \mu \geq t\sigma] \;\leq\; \frac{1}{1 + t^2},$$

and hence

$$F(k) = \Pr[X \leq k] \;\geq\; 1 - \frac{1}{1 + t^2} = \frac{t^2}{1 + t^2}.$$

If $t \geq 1$, then $F(k) \geq \frac{1}{2}$ in this regime. Thus, it follows that

$$m_k = F(k) - \tfrac{1}{2} \;\geq\; \frac{t^2}{1 + t^2} - \tfrac{1}{2} = \frac{t^2 - 1}{2(1 + t^2)}. \tag{34}$$

Choosing

$$t_\gamma := \sqrt{\frac{1 + 2\gamma}{1 - 2\gamma}} \;\geq 1,$$

since $\gamma > 0$, makes the right-hand side of (34) equal to γ, so $m_k \geq \gamma$ whenever $k \geq \mu + t_\gamma \sigma$.

By symmetry, applying the same argument to $-X$ yields that $m_k \geq \gamma$ whenever $k \leq \mu - t_\gamma \sigma$. Equivalently, by contraposition,

$$k \in K_\gamma = \{k : m_k < \gamma\} \quad \Longrightarrow \quad |k - \mu| < t_\gamma \sigma.$$

Therefore,

$$K_\gamma \;\subseteq\; (\mu - t_\gamma \sigma,\; \mu + t_\gamma \sigma),$$

and since k ranges over integers, the number of such indices is bounded by

$$\begin{aligned} |K_\gamma| &\le \lceil 2t_\gamma \sigma \rceil + 1 \\ &\le 2t_\gamma \sigma + 2 \\ &= 2t_\gamma \sqrt{np(1-p)} + 2. \end{aligned} \tag{35}$$

Note that this Chebyshev step is deliberately distribution-free and can be loose for small n; sharper binomial concentration would improve constants without changing the selection argument.

4. From median error to transition-probability error. Define $\widehat{p} := \widehat{P}_{ss'}(a)$ as in (8), i.e. $\widehat{p} = \frac{1}{n}(\widehat{k}_{\mathrm{med}} - \frac{1}{2})$ by definition (32). A standard binomial fact is that the median differs from the mean by at most 1: $|k_{\mathrm{med}} - np| \le 1$; equivalently, $k_{\mathrm{med}} \in \{\lfloor np \rfloor, \lceil np \rceil\}$. Hence,

$$\begin{aligned} |\widehat{p} - p| &\le \left|\widehat{p} - \frac{k_{\mathrm{med}}}{n}\right| + \left|\frac{k_{\mathrm{med}}}{n} - p\right| \\ &\le \frac{|\widehat{k}_{\mathrm{med}} - k_{\mathrm{med}}| + \frac{1}{2}}{n} + \frac{1}{n} \\ &\le \frac{|\widehat{k}_{\mathrm{med}} - k_{\mathrm{med}}|}{n} + O\left(\frac{1}{n}\right). \end{aligned}$$

Combining (33) and (35) gives, for this fixed (s, a, s'),

$$\begin{aligned} |\widehat{p} - p| \le 2t_\gamma \sqrt{\frac{p(1-p)}{n}} &+ \frac{1}{nc(\gamma)} \sum_{k=0}^{n} \delta_{s,a,s',k}(\pi; s_0) \\ &+ O\left(\frac{1}{n}\right). \end{aligned}$$

Finally, average over $(s, a, s') \sim \mathrm{Unif}(\mathcal{S} \times \mathcal{A} \times \mathcal{S})$ and use the global assumption (9) to bound

$$\begin{aligned} &\mathbb{E}_{(s,a,s')}\left[\frac{1}{n} \sum_{k=0}^{n} \delta_{s,a,s',k}(\pi; s_0)\right] \\ &= \frac{n+1}{n} \mathbb{E}_{(s,a,s',k)}[\delta_{s,a,s',k}(\pi; s_0)] \\ &\le \frac{n+1}{n} \bar{\delta}, \end{aligned}$$

which yields (10). □

C PROOF OF COROLLARIES 1 AND 2

C.1 PROOF OF COROLLARY 1

Proof. By the $\varepsilon_{\mathrm{cMP}}$-*approximate* causal Markov-process assumption (11), for each (s, a, s') we have $|P_{ss'}(a) - P^{\mathrm{do}}_{ss'}(a)| \le \varepsilon_{\mathrm{cMP}}$. Thus, by the triangle inequality,

$$\begin{aligned} |\widehat{P}_{ss'}(a) - P^{\mathrm{do}}_{ss'}(a)| &\le |\widehat{P}_{ss'}(a) - P_{ss'}(a)| + |P_{ss'}(a) - P^{\mathrm{do}}_{ss'}(a)| \\ &\le |\widehat{P}_{ss'}(a) - P_{ss'}(a)| + \varepsilon_{\mathrm{cMP}}. \end{aligned}$$

Taking expectations over $(s, a, s') \sim \mathrm{Unif}(\mathcal{S} \times \mathcal{A} \times \mathcal{S})$, the bound follows immediately from Theorem 1 by substitution. □

C.2 PROOF OF COROLLARY 2

Proof. Two structural causal models can share the same interventional kernel $P(S_{t+1} \mid S_t = s, \mathrm{do}(A_t = a))$ for all (s, a) while differing in counterfactual couplings.

Fix a single state s, binary actions $\{0, 1\}$, and binary next state. Let $U \sim \mathrm{Bernoulli}(1/2)$. Model (I): $S_{t+1} = U$. Model (II): $S_{t+1} = A_t \oplus U$.

Both satisfy $P(S_{t+1} = 1 \mid S_t = s, \mathrm{do}(A_t = a)) = 1/2$ for $a \in \{0, 1\}$, so their interventional kernels coincide. However, conditioning on $A_t = 0$ and $S_{t+1} = 1$ (hence $U = 1$), the counterfactual under $A_t = 1$ gives $S^1_{t+1} = 1$ in Model (I) but $S^1_{t+1} = 0$ in Model (II). Thus Level 3 counterfactuals are not identified by the interventional kernel. □

D PROOF OF THEOREM 2

Proof. Fix (h, T) and write $p := p_T(h)$, $m := m_T(h)$, $q := q_T(h)$. If $p \geq \frac{1}{2}$ then $p = \frac{1}{2} + m$ and the suboptimal mass is $w_T(h) = 1 - q$. Using (13)-(14):

$$V^{\pi}(h; g_T) = q(\tfrac{1}{2} + m) + (1 - q)(\tfrac{1}{2} - m) = \tfrac{1}{2} - m + 2mq,$$
$$V^{\star}(h; g_T) = \tfrac{1}{2} + m.$$

Thus

$$\begin{aligned} \delta_T(\pi; h) = 1 - \frac{\frac{1}{2} - m + 2mq}{\frac{1}{2} + m} &= \frac{2m(1 - q)}{\frac{1}{2} + m} \\ &= w_T(h) \frac{4m}{1 + 2m}. \end{aligned} \tag{36}$$

If $p < \frac{1}{2}$, the symmetric calculation yields the same identity $\delta_T(\pi; h) = w_T(h) \cdot \frac{4m}{1+2m}$ (now $w_T(h) = q$). Therefore, on the event $\{m \geq \gamma\}$ we have

$$\delta_T(\pi; h) \;\geq\; w_T(h) \cdot \frac{4\gamma}{1 + 2\gamma} \;=\; w_T(h)\, c(\gamma).$$

Taking expectations over $h \sim \mathcal{H}$ and $T \sim D$ and using (15):

$$\begin{aligned} \bar{\delta} \geq\; \mathbb{E}[\delta_T(\pi; h)] &\geq \mathbb{E}[\delta_T(\pi; h)\, \mathbf{1}\{m_T(h) \geq \gamma\}] \\ &\geq\; c(\gamma)\, \mathbb{E}[w_T(h)\, \mathbf{1}\{m_T(h) \geq \gamma\}], \end{aligned}$$

which proves (17). The conditional bound (18) follows by dividing by $q_\gamma = \Pr(m_T(H) \geq \gamma)$. □

E PROOF OF PROPOSITION 1

Proof. Fix any $p, q \in (\frac{1}{2}, 1)$ with $p \neq q$. Let $\mathcal{A}$ be any finite action space with $|\mathcal{A}| \geq 2$. For each $r \in \{p, q\}$, define a POMDP $E_r = (\mathcal{X}, \mathcal{A}, \mathcal{O}, T, Z, \mu_0)$ as follows:

$$\mathcal{X} = \{x_0, x_1, y_0, y_1\}, \qquad \mathcal{O} = \{u, 0, 1\},$$

with initial distribution

$$\mu_0(x_0) = r, \qquad \mu_0(x_1) = 1 - r, \qquad \mu_0(y_0) = \mu_0(y_1) = 0.$$

The observation kernel is

$$Z(u \mid x_0) = Z(u \mid x_1) = 1, \qquad Z(0 \mid y_0) = 1, \qquad Z(1 \mid y_1) = 1.$$

For every action $a \in \mathcal{A}$, the transition kernel is

$$T(y_0 \mid x_0, a) = 1, \qquad T(y_1 \mid x_1, a) = 1, \qquad T(y_0 \mid y_0, a) = 1, \qquad T(y_1 \mid y_1, a) = 1.$$

Let $h = (u)$ be the initial history. Fix any test $T = (\alpha, W)$ with $|\alpha| = k$. Since actions do not affect the dynamics, conditional on h the future observation sequence is deterministically either 0^k or 1^k. Therefore

$$p_T^{E_r}(h) = r\,\mathbf{1}\{0^k \in W\} + (1-r)\,\mathbf{1}\{1^k \in W\}. \tag{37}$$

Hence $p_T^{E_r}(h)$ can only take one of the four values

$$0, \qquad 1-r, \qquad r, \qquad 1.$$

Since $r > \frac{1}{2}$, the unique optimal fair bet is:

- report L if $0^k \in W$;
- report R if $0^k \notin W$.

This rule is independent of the value of $r \in (\frac{1}{2}, 1)$. Therefore every optimal policy for the fair betting goals g_T induces exactly the same answers on all tests at h in E_p and E_q.

On the other hand, if $T^\star = (\alpha^\star, \{0^{|\alpha^\star|}\})$ for any fixed action sequence $\alpha^\star$, then by (37),

$$p_{T^\star}^{E_p}(h) = p \qquad \text{and} \qquad p_{T^\star}^{E_q}(h) = q,$$

so $p_{T^\star}^{E_p}(h) \neq p_{T^\star}^{E_q}(h)$. Thus

$$\eta_{\mathcal{T}}^{E_p}(h) \neq \eta_{\mathcal{T}}^{E_q}(h).$$

This proves that exact optimal query access to the fair betting goals does not, in general, identify the predictive state. □

F PROOF OF THEOREM 3

Proof. Fix (h, T) and write

$$p := p_T(h), \qquad q_k := q_{T,\lambda_k}(h), \qquad \hat{p} := \frac{1}{K}\sum_{k=1}^{K} q_k.$$

Let

$$r_k := V^\star(h; g_{T,\lambda_k}) - V^\pi(h; g_{T,\lambda_k})$$

denote the unnormalized regret at threshold λ_k. Then

$$r_k = \begin{cases} (1-q_k)(p-\lambda_k), & \lambda_k \le p, \\ q_k(\lambda_k - p), & \lambda_k > p. \end{cases} \tag{38}$$

Let

$$J := \#\{k : \lambda_k \le p\} = \left\lfloor Kp + \frac{1}{2} \right\rfloor.$$

Averaging (38) over k gives

$$R := \frac{1}{K}\sum_{k=1}^{K} r_k = \frac{1}{K}\sum_{k \le J}(p - \lambda_k) - p\hat{p} + \frac{1}{K}\sum_{k=1}^{K}\lambda_k q_k. \tag{39}$$

We now bound the two terms on the right-hand side of (39).

First,

$$\frac{1}{K}\sum_{k \le J}(p-\lambda_k) = \frac{Jp}{K} - \frac{1}{K}\sum_{k=1}^{J}\frac{k-\frac{1}{2}}{K} = \frac{Jp}{K} - \frac{J^2}{2K^2}.$$

Since

$$\frac{Jp}{K} - \frac{J^2}{2K^2} = \frac{p^2}{2} - \frac{(Kp-J)^2}{2K^2},$$

and $|Kp - J| \le \frac{1}{2}$, it follows that

$$\frac{1}{K}\sum_{k\le J}(p-\lambda_k) \ge \frac{p^2}{2} - \frac{1}{8K^2}. \tag{40}$$

Second, among all choices $q_k \in [0,1]$ with fixed average $\hat{p}$, the weighted sum $\sum_k \lambda_k q_k$ is minimized by placing as much mass as possible on the smallest thresholds. Let $L = \lfloor K\hat{p} \rfloor$ and $\beta = K\hat{p} - L \in [0,1]$. The minimum is attained by

$$q_1 = \cdots = q_L = 1, \qquad q_{L+1} = \beta, \qquad q_{L+2} = \cdots = q_K = 0,$$

(with the convention that if $\beta = 0$ the partial entry is omitted), and equals

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\lambda_k q_k &\ge \frac{1}{K}\left(\sum_{k=1}^{L}\frac{k-\frac{1}{2}}{K} + \beta\frac{L+\frac{1}{2}}{K}\right)\\ &= \frac{(L+\beta)^2 + \beta(1-\beta)}{2K^2}\\ &\ge \frac{\hat{p}^2}{2}.\end{aligned} \tag{41}$$

Substituting (40) and (41) into (39) yields

$$R \ge \frac{p^2}{2} - \frac{1}{8K^2} - p\hat{p} + \frac{\hat{p}^2}{2} = \frac{(\hat{p}-p)^2}{2} - \frac{1}{8K^2}. \tag{42}$$

Now

$$r_k = \delta_{T,\lambda_k}(\pi;h)\, V^\star(h; g_{T,\lambda_k}) \le \delta_{T,\lambda_k}(\pi;h),$$

since $V^\star(h; g_{T,\lambda_k}) \le 1$. Therefore

$$R \le \frac{1}{K}\sum_{k=1}^{K}\delta_{T,\lambda_k}(\pi;h). \tag{43}$$

Combining (42) and (43),

$$\big(\hat{p} - p_T(h)\big)^2 \le 2\Big(\frac{1}{K}\sum_{k=1}^{K}\delta_{T,\lambda_k}(\pi;h)\Big) + \frac{1}{4K^2}. \tag{44}$$

Finally, average (44) over $h \sim \mathcal{H}$ and $T \sim D$, and use (20), obtaining

$$\mathbb{E}_{h\sim\mathcal{H}}\,\mathbb{E}_{T\sim D}\Big[\big(\hat{p}_T(h) - p_T(h)\big)^2\Big] \le 2\bar{\delta}_K + \frac{1}{4K^2},$$

which proves (21).

For the vector statement, if D is uniform on $\mathcal{T}_\ell = \{T_1, \ldots, T_d\}$ then

$$\begin{aligned}\mathbb{E}_{h\sim\mathcal{H}}\Big[\frac{1}{d}\big\|\hat{\eta}_{\mathcal{T}_\ell}(h) - \eta_{\mathcal{T}_\ell}(h)\big\|_2^2\Big] &= \mathbb{E}_{h\sim\mathcal{H}}\Big[\frac{1}{d}\sum_{j=1}^{d}\big(\hat{p}_{T_j}(h) - p_{T_j}(h)\big)^2\Big]\\ &= \mathbb{E}_{h\sim\mathcal{H}}\,\mathbb{E}_{T\sim D}\Big[\big(\hat{p}_T(h) - p_T(h)\big)^2\Big]\\ &\le 2\bar{\delta}_K + \frac{1}{4K^2},\end{aligned}$$

proving (22). □

G PROOF OF THEOREM 4

Proof. Apply Theorem 3 with $\mathcal{H}$ uniform on $\{h^1, \dots, h^d\}$ and with the test family defined as the indexed collection

$$\mathcal{T}^+_{\text{idx}} := \{T_1, \dots, T_d\} \ \cup \ \{\sigma \circ T_j : \sigma \in \mathcal{A} \times \mathcal{O},\ 1 \le j \le d\},$$

counting each pair (σ, j) separately. Then

$$|\mathcal{T}^+_{\text{idx}}| = d\big(1 + |\mathcal{A}||\mathcal{O}|\big).$$

By (21),

$$\frac{1}{d\,|\mathcal{T}^+_{\text{idx}}|} \sum_{i=1}^{d} \sum_{T \in \mathcal{T}^+_{\text{idx}}} \big(\hat{p}_T(h^i) - p_T(h^i)\big)^2 \ \le \ \varepsilon_K. \tag{45}$$

By construction, the sum over $T \in \mathcal{T}^+_{\text{idx}}$ exactly enumerates all entries of S (from T_j) and all entries of each Y_σ (from $\sigma \circ T_j$), so the left-hand side of (45) equals

$$\frac{1}{d^2(1 + |\mathcal{A}||\mathcal{O}|)} \left(\|\hat{S} - S\|_F^2 + \sum_{\sigma} \|\hat{Y}_\sigma - Y_\sigma\|_F^2 \right),$$

which proves (24).

Let $\kappa := \|S^{-1}\|_2$. From (24),

$$\|\hat{S} - S\|_2 \le \|\hat{S} - S\|_F \le d\sqrt{(1 + |\mathcal{A}||\mathcal{O}|)\,\varepsilon_K}.$$

Under (25), this implies

$$\|S^{-1}(\hat{S} - S)\|_2 \le \frac{1}{2}.$$

Hence $\hat{S}$ is invertible, and the standard Neumann series perturbation bound gives

$$\|\hat{S}^{-1}\|_2 \le 2\kappa, \qquad \|\hat{S}^{-1} - S^{-1}\|_2 \le 2\kappa^2 \|\hat{S} - S\|_F. \tag{46}$$

For each σ,

$$\hat{B}_\sigma - B_\sigma = (\hat{Y}_\sigma - Y_\sigma)\hat{S}^{-1} + Y_\sigma(\hat{S}^{-1} - S^{-1}).$$

Using (46),

$$\|\hat{B}_\sigma - B_\sigma\|_F \le 2\kappa\, \|\hat{Y}_\sigma - Y_\sigma\|_F + 2\kappa^2\, \|Y_\sigma\|_2\, \|\hat{S} - S\|_F.$$

Squaring and using $(u + v)^2 \le 2u^2 + 2v^2$,

$$\|\hat{B}_\sigma - B_\sigma\|_F^2 \le 8\kappa^2 \|\hat{Y}_\sigma - Y_\sigma\|_F^2 + 8\kappa^4 \|Y_\sigma\|_2^2 \|\hat{S} - S\|_F^2.$$

Summing over σ and applying (24) yields (26). □

H PROOF OF THEOREM 5

Proof. Fix any M-based policy π. Consider any pair (h, h') with $(h, h') \in \mathsf{Alias}_M$. Because π is M-based and $M(h) = M(h')$, for every test T we have identical bet distributions: $q_T(h) = q_T(h') =: q_T$.

Now fix a test $T \in S_\gamma(h, h')$. By assumption, $p_T(h) \ge \frac{1}{2} + \gamma$ so the optimal bet at h is L and thus $w_T(h) = 1 - q_T$; while $p_T(h') \le \frac{1}{2} - \gamma$ so the optimal bet at h' is R and thus $w_T(h') = q_T$. Therefore,

$$\frac{1}{2}\big(w_T(h) + w_T(h')\big) \ = \ \frac{1}{2}.$$

By the pointwise identity (36) from the proof of Theorem 2, when $m_T(\cdot) \geq \gamma$ we have $\delta_T(\pi;\cdot) \geq c(\gamma)\, w_T(\cdot)$. Hence for $T \in S_\gamma(h, h')$,

$$\begin{aligned}&\frac{1}{2}\big(\delta_T(\pi;h) + \delta_T(\pi;h')\big)\\ &\geq \frac{1}{2}\, c(\gamma)\big(w_T(h) + w_T(h')\big) \;=\; \frac{c(\gamma)}{2}.\end{aligned}$$

Taking expectations over $(h, h') \sim \mathcal{P}$ and $T \sim D$ and restricting to the event $\{(h, h') \in \mathsf{Alias}_M,\ T \in S_\gamma(h, h')\}$ yields

$$\bar{\delta}_{\mathcal{P}}(\pi) \;\geq\; q_\gamma^{\mathsf{Alias}}(M) \cdot \frac{c(\gamma)}{2},$$

which proves (27). □

I PROOF OF COROLLARY 3

Proof. For each i, define the *blockwise* pair-averaged regret

$$\bar{\delta}_{\mathcal{P},i}(\pi) := \mathbb{E}_{(h,h')\sim\mathcal{P}}\, \frac{1}{2}\Big(\mathbb{E}_{T\sim D_i}[\delta_T(\pi;h)] + \mathbb{E}_{T\sim D_i}[\delta_T(\pi;h')]\Big).$$

Applying Theorem 5 with test distribution D_i yields

$$\bar{\delta}_{\mathcal{P},i}(\pi) \;\geq\; q_{\gamma,i}^{\mathsf{Alias}}(M)\frac{c(\gamma)}{2},$$

so

$$q_{\gamma,i}^{\mathsf{Alias}}(M) \leq \frac{2\,\bar{\delta}_{\mathcal{P},i}(\pi)}{c(\gamma)}. \tag{47}$$

Now relate $\bar{\delta}_{\mathcal{P}}(\pi)$ (under D) to the $\bar{\delta}_{\mathcal{P},i}(\pi)$. Because $D = \sum_{i=1}^K p_i D_i$, for any fixed history h we have

$$\mathbb{E}_{T\sim D}[\delta_T(\pi;h)] = \sum_{i=1}^K p_i\, \mathbb{E}_{T\sim D_i}[\delta_T(\pi;h)].$$

Substituting this identity into the definition of $\bar{\delta}_{\mathcal{P}}(\pi)$ and exchanging sums/expectations gives

$$\bar{\delta}_{\mathcal{P}}(\pi) = \sum_{i=1}^K p_i\, \bar{\delta}_{\mathcal{P},i}(\pi).$$

Since all terms are nonnegative, $\bar{\delta}_{\mathcal{P}}(\pi) \geq p_i\, \bar{\delta}_{\mathcal{P},i}(\pi)$, hence

$$\bar{\delta}_{\mathcal{P},i}(\pi) \;\leq\; \frac{\bar{\delta}_{\mathcal{P}}(\pi)}{p_i}. \tag{48}$$

Combining (47) and (48) yields

$$q_{\gamma,i}^{\mathsf{Alias}}(M) \;\leq\; \frac{2}{c(\gamma)} \cdot \frac{\bar{\delta}_{\mathcal{P}}(\pi)}{p_i},$$

as claimed. □

J PROOF OF COROLLARIES 4 AND 5

For the next two corollaries, it will be useful to have the following lemma:

Lemma 2 (Low pair-regret ⇒ small aliasing mass on γ-separations)**.** *Work in the setting of Theorem 5 with $(\mathcal{P}, D, \gamma)$ and witness sets $S_\gamma(h, h')$. Let π be M-based. Then*

$$\Pr_{(h,h')\sim\mathcal{P},\ T\sim D}\Big(M(h) = M(h')\ \wedge\ T \in S_\gamma(h, h')\Big) \le \frac{2\,\bar{\delta}_{\mathcal{P}}(\pi)}{c(\gamma)}.$$

Equivalently,

$$\mathbb{E}_{(h,h')\sim\mathcal{P}}\Big[\mathbf{1}\{M(h) = M(h')\} \cdot D(S_\gamma(h, h'))\Big] \ \le\ \frac{2\,\bar{\delta}_{\mathcal{P}}(\pi)}{c(\gamma)}.$$

In particular, if $\bar{\delta}_{\mathcal{P}}(\pi) = 0$ then $\mathbf{1}\{M(h) = M(h')\} \cdot D(S_\gamma(h, h')) = 0$ for $\mathcal{P}$ almost everywhere on (h, h').

Proof. Theorem 5 gives the lower bound

$$\bar{\delta}_{\mathcal{P}}(\pi)\ \ge\ q_\gamma^{\mathsf{Alias}}(M) \cdot \frac{c(\gamma)}{2},$$

where

$$q_\gamma^{\mathsf{Alias}}(M) = \Pr_{(h,h')\sim\mathcal{P},\ T\sim D}\Big(M(h) = M(h')\ \wedge\ T \in S_\gamma(h, h')\Big).$$

Rearranging yields the first inequality. For the second display, note that

$$q_\gamma^{\mathsf{Alias}}(M) = \mathbb{E}_{(h,h')\sim\mathcal{P}}\Big[\mathbf{1}\{M(h) = M(h')\} \cdot D(S_\gamma(h, h'))\Big],$$

by the law of total expectation over $T \sim D$. The final claim follows because a nonnegative random variable with zero expectation is zero almost surely. □

J.1 PROOF OF COROLLARY 4

Proof. By assumption, the witness set is supported on regime-mismatched pairs in the sense that

$$T \in S_\gamma(h, h')\ \Longrightarrow\ I(h) \neq I(h').$$

Therefore the event in the corollary simplifies:

$$\{M(h) = M(h')\ \wedge\ I(h) \neq I(h')\ \wedge\ T \in S_\gamma(h, h')\} = \{M(h) = M(h')\ \wedge\ T \in S_\gamma(h, h')\},$$

since $T \in S_\gamma(h, h')$ already implies $I(h) \neq I(h')$. Taking probabilities under $(h, h') \sim \mathcal{P}$ and $T \sim D$ yields

$$\Pr\big(M(h) = M(h')\ \wedge\ I(h) \neq I(h')\ \wedge\ T \in S_\gamma(h, h')\big) = \Pr\big(M(h) = M(h')\ \wedge\ T \in S_\gamma(h, h')\big).$$

Now apply Lemma 2 with the same $(\mathcal{P}, D, \gamma, S_\gamma)$ to obtain

$$\Pr_{(h,h')\sim\mathcal{P},\ T\sim D}\Big(M(h) = M(h')\ \wedge\ T \in S_\gamma(h, h')\Big) \ \le\ \frac{2\,\bar{\delta}_{\mathcal{P}}(\pi)}{c(\gamma)}.$$

□

J.2 PROOF OF COROLLARY 5

Proof. For convenience, write $\ell(h) := \ell_D^\gamma(h)$.

1. (ii) implies each M_j is a function of ℓ. Fix $j \in \{1, 2\}$. Assumption (ii) says $\ell(h) = \ell(h') \Rightarrow M_j(h) = M_j(h')$, so we may define a map a_j on the range of ℓ by $a_j(\ell(h)) := M_j(h)$; this is well-defined by (ii). Thus,

$$M_j(h) = a_j(\ell(h)), \tag{49}$$

almost surely under the history distribution.

2. (i) vanishing pair-regret + (iii) implies ℓ is a function of each M_j. Fix j. By (i) and Lemma 2 applied to π_j and M_j,

$$\begin{aligned}&\mathbb{E}_{(h,h')\sim\mathcal{P}}\Big[\mathbf{1}\{M_j(h) = M_j(h')\} \cdot D(S_\gamma(h, h'))\Big] \\ &\leq \frac{2\,\bar{\delta}_{\mathcal{P}}(\pi_j)}{c(\gamma)} \;\to\; 0.\end{aligned}$$

Since the integrand is nonnegative, this implies

$$\mathbf{1}\{M_j(h) = M_j(h')\} \cdot D(S_\gamma(h, h')) = 0, \tag{50}$$

for $\mathcal{P}$ almost everywhere on (h, h'). Now suppose (for $\mathcal{P}$ almost everywhere pairs) that $M_j(h) = M_j(h')$. Then (50) gives $D(S_\gamma(h, h')) = 0$. By γ-completeness (iii), $D(S_\gamma(h, h')) = 0$ implies $\ell(h) = \ell(h')$ (contrapositive). Therefore, $M_j(h) = M_j(h') \Rightarrow \ell(h) = \ell(h')$ for $\mathcal{P}$ almost everywhere on (h, h'), which means that ℓ is almost surely a function of M_j. Concretely, we can define b_j on the range of M_j by $b_j(M_j(h)) := \ell(h)$; this is well-defined almost surely because $M_j(h) = M_j(h')$ forces $\ell(h) = \ell(h')$. Hence

$$\ell(h) = b_j(M_j(h)), \tag{51}$$

almost surely.

3. Composition to obtain mutual recodings. Using (49) for $j = 1$ and (51) for $j = 2$,

$$M_1(h) = a_1(\ell(h)) = a_1(b_2(M_2(h))) \;:=\; \varphi(M_2(h)),$$

where $\varphi := a_1 \circ b_2$. Symmetrically,

$$M_2(h) = a_2(\ell(h)) = a_2(b_1(M_1(h))) \;:=\; \psi(M_1(h)),$$

where $\psi := a_2 \circ b_1$. This proves the claimed mutual recodability, almost surely on the support of $\mathcal{P}$. □