REINFORCING THE WORLD'S EDGE: A CONTINUAL LEARNING PROBLEM IN THE MULTI-AGENT-WORLD BOUNDARY

Anonymous authors
Paper under double-blind review

ABSTRACT

Ordinary decentralized multi-agent reinforcement learning presents each focal agent with a continual learning problem: peer updates change its induced rewards and dynamics even when the joint Markov game is stationary. We connect the lifetime of success-conditioned reusable structure to peer learning and policy reuse. An invariant core represents maximal abstract patterns shared by a high fraction of successful trajectories; survival refers to a fixed pattern's coverage, not an unchanged maximal frontier. An established sharp conditioning bound limits coverage loss to $\frac{\varepsilon}{p_0}$, where ε is trajectory-law drift, p_0 is reference success mass, and drift is smaller than the initial compatible-success mass. Combining this bound with peer-policy movement certifies $\Omega(\frac{1}{\eta})$ survival under bounded peer updates of size η and positive coverage margin. An effective-conflict condition yields a matching $\Theta(\frac{1}{\eta})$ first-exit law and holds in an analytic exact-policy-gradient class. With success-mass and performance calibration, structural survival also yields policy-value and finite-library transfer guarantees. An exactly solvable corridor tests the structural predictions. Two registered 64-stream studies in continual control and cue-MNIST show that coverage erosion predicts impending failure and enables near-oracle intervention. An exploratory reanalysis of eight learned-partner Level-Based Foraging pairings suggests the same erosion–failure link under peer learning.

1 INTRODUCTION

In ordinary Dec-MARL, the joint Markov game may remain stationary while the transitions and rewards experienced by one agent change as its peers update. This peer-induced non-stationarity is a central difficulty of multi-agent learning (Littman, 1994; Claus & Boutilier, 1998; Bowling & Veloso, 2002). Using the standard best-response lens (Shoham & Leyton-Brown, 2008; Busoniu et al., 2008), marginalizing the peers' policies in episode e gives a focal induced MDP $\mathcal{M}_e$ with the same focal trajectory law and expected return. Peer learning generates the sequence $(\mathcal{M}_0, \mathcal{M}_1, \ldots)$, an agent-centric continual reinforcement-learning problem (Khetarpal et al., 2022). The focal agent need not itself be learning: ordinary peer updates supply the endogenous change.

Which reusable structure survives that change? We study abstract patterns shared by successful focal trajectories. Their maximal representatives form an *invariant core*, whose coverage can erode as the trajectory law changes. Established conditioning results give a sharp amplification factor $\frac{1}{p_0}$ for reference success mass p_0 (Pelessoni & Vicig, 2022, Proposition 3.2). We combine this control with peer-learning dynamics to derive structural lifetimes and, with performance calibration, value guarantees.

For a fixed focal probe and candidate u, combining the induced-MDP representation with the conditioning theorem gives

$$c_{\mu_E}(u) \geq c_{\mu_0}(u) - \frac{\varepsilon_E^{\text{peer}}}{p_0}, \qquad \varepsilon_E^{\text{peer}} < p_0 c_{\mu_0}(u),$$
$$\varepsilon_E^{\text{peer}} := (H-1) \sum_{e=1}^{E} \sup_s \mathrm{TV}\big(\pi_2^e(\cdot \mid s), \pi_2^{e-1}(\cdot \mid s)\big)\,. \tag{$\star$}$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

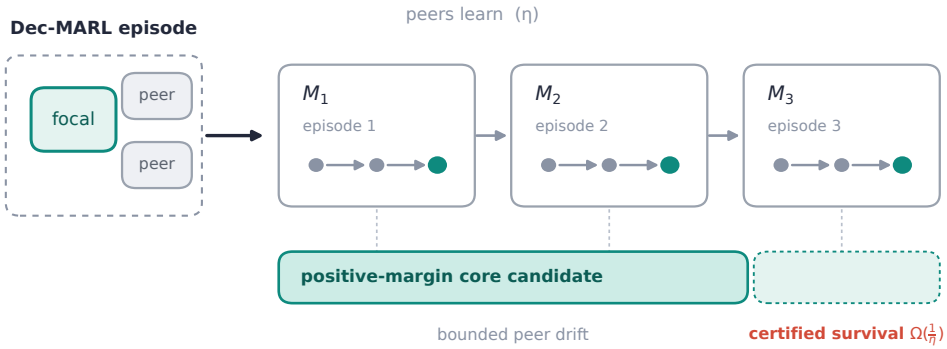

Figure 1: **Structural lifetime under peer learning.** Peer updates generate a continual sequence of induced MDPs for the focal agent. Sharp success-conditioning control converts their variation into an $\Omega(\frac{1}{\eta})$ survival certificate for a fixed candidate with positive coverage margin.

Here $c_{\mu_e}(u)$ is coverage among successes in episode e, $H-1$ bounds the number of transitions, and π_2^e is the policy in the one-peer specialization. The comparison fixes the initial law, probe, success event, and abstraction; the drift bound must be smaller than the reference mass of successful trajectories containing u (Lemma 3.2 and Theorem 3.1). If each update moves the peer policy by at most $c\eta$, positive initial coverage margin becomes an inverse-rate survival horizon (Corollary 3.1). The chain makes the moving agent–world boundary (Abel et al., 2025) quantitatively consequential.

Results.

1. **Structural lifetimes under peer learning.** Positive coverage margins give unconditional $\Omega(\frac{1}{\eta})$ survival; effective conflict gives a non-vacuous conditional $\Theta(\frac{1}{\eta})$ first exit, despite exact-core discontinuity (Proposition 3.2, Theorems 3.1 and 3.2, and Corollary 3.1).
2. **Performance consequences.** Calibrated success mass and executability convert surviving structure into policy-value, library-selection, and transfer-regret guarantees (Section 3.6).
3. **Empirical consequences.** Exact population tests examine the structural laws; registered continual-control and cue-MNIST studies test prediction and intervention; exploratory LBF connects erosion to learned-partner failure (Section 4).

Related work. Planning landmarks identify necessary facts and orderings (Hoffmann et al., 2004; Richter & Westphal, 2010); sequence mining supplies high-support candidates (Beedkar et al., 2019); and non-stationary-MDP analyses connect variation budgets to regret (Cheung et al., 2020; Mao et al., 2021). Our analysis combines established conditioning control with peer-induced variation to quantify structural lifetimes and performance consequences. Appendix B details the connections to conditioning, continual RL, abstraction, options, and partner-robust learning.

2 The Invariant Core in Stationary Single-Agent MDPs

The invariant core is defined first in a stationary MDP; Section 3 then introduces learning peers and the resulting environment drift. The two-route corridor of Figure 3 provides a running instance whose shared start-to-goal structure is captured by the core.

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, H, \mathcal{G})$ be a finite-horizon goal-conditioned MDP with finite $\mathcal{S}, \mathcal{A}$, start state $s_0 \notin \mathcal{G}$, and successful trajectories S terminating on first entry into $\mathcal{G}$. Here H is the maximum number of labels, including the terminal label, so there are at most $H-1$ stochastic transitions. Write $u \preccurlyeq v$ when string u is a not-necessarily-contiguous subsequence of v.

An optional abstraction $\phi : \mathcal{S} \times \mathcal{A} \to \Sigma$ maps each decision pair to an abstract symbol (Li et al., 2006). For a trajectory $\tau = (s_0, a_0, \ldots, s_T)$, the induced string contains $\phi(s_t, a_t)$ for $0 \le t < T$ and, on success, appends a distinguished terminal symbol g_Σ for the transition into $\mathcal{G}$. Hence every

string lies in $\Sigma^{\leq H}$ (the empty string included), and every $\tau \in \mathsf{S}$ satisfies $g_\Sigma \preccurlyeq \phi(\tau)$. The identity abstraction recovers the raw decision-pair alphabet $\mathcal{S} \times \mathcal{A}$ together with g_Σ.

Definition 2.1 (Coverage and the δ-core)**.** Let μ be a reference distribution over trajectories (e.g. induced by a fixed *probe* policy in $\mathcal{M}$), with $\mu(\mathsf{S}) > 0$. For $u \in \Sigma^{\leq H}$ define its *coverage*

$$c_\mu(u) \;=\; \Pr_{\tau\sim\mu}\big[\, u \preccurlyeq \phi(\tau) \;\mid\; \tau \in \mathsf{S} \,\big].$$

For $\delta \in [0, 1)$ the *high-coverage family* and the δ-*core* are

$$\mathcal{H}^\delta_\phi(\mu) \;=\; \big\{\, u \in \Sigma^{\leq H} \;:\; c_\mu(u) \geq 1 - \delta \,\big\}, \qquad \mathrm{Core}^\delta_\phi(\mu) \;=\; \max_{\preccurlyeq} \mathcal{H}^\delta_\phi(\mu),$$

the $\preccurlyeq$-maximal elements of the family. The *support core* is $\delta{=}0$. When μ has full support on S, $\mathcal{H}^0_\phi(\mu)$ is the universal family $\{u : \forall \tau \in \mathsf{S},\, u \preccurlyeq \phi(\tau)\}$ of subsequences shared by all successes, and $\mathrm{Core}^0_\phi(\mu)$ is its maximal frontier.

In words: coverage $c_\mu(u)$ is the fraction of *successful* trajectories in which the pattern u appears; the family $\mathcal{H}^\delta_\phi(\mu)$ collects the patterns present in all but a δ-fraction of successes; and the core is the shortlist of maximal such patterns, from which every family member is recovered as a subsequence.

Lemma 2.1 (The core generates its family)**.** *$\mathcal{H}^\delta_\phi(\mu)$ is hereditary: if $u \preccurlyeq w$ and $w \in \mathcal{H}^\delta_\phi(\mu)$ then $u \in \mathcal{H}^\delta_\phi(\mu)$. Consequently $\mathrm{Core}^\delta_\phi(\mu)$ is its finite generating frontier,*

$$\mathcal{H}^\delta_\phi(\mu) \;=\; \big\{\, u : \exists\, v \in \mathrm{Core}^\delta_\phi(\mu),\; u \preccurlyeq v \,\big\}.$$

Thus membership means $u \in \mathcal{H}^\delta_\phi(\mu)$, equivalently $u \preccurlyeq v$ for some frontier element. Proofs are in Appendices D and E.

The terminal symbol gives a nonempty core. For example, two positive-mass success strings (s_0, t, g_Σ) and (s_0, b, g_Σ) share frontier (s_0, g_Σ); a top-only law additionally generates t. Theorem C.1 and Propositions C.1 and C.2 develop bottlenecks, admissible pruning, and order constraints.

An ordered pattern requires options that can start, emit their symbols, chain, and reach the goal within the horizon. Assumption C.1 and Proposition C.3 give sufficient execution conditions and a Bellman decomposition for supplied options attaining the segment optima. Proposition 3.4 separates structural survival from execution.

2.1 Definitional invariance vs. estimated drift

In a stationary MDP, probes with full support on all feasible successes share the same support core. Thresholded coverage and finite empirical cores can change with the probe or its sampling coverage; Theorem C.2 controls fixed-probe i.i.d. evaluation batches, not pooled rollouts from an adapting policy. In Dec-MARL, by contrast, the true core itself can move.

Computation and measurement. Exact longest support-core computation is NP-hard (Maier, 1978). For a fixed candidate set $\mathcal{U}$, subsequence scans evaluate N trajectories in $O(N|\mathcal{U}|H)$ time; the experiments use bounded-length candidates. Candidates are specified before evaluation or discovered on separate data. Theorem C.2 gives their confidence bounds. If $\widehat{c}_0(u)$ estimates reference coverage with error at most t, then sampling and drift consume the same margin:

$$\widehat{c}_0(u) - t - \frac{\varepsilon_e}{p_0} \geq 1 - \bar{\delta} \quad\Longrightarrow\quad u \in \mathcal{H}^{\bar{\delta}}_\phi(\mu_e).$$

Here $p_0 = \mu_0(\mathsf{S}) > 0$ is known, $\bar{\delta} < 1$, and ε_e bounds trajectory TV. For unknown p_0, Proposition E.1 replaces it with a lower confidence bound and budgets both estimation errors jointly.

3 Focal Induced MDPs and Peer-Drift Budgets

Let $\mathfrak{G} = (\mathcal{S}, \mathcal{A}_1, \ldots, \mathcal{A}_n, P, R_1, H, \mathcal{G})$ be a stationary decentralized Markov game with focal agent 1 (Littman, 1994). In episode e, let the other agents follow a joint state-based Markov ker-

nel $\pi^e_{-1}(a_{-1} \mid s)$ that is fixed within the episode. Marginalizing them gives the induced MDP $\mathcal{M}_e = (\mathcal{S}, \mathcal{A}_1, P_e, R_e, H, \mathcal{G})$:

$$P_e(s' \mid s, a_1) = \sum_{a_{-1}} P(s' \mid s, a_1, a_{-1})\, \pi^e_{-1}(a_{-1} \mid s),$$
$$R_e(s, a_1) = \sum_{a_{-1}} R_1(s, a_1, a_{-1})\, \pi^e_{-1}(a_{-1} \mid s).$$

Proposition 3.1 (Focal induced-MDP equivalence)**.** *For every episode e and every focal policy based on its state–action history, the marginal focal state–action trajectory law in $\mathfrak{G}$ equals its law in $\mathcal{M}_e$, and expected additive return agrees. Hence every fixed success event, abstraction, coverage, and core transports exactly to $\mathcal{M}_e$. Put $D_e := \sup_s \mathrm{TV}(\pi^e_{-1}(\cdot \mid s), \pi^{e-1}_{-1}(\cdot \mid s))$ and $R_{\max} := \max_{s,a_1,a_{-1}} |R_1(s, a_1, a_{-1})|$. Then*

$$\|P_e - P_{e-1}\|_{1,\infty} \le 2D_e, \qquad \sup_{s,a_1} |R_e(s, a_1) - R_{e-1}(s, a_1)| \le 2R_{\max} D_e.$$

If decentralized peers act conditionally independently, then $D_e \le \sum_{j=2}^{n} \sup_s \mathrm{TV}(\pi^e_j(\cdot \mid s), \pi^{e-1}_j(\cdot \mid s))$.

Thus $\mathcal{C} = (\mathcal{M}_0, \mathcal{M}_1, \ldots)$ is the agent-centric continual RL process faced by the focal agent: the multi-agent task and joint game stay fixed, while peer learning generates drift in the focal agent's induced rewards and dynamics. Attributing this drift to peers requires a fixed focal probe. The notation henceforth specializes to one peer ($n = 2$), with policy π^e_2 and fixed goal-reaching event and probe law S, μ_e in $\mathcal{M}_e$. Persistent peer memory, shared randomness, or within-episode learning must be included in an augmented state; otherwise they need not induce a Markov kernel on s.

Scope of the reduction and lifetime bound. Every comparison fixes a common trajectory space, initial law, focal probe, success event S, abstraction ϕ, and candidate. The kernel-based results assume a fully observed Markov state and peers fixed within each episode. Augmenting latent peer memory gives a latent-state model, not necessarily an observable focal MDP. Applying these results to a history or belief state requires a common sufficient representation and a separately established kernel-drift bound on that representation. The law-level Lemma 3.2 applies directly whenever a common-space trajectory-TV bound is available. Reward-only drift enters the value analysis through $G_{e,k}$ and its calibration radius.

Dependency chain. For a fixed candidate, the argument follows the chain

$$\text{peer movement} \Longrightarrow \text{induced-kernel drift} \Longrightarrow \text{trajectory-law drift}$$
$$\Longrightarrow \text{conditional coverage loss} \Longrightarrow \text{survival time} \Longrightarrow \text{policy value}.$$

Proposition 3.1 establishes the first link, Lemma 3.1 the second, and Lemma 3.2 the third. Positive margin yields survival through Theorem 3.1 and Corollary 3.1; effective peer progress yields the two-sided law in Theorem 3.2; and success mass plus performance calibration yield value in Theorem 3.3. Each conclusion retains the premises of the preceding link.

3.1 Impossibility: the exact core is discontinuous

Proposition 3.2 (No modulus of continuity for exact membership)**.** *There is a family of games and peer policies $\{\pi^\alpha_2\}_{\alpha \in [0, \frac{1}{2}]}$ with induced kernels P_α such that $\|P_\alpha - P_0\|_{1,\infty} \le 2\alpha \to 0$, yet a prototype $\sigma^\star$ generated by the exact core ($\sigma^\star \in \mathcal{H}^0_\phi(\mu_0)$) satisfies $\sigma^\star \notin \mathcal{H}^0_\phi(\mu_\alpha)$ for* every *$\alpha > 0$. Hence the map $P \mapsto \mathbf{1}[c_{\mu_P}(u) = 1]$ from the induced kernel to the exact-membership indicator admits no modulus of continuity, and the maximal frontier itself jumps.*

Two routes sharing only start and goal witness the result: any positive probability on the bypass removes a top-route symbol from coverage one, although the induced kernel converges to the top-only kernel.

3.2 A COVERAGE MARGIN MAKES δ-CORE MEMBERSHIP LIPSCHITZ-STABLE

Lemma 3.1 (Trajectory simulation). *For a fixed focal probe policy and common initial law, the trajectory laws under $\mathcal{M}_e$ and a reference $\mathcal{M}_0$ satisfy $\|\mu_e - \mu_0\|_{\mathrm{TV}} \le \varepsilon_e := \frac{1}{2}(H-1)\ \|P_e - P_0\|_{1,\infty}$.*

Lemma 3.2 (Sharp success-conditioning step). *Let μ, ν be trajectory laws, let $A \subseteq \mathsf{S}$, and put $p = \mu(\mathsf{S}) > 0$. If $\mathrm{TV}(\mu, \nu) \le \varepsilon < \mu(A)$, then $\nu(\mathsf{S}) > 0$ and*

$$\nu(A \mid \mathsf{S}) \ge \mu(A \mid \mathsf{S}) - \frac{\varepsilon}{p}.$$

The coefficient $\frac{1}{p}$ is best possible.

Key proof step. Put $a = \mu(A)$ and $b = \mu(\mathsf{S} \setminus A)$, so $a + b = p$. Total variation gives $\nu(A) \ge a - \varepsilon > 0$ and $\nu(\mathsf{S} \setminus A) \le b + \varepsilon$. Since $\frac{x}{x+y}$ increases in x and decreases in y,

$$\nu(A \mid \mathsf{S}) \ge \frac{a - \varepsilon}{(a - \varepsilon) + (b + \varepsilon)} = \frac{a}{p} - \frac{\varepsilon}{p}.$$

When $\mathsf{S} \setminus A \neq \varnothing$, moving ε mass there from A attains equality.

Theorem 3.1 (δ-family stability). *Let $p_0 = \mu_0(\mathsf{S}) > 0$, $\varepsilon_e = \frac{1}{2}(H-1)\ \|P_e - P_0\|_{1,\infty}$, and suppose $\varepsilon_e < (1-\delta)p_0$. If $u \in \mathcal{H}_\phi^\delta(\mu_0)$ (i.e. u is generated by the δ-core of μ_0) then $u \in \mathcal{H}_\phi^{\delta'}(\mu_e)$ with*

$$\delta' \ \le\ \delta + \frac{\varepsilon_e}{p_0} \ <\ 1.$$

Equivalently, $\mathcal{H}_\phi^\delta(\mu_0) \subseteq \mathcal{H}_\phi^{\delta + \frac{\varepsilon_e}{p_0}}(\mu_e)$. The trajectory-law constant is worst-case tight.

The proof follows Lemma 3.1 $\Longrightarrow$ Lemma 3.2 $\Longrightarrow$ Theorem 3.1. The middle lemma specializes established total-variation conditioning (Pelessoni & Vicig, 2022, Proposition 3.2); Appendix D gives the correspondence and a self-contained proof. The theorem controls coverage; membership in a fixed target family requires positive initial margin, and the maximal frontier may change.

Fix reference slack $\delta_{\mathrm{ref}} \in [0, 1)$ and target slack $\bar{\delta} \in [0, 1)$. Define the surviving reference family and the target-family exit time by

$$\mathcal{K}_e^{\delta_{\mathrm{ref}} \to \bar{\delta}} := \mathcal{H}_\phi^{\delta_{\mathrm{ref}}}(\mu_0) \cap \mathcal{H}_\phi^{\bar{\delta}}(\mu_e), \qquad E_{\mathrm{exit}}^{\bar{\delta}}(u) := \inf\{e \ge 0 : u \notin \mathcal{H}_\phi^{\bar{\delta}}(\mu_e)\},$$

with $\inf \varnothing = \infty$.

Proposition 3.3 (Ordered erosion: the surviving structure is again a core). *At each episode $\mathcal{K}_e^{\delta_{\mathrm{ref}} \to \bar{\delta}}$ is hereditary and therefore has a finite maximal frontier. If $u \preccurlyeq v$ and both lie in the reference family, then $E_{\mathrm{exit}}^{\bar{\delta}}(u) \ge E_{\mathrm{exit}}^{\bar{\delta}}(v)$. Re-entry can occur under non-monotone peer updates.*

3.3 FROM PEER LEARNING RATE TO BOUNDARY DRIFT

The inequalities $\|P_e - P_0\|_{1,\infty} \le \sum_{j \le e} \|P_j - P_{j-1}\|_{1,\infty} =: V_e$ (a variation budget as in Even-Dar et al., 2009; Cheung et al., 2020; Mao et al., 2021) and $\|P_e - P_{e-1}\|_{1,\infty} \le 2 \sup_s \mathrm{TV}(\pi_2^e(\cdot \mid s), \pi_2^{e-1}(\cdot \mid s))$ link the peer's learning dynamics directly to induced-kernel drift.

Corollary 3.1 (Certified survival horizon $\Omega(\frac{1}{\eta})$). *Let $H \ge 2$, $\delta_0 := 1 - c_{\mu_0}(u) < \bar{\delta}$, and $m := \bar{\delta} - \delta_0 > 0$. If $\sup_s \mathrm{TV}(\pi_2^e, \pi_2^{e-1}) \le c\eta$ for every $e \ge 1$, then $V_E \le 2c\eta E$ and $u \in \mathcal{H}_\phi^{\bar{\delta}}(\mu_E)$ for every*

$$E \ \le\ E^\star = \left\lfloor \frac{m\, p_0}{c\, \eta\, (H-1)} \right\rfloor = \Omega\left(\frac{1}{\eta}\right),$$

with c, p_0, m, H fixed as η varies.

Remark 3.1 (One-sided guarantee). Membership holds through $E^\star$, hence $E_{\mathrm{exit}}^{\bar{\delta}}(u) \ge E^\star + 1$. A zero-gradient peer may preserve u forever, so this unconditional result is $\Omega(\frac{1}{\eta})$, not $\Theta(\frac{1}{\eta})$.

Remark 3.2 (Score-to-TV control). Suppose π_θ is differentiable, $\sup_{s,a,\vartheta} \|\nabla_\theta \log \pi_\vartheta(a \mid s)\|_2 \le B$ along the update segment, and $\|d_e\|_2 \le G$. Integrating the policy derivative along $\theta_e + t\eta d_e$ gives, for every state,

$$\mathrm{TV}(\pi_{\theta_{e+1}}, \pi_{\theta_e}) \le \frac{\eta}{2} \int_0^1 \sum_a \pi_{\theta_e + t\eta d_e}(a \mid s) \left| \langle \nabla \log \pi_{\theta_e + t\eta d_e}(a \mid s), d_e \rangle \right| dt \le \frac{BG}{2} \eta.$$

Thus the corollary holds with $c = \frac{BG}{2}$. A softmax policy with feature norm at most F has score norm at most $2F$, giving $c = FG$.

3.4 A TWO-SIDED LIFETIME LAW UNDER PEER PROGRESS

Corollary 3.1 is one-sided because its premise $\sup_s \mathrm{TV}(\pi_2^e, \pi_2^{e-1}) \le c\eta$ is an *upper* bound: a peer that stops moving satisfies it without eroding coverage. A matching upper bound on the lifetime therefore requires the peer to keep moving *and* to move in a direction that actually erodes the pattern. Let $A = \{u \preccurlyeq \phi(\tau)\} \cap \mathsf{S}$ and write

$$c(\theta) \;=\; \frac{\mu_\theta(A)}{\mu_\theta(\mathsf{S})}$$

for the coverage of u as a function of the peer parameter θ; on a finite MDP with a smooth (e.g. softmax) parameterization, c is smooth wherever $\mu_\theta(\mathsf{S}) > 0$.

Assumption 3.1 (Effective peer progress)**.** Write $E_{\mathrm{exit}} := E^{\bar\delta}_{\mathrm{exit}}(u)$. For updates $\theta_{e+1} = \theta_e + \eta d_e$, assume before first exit that

(P1) **Bounded step.** $\|d_e\|_2 \le G$;

(P2) **Effective conflict.** For every $1 \le E < E_{\mathrm{exit}}$, $\frac{1}{E}\sum_{e<E} -\langle \nabla c(\theta_e), d_e \rangle \ge \lambda > 0$;

(P3) **Smoothness.** $\nabla_\theta c$ is L_c-Lipschitz on the segment $[\theta_e, \theta_{e+1}]$.

Proposition C.5 gives an analytic two-route policy-gradient class with $\lambda = \Delta r_\star^2 > 0$, verifying non-vacuity.

Lemma 3.3 (Cumulative erosion)**.** *Under Assumption 3.1, if* $\eta \le \frac{\lambda}{L_c G^2}$ *then for every integer* $1 \le E < E_{\mathrm{exit}}$

$$c(\theta_E) \;\le\; c(\theta_0) - \kappa\,\eta E, \qquad \kappa := \frac{\lambda}{2}.$$

Theorem 3.2 (Two-sided lifetime law)**.** *Fix a target family* $\mathcal{H}^{\bar\delta}_\phi$*, let* $\delta_0 := 1 - c_{\mu_0}(u)$ *be the* exact *initial slack of* u*, and let* $m := \bar\delta - \delta_0 > 0$*. Under the premise of Corollary 3.1 together with Assumption 3.1 for the same target* $\bar\delta$*, and for* $\eta \le \frac{\lambda}{L_c G^2}$*, the first exit episode obeys*

$$\left\lfloor \frac{m\,p_0}{c\,\eta\,(H-1)} \right\rfloor + 1 \;\le\; E_{\mathrm{exit}} \;\le\; \left\lfloor \frac{2m}{\lambda\,\eta} \right\rfloor + 1.$$

Hence $E_{\mathrm{exit}} = \Theta(\frac{1}{\eta})$ *when all displayed constants are uniform in* η*.*

Remark 3.3 (Effective conflict). (P2) requires persistent coverage-eroding progress; the bounded-step survival certificate also permits cycling or stalled peers. For stochastic updates, the two-sided conclusion holds on realized paths satisfying (P2); a mean-gradient condition alone is insufficient. Proposition C.5 verifies (P2) pointwise for exact policy gradient.

From cumulative erosion to a windowed monitor. Writing $a_j = -\langle \nabla c(\theta_j), d_j \rangle$, smoothness and bounded steps give

$$\left| c(\theta_{t-w}) - c(\theta_t) - \eta \sum_{j=t-w}^{t-1} a_j \right| \le \frac{L_c G^2}{2} \eta^2 w.$$

Proposition E.2 carries this bound through the positive part and confidence intervals. It yields a windowed erosion signal under progress on that window; predicting a future exit additionally requires continued progress.

3.5 MEMBERSHIP DRIFT VS. EXECUTABILITY DRIFT

Proposition 3.4 (Two drift modes)**.** *A peer update can cause two distinct failures. Under* membership drift*, u leaves $\mathcal{H}^0_\phi(\mu_{e+1})$. Under* executability drift*, u remains in the family, but its options violate chainability or bounded-horizon goal termination (Assumption C.1). Membership monitoring cannot detect the second failure: the ordered chain's success probability can become arbitrarily small while $\mathcal{H}^0_\phi$ and its frontier remain unchanged.*

3.6 FROM CORE SURVIVAL TO POLICY VALUE

Membership alone cannot provide a value bound because execution can still fail (Proposition 3.4). Value control therefore also requires executability of the preserved pattern. Let $\{\pi^k\}_{k=1}^K$ be a supplied library of focal policies or option chains, with fixed candidates u_k. At episode e, let $\mu_{e,k}$ be the monitoring law and $\nu_{e,k}$ the deployment law for π^k on a common padded trajectory space. Put

$$A_k := \mathsf{S} \cap \{u_k \preccurlyeq \phi(\tau)\}, \qquad \Gamma_{e,k} := \mu_{e,k}(A_k) = p_{e,k} c_{\mu_{e,k}}(u_k),$$

where $p_{e,k} := \mu_{e,k}(\mathsf{S}) > 0$. Thus $\Gamma_{e,k}$ is the *core-compatible success mass*, not success-conditional coverage alone. Let $G_{e,k}(\tau)$ be total return and $V_{e,k} := \mathbb{E}_{\nu_{e,k}} G_{e,k}$.

The value bridge has three explicit links. First, structural survival supplies $c_{\mu_{e,k}}(u_k) \geq 1 - \delta$. Second, a success-mass floor supplies $\Gamma_{e,k} = p_{e,k} c_{\mu_{e,k}}(u_k) \geq \underline{p}_k(1 - \delta)$. Third, performance calibration converts this compatible mass into deployment value. The next assumption states exactly the quantities required for the third link.

Assumption 3.2 (Performance calibration)**.** For each (e, k) there are upper bounds $\chi_{e,k}, \xi_{e,k}, b_{e,k} \geq 0$ and a scale $r_k \geq 0$ such that

$$|\nu_{e,k}(A_k) - \mu_{e,k}(A_k)| \leq \chi_{e,k}, \qquad \mathbb{E}_{\nu_{e,k}}[|G_{e,k}|\mathbf{1}[A_k^c]] \leq b_{e,k},$$
$$|G_{e,k} - r_k|\mathbf{1}[A_k] \leq \xi_{e,k}\mathbf{1}[A_k] \quad \nu_{e,k}\text{-a.s.}$$

Write $\zeta_{e,k} := \xi_{e,k} + b_{e,k} + r_k \chi_{e,k}$. The radius $\zeta_{e,k}$ records execution error, reward outside the compatible event, and shift in that event's probability.

The event-specific shift is sufficient; full trajectory TV also upper-bounds it. On-policy monitoring has $\chi_{e,k} = 0$. Independent monitoring and deployment samples can bound this scalar event difference (Appendix E).

Theorem 3.3 (Core-compatible value and transfer)**.** *Under Assumption 3.2,*

$$|V_{e,k} - r_k \Gamma_{e,k}| \leq \zeta_{e,k}.$$

Hence if $u_k \in \mathcal{H}^\delta_\phi(\mu_{e,k})$ and $p_{e,k} \geq \underline{p}_k > 0$, then

$$V_{e,k} \geq r_k \underline{p}_k (1 - \delta) - \zeta_{e,k}.$$

For selection, suppose the scales and valid calibration bounds are available before choosing a policy. For a finite episode set $\mathcal{E}$, let $\mathcal{I} := \mathcal{E} \times \{1, \ldots, K\}$. Estimate $\Gamma_{e,k}$ from $N_{e,k}$ independent monitoring rollouts by the empirical mean $\widehat{\Gamma}_{e,k}$ of $\mathbf{1}[A_k]$, and set

$$\beta_{e,k} := \sqrt{\frac{\log(\frac{2|\mathcal{I}|}{\rho})}{2N_{e,k}}}, \qquad L_{e,k} := r_k \widehat{\Gamma}_{e,k} - r_k \beta_{e,k} - \zeta_{e,k}.$$

With probability at least $1 - \rho$, the selector $\widehat{k}_e \in \arg\max_k L_{e,k}$ satisfies simultaneously for every $e \in \mathcal{E}$ and comparator k,

$$V_{e,k} - V_{e,\widehat{k}_e} \leq 2(r_k \beta_{e,k} + \zeta_{e,k}).$$

Consequently, taking $k_e^\star \in \arg\max_k V_{e,k}$ *and summing,*

$$\sum_{e\in\mathcal{E}}\left(V_{e,k_e^\star} - V_{e,\widehat{k}_e}\right) \le 2\sum_{e\in\mathcal{E}}\left(r_{k_e^\star}\beta_{e,k_e^\star} + \zeta_{e,k_e^\star}\right).$$

Corollary 3.2 (Value survives with the core)**.** *Suppose the monitoring law for* (π^k, u_k) *is the fixed-probe law of Corollary 3.1,* $p_{e,k} \ge \underline{p}_k$ *and* $\zeta_{e,k} \le \bar{\zeta}_k$ *through its certified horizon* $E^\star$*. Then, for every* $e \le E^\star$*,*

$$V_{e,k} \ge r_k \underline{p}_k(1-\bar{\delta}) - \bar{\zeta}_k.$$

Within this horizon, the same drift bound supplies $p_{e,k} \ge p_{0,k} - \varepsilon_{e,k} \ge p_{0,k}(1-m) > 0$*, so one may take* $\underline{p}_k = p_{0,k}(1-m)$*. Execution and reward calibration then give an* $\Omega(\frac{1}{\eta})$ *policy-value floor.*

A supplied policy or option library can be evaluated by monitoring core-compatible success mass together with its execution radius. Theorem 3.3 converts these quantities into selection and transfer-regret guarantees, while Proposition 3.4 requires monitoring both. The partner-universal extension in Definition C.1 and Proposition C.4 characterizes structure shared across success-enabling peers.

4 Empirical Validation

The experiments serve three roles. E1–E9 test the structural laws in the two-player Switch–Rock Corridor (SRC), including discontinuity, stability, inverse-rate lifetime, sharpness, estimation, and executability. Exact population coverage gives first-exit time linear in $\frac{1}{\eta}$ ($R^2 > 0.9999$), and all 50 finite-rollout grid exits match their exact-grid counterparts. E10 and E12 prospectively test whether coverage erosion predicts failure and supports intervention in continual control and cue-MNIST. E11 explores the complete peer-update–coverage–failure chain in LBF. In these experiments, *core erosion* means a decline in the coverage, or its lower confidence bound, of a fixed core candidate rather than a distance between two maximal frontiers.

Average precision (AP) summarizes the precision–recall ranking of prospective failure labels; higher AP means better separation of impending failures from eligible non-failures. Control and cue-MNIST are frozen-design confirmations: their protocols and analyses were fixed on development seeds before evaluation on disjoint confirmation seeds.

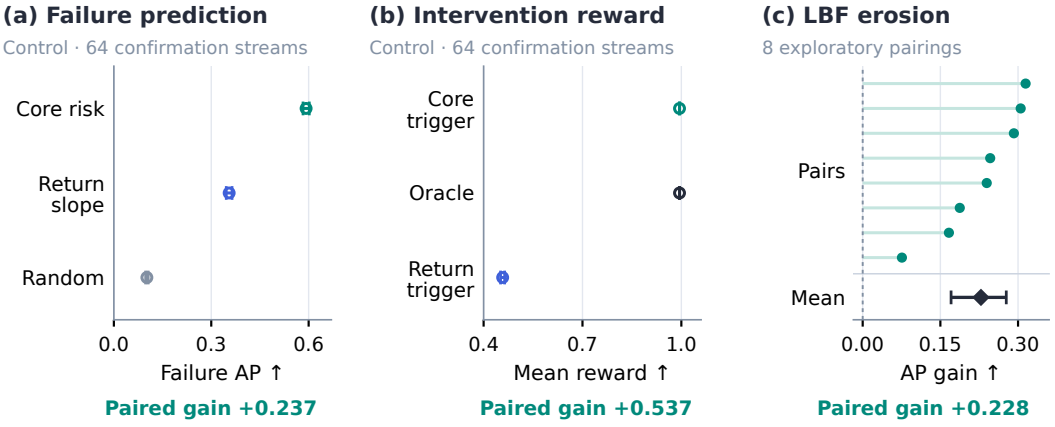

Figure 2: **Core erosion predicts failure and supports intervention.** *(a)–(b), E10 confirmation:* Mean outcomes with 95% intervals, some narrower than markers. *(c), E11 exploratory reanalysis:* Each unlabelled row is one development pairing's AP gain over return slope, ordered by gain; the diamond shows the mean with its 95% interval. Paired core-over-return gains appear below each panel; their 95% intervals are (a) $[0.223, 0.251]$, (b) $[0.531, 0.543]$, and (c) $[0.171, 0.277]$. Intervals resample complete streams or pairings. Panels (a) and (c) report AP gains; (b) reports mean-reward gain. Figures 17 and 18 give extended results and cue-MNIST confirmation.

Tasks and measured structure. In SRC, the peer changes route availability and coverage measures a fixed pattern among goal-reaching trajectories. In continual control, actuator drift changes the fraction of reference states that stay in a safe interval under both extreme disturbances; this is a length-one coverage monitor with reference success mass one. In cue-MNIST, a 64-unit tanh classifier is probed for correctness under all ten cue identities among reference-cue-correct examples. In LBF, partner learning changes successful left-route coordination under a fixed focal probe. Each score monitors a fixed event or pattern; a decline in coverage or its lower confidence bound supplies the erosion signal. E10 has $p_t = 1$; E12 conditions on reference-correct examples and has $p_t = \frac{|B_t|}{2000}$. Task schematics appear in Figures 3 and 19, and Appendix F.1 gives the exact event-space mappings.

Intervals resample complete independent streams or peer/focal pairings, preserving dependent checkpoints within each unit. Prediction and intervention are compared within the same stream or pairing; the reported differences therefore use paired confidence intervals.

The registered studies compare specified monitoring signals and intervention triggers within each task. In continual control, 64 confirmation streams give core-risk AP 0.592 versus 0.355 for return slope. Core-triggered reward is 0.9944 versus 0.4578 for the return trigger, 0.4604 without intervention, and 0.5356 for random-matched switching, close to the 0.9945 oracle (Figure 2).

In continual cue-MNIST, 64 untouched supervised streams give erosion AP 0.384 versus 0.345 for cue association (paired $+0.039$, 95% CI $[0.022, 0.055]$). Core-triggered replay obtains 0.9500 reward versus 0.9382 for an accuracy trigger (paired $+0.0117$, $[0.0113, 0.0121]$), matching the 0.9502 oracle (Figure 18).

In a controlled Level-Based Foraging coordination fork, peers use a feed-forward logistic route policy updated by sampled REINFORCE. An exploratory reanalysis of eight development pairings suggests that fixed-probe, three-checkpoint core erosion predicts subsequent old-partner failure, reaching AP 0.971 versus 0.743 for return slope (paired difference 0.228, 95% CI $[0.171, 0.277]$; Figure 2). The registered absolute-risk score attained AP 0.526, below return slope; the post-hoc switch to erosion used a window matching the already-frozen three-checkpoint failure horizon. Evaluation on the same development pairings remains exploratory. The coordination-fork schematic is shown at right in Figure 19.

5 Conclusion

Ordinary Dec-MARL presents each focal agent with a continual RL problem as its peers update, even though the joint Markov game remains stationary. Combining established sharp conditioning bounds with peer-induced trajectory drift gives a structural-lifetime analysis: positive coverage margins certify unconditional $\Omega(\frac{1}{\eta})$ survival, effective conflict yields a non-vacuous conditional $\Theta(\frac{1}{\eta})$ first-exit law, and calibrated executability gives policy-value and transfer guarantees. This chain connects peer learning to the persistence and usefulness of reusable structure. Registered E10 and E12 show that core erosion predicts failure and enables near-oracle intervention; exploratory E11 suggests that erosion also anticipates peer-induced old-partner failure.

Reproducibility statement. Complete proofs, protocols, per-seed statistics, and fixed-seed figure-regeneration code are supplied in Appendices D and F.

REFERENCES

David Abel, David Hershkowitz, and Michael Littman. Near optimal behavior via approximate state abstraction. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2915–2923, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/abel16.html.

David Abel, André Barreto, Michael Bowling, Will Dabney, Shi Dong, Steven Hansen, Anna Harutyunyan, Khimya Khetarpal, Clare Lyle, Razvan Pascanu, et al. Agency is frame-dependent. *arXiv preprint arXiv:2502.04403*, 2025.

Ryan Prescott Adams and David J. C. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Kaustubh Beedkar, Rainer Gemulla, and Wim Martens. A unified framework for frequent sequence mining with subsequence constraints. *ACM Transactions on Database Systems*, 44(3):11:1–11:42, 2019. doi: 10.1145/3321486.

Michael Bowling and Manuela Veloso. Multiagent learning using a variable learning rate. *Artificial intelligence*, 136(2):215–250, 2002.

Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 38(2): 156–172, 2008.

David H. Chan, Mark Roberts, and Dana S. Nau. Landmark-assisted Monte Carlo planning. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI)*, 2025. doi: 10.48550/arXiv.2508.11493. arXiv:2508.11493.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Reinforcement learning for non-stationary Markov decision processes: The blessing of (More) optimism. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1843–1854. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/cheung20a.html.

Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minigrid: A minimalistic gridworld environment for openai gym. https://github.com/Farama-Foundation/Minigrid, 2018.

Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. *Proceedings of the National Conference on Artificial Intelligence*, 15(2):746–752, 1998.

Thomas Dean and Robert Givan. Model minimization in markov decision processes. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, AAAI'97/IAAI'97, pp. 106–111. AAAI Press, 1997. ISBN 0262510952.

Esraa Elelimy, David Szepesvari, Martha White, and Michael Bowling. Rethinking the foundations for continual reinforcement learning. *Reinforcement Learning Journal*, 6:2174–2194, 2025.

Eyal Even-Dar, Sham Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Jakob Foerster, Nantas Nardelli, Gregory Farquhar, Triantafyllos Afouras, Philip HS Torr, Pushmeet Kohli, and Shimon Whiteson. Stabilising experience replay for deep multi-agent reinforcement learning. In *International conference on machine learning*, pp. 1146–1155. PMLR, 2017.

Jakob Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 122–130, 2018.

Sertan Girgin and Faruk Polat. Option discovery in reinforcement learning using frequent common subsequences of actions. In *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC)*, volume 1, pp. 371–376. IEEE, 2005.

He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. Opponent modeling in deep reinforcement learning. In *International conference on machine learning*, pp. 1804–1813. PMLR, 2016.

Pablo Hernandez-Leal, Michael Kaisers, Tim Baarslag, and Enrique Munoz de Cote. A survey of learning in multiagent environments: Dealing with non-stationarity. *arXiv preprint arXiv:1707.09183*, 2017.

Jörg Hoffmann, Julie Porteous, and Laura Sebastia. Ordered landmarks in planning. *Journal of Artificial Intelligence Research*, 22:215–278, 2004.

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "Other-play" for zero-shot coordination. In *International Conference on Machine Learning*, pp. 4399–4410. PMLR, 2020.

Chia-Chun Hung, Timothy Lillicrap, Josh Abramson, Yori Wu, Adam Marblestone, and Greg Wayne. Optimizing agent behavior over long time scales by transporting value. *Nature Communications*, 10(1):1362, 2019.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2022. doi: 10.1613/JAIR.1.13673. URL https://doi.org/10.1613/JAIR.1.13673.

George Konidaris and Andrew Barto. Skill discovery in continuous reinforcement learning domains using skill chaining. *Advances in neural information processing systems*, 22, 2009.

Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for mdps. *AI&M*, 1(2):3, 2006.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.

Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. Trajectory diversity for zero-shot coordination. In *International Conference on Machine Learning*, pp. 7204–7213. PMLR, 2021.

David Maier. The complexity of some problems on subsequences and supersequences. *Journal of the ACM*, 25(2):322–336, 1978. doi: 10.1145/322063.322075.

Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, and Tamer Basar. Near-optimal model-free reinforcement learning in non-stationary episodic mdps. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7447–7458. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/mao21b.html.

Amy McGovern and Andrew G Barto. Automatic discovery of subgoals in reinforcement learning using diverse density. In *International Conference on Machine Learning*, pp. 361–368, 2001.

Georgios Papoudakis, Filippos Christianos, Arrasy Rahman, and Stefano V Albrecht. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737*, 2019.

Renato Pelessoni and Paolo Vicig. Dilation properties of coherent Nearly-Linear models. *International Journal of Approximate Reasoning*, 140:211–231, 2022. doi: 10.1016/j.ijar.2021.10.009. URL https://doi.org/10.1016/j.ijar.2021.10.009.

Roberta Raileanu, Emily Denton, Arthur Szlam, and Rob Fergus. Modeling others using oneself in multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Silvia Richter and Matthias Westphal. The LAMA planner: Guiding cost-based anytime planning with landmarks. *Journal of Artificial Intelligence Research*, 39:127–177, 2010.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/fa7cdfad1a5aaf8370ebeda47a1ff1c3-Abstract.html.

Yoav Shoham and Kevin Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2008.

Özgür Şimşek and Andrew Barto. Skill characterization based on betweenness. *Advances in Neural Information Processing Systems*, 21, 2009.

Peter Stone, Gal A Kaminka, Sarit Kraus, and Jeffrey S Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1):1504–1509, 2010.

DJ Strouse, Kevin R. McKee, Matt Botvinick, Edward Hughes, and Richard Everett. Collaborating with humans without human data. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2110.08176.

Chen Sun, Wannan Yang, Thomas Jiralerspong, Dane Malenfant, Benjamin Alsbury-Nealy, Yoshua Bengio, and Blake Aaron Richards. Contrastive retrospection: Honing in on critical steps for rapid learning and generalization in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

A NOTATION

Task and trajectories	
$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, H, \mathcal{G})$	finite-horizon goal-conditioned MDP: states, actions, kernel, reward, horizon, goal set; fixed start s_0
$\mathcal{G} \subseteq \mathcal{S}$	goal set; episodes terminate on first visit
H	label horizon: at most H abstract labels and $H-1$ stochastic transitions
$\tau = (s_0, a_0, \ldots, s_T)$	a raw trajectory with $T \le H-1$ transitions; S is the set of trajectories whose terminal state lies in $\mathcal{G}$
$\phi : \mathcal{S} \times \mathcal{A} \to \Sigma$	optional state–action abstraction; $\phi(\tau)$ is the induced string over Σ
$g_\Sigma \in \Sigma$	distinguished terminal symbol appended on first entry into $\mathcal{G}$, so $\lvert\phi(\tau)\rvert \le H$
$u \preccurlyeq v$	u is a (not necessarily contiguous) subsequence of v
μ has full conditional support	every feasible successful trajectory has positive μ-mass
Coverage, core, and family (Definition 2.1, Lemma 2.1)	
μ	reference (probe-induced) trajectory law; $p_0 = \mu_0(\mathsf{S}) > 0$
$c_\mu(u)$	*coverage*: $\Pr_{\tau\sim\mu}[u \preccurlyeq \phi(\tau) \mid \tau \in \mathsf{S}]$
$\delta \in [0,1)$	coverage slack: $u \in \mathcal{H}^\delta$ iff $c_\mu(u) \ge 1-\delta$. *Margin* is reserved for $m = \bar{\delta} - \delta_0$, the slack to a fixed target family (Corollary 3.1)
$\mathcal{H}^\delta_\phi(\mu)$	*high-coverage family*: $\{u : c_\mu(u) \ge 1-\delta\}$ (hereditary under $\preccurlyeq$)
$\mathrm{Core}^\delta_\phi(\mu)$	*δ-core*: the $\preccurlyeq$-maximal *frontier* $\max_\preccurlyeq \mathcal{H}^\delta_\phi(\mu)$, which generates the family
"generated by the core"	$u \in \mathcal{H}^\delta_\phi(\mu)$, i.e. $u \preccurlyeq v$ for some frontier $v \in \mathrm{Core}^\delta_\phi(\mu)$
ϕ-bottleneck	symbol σ with $c_\mu(\sigma) = 1$; equivalently a step-cut of G_S (Theorem C.1)
G_S	action-labelled time-unrolled success graph: vertices (s,t) and edges $(s,t) \xrightarrow{a} (s', t+1)$ from μ-positive successes; each edge carries $\phi(s,a)$
o_σ	time-aware option $(I_\sigma, \pi_\sigma, \beta_\sigma)$ realizing σ: initiation set, policy, termination (Assumption C.1)
$V^\star$, $\mathcal{P}_i(t)$	optimal almost-sure goal-reaching cost and admissible policies for landmark segment i (Proposition C.3)
Estimation (Appendix C.3.1)	
$\widehat{c}_N(u)$	empirical coverage from N probe rollouts; N_S successful among them
t, γ, ρ	estimation tolerance, gap to the threshold $1-\delta$, and failure probability
Decentralized game and drift (Section 3)	
π^e_{-1}, π^e_2	joint peer policy and its one-peer specialization at episode e; state-based and fixed within the episode
$\mathfrak{G}$	stationary n-player Markov game; distinct from the goal set $\mathcal{G}$
$\mathcal{M}_e$, P_e, μ_e	focal-induced MDP, kernel, and probe law at episode e
$\mathcal{C} = (\mathcal{M}_e)_{e\ge 0}$	focal continual RL process induced by the peers (Proposition 3.1)
$\lVert P - Q\rVert_{1,\infty}$	rowwise difference $\sup_{s,a_1} \sum_{s'} \lvert P(s' \mid s,a_1) - Q(s' \mid s,a_1)\rvert$
$\mathrm{TV}(\cdot,\cdot)$	total variation distance
ε_e	trajectory-law drift bound $\frac{1}{2}(H-1)\lVert P_e - P_0\rVert_{1,\infty}$ (Lemma 3.1)
V_e	variation budget $\sum_{j=1}^e \lVert P_j - P_{j-1}\rVert_{1,\infty}$
δ'	drifted slack; $\delta' \le \delta + \frac{\varepsilon_e}{p_0}$ (Theorem 3.1)
η, c	peer step size and score-to-TV constant (Remark 3.2)
$\mathcal{K}_e^{\delta_{\mathrm{ref}}\to\bar{\delta}}$	reference patterns from $\mathcal{H}^{\delta_{\mathrm{ref}}}_\phi(\mu_0)$ that remain in the target family $\mathcal{H}^{\bar{\delta}}_\phi(\mu_e)$
$E^\star$, $E^\delta_{\mathrm{exit}}(u)$	certified survival horizon and first episode outside the target family (Proposition 3.3 and Corollary 3.1); membership holds through $E^\star$, so exit is at least $E^\star+1$
Π_2^+	peer policies with positive probe success, $\{\pi_2 : \mu(\pi_2)(\mathsf{S}) > 0\}$
$\mathrm{Supp}_\mathsf{S}(\pi_2)$	positive-mass successful trajectories under peer policy π_2
$\mathcal{H}_{\mathrm{rob}}$, $\mathrm{Core}_{\mathrm{rob}}$	robust (partner-universal) family $\bigcap_{\pi_2\in\Pi_2^+} \mathcal{H}^0_\phi(\mu(\pi_2))$ and its frontier (Definition C.1)
Policy-value transfer (Section 3.6)	
$\mu_{e,k}$, $\nu_{e,k}$	monitoring and deployment trajectory laws for supplied policy or option chain π^k
A_k, $\Gamma_{e,k}$	core-compatible success event and mass: $A_k = \mathsf{S} \cap \{u_k \preccurlyeq \phi(\tau)\}$, $\Gamma_{e,k} = \mu_{e,k}(A_k) = p_{e,k} c_{\mu_{e,k}}(u_k)$
$G_{e,k}$, $V_{e,k}$	total trajectory return and deployment value $\mathbb{E}_{\nu_{e,k}} G_{e,k}$
$\chi_{e,k}$, $\xi_{e,k}$, $b_{e,k}$	shift in the probability of A_k, compatible-return error, and absolute return mass outside A_k (Assumption 3.2)
$\zeta_{e,k}$	performance-calibration radius $\xi_{e,k} + b_{e,k} + r_k\chi_{e,k}$
$\widehat{\Gamma}_{e,k}$, $L_{e,k}$	empirical compatible mass and its calibrated lower confidence value (Theorem 3.3)
Domain (SRC-L-N, Appendix F)	
L, N	corridor length and number of peers
α, q	per-peer clear probability; induced scalar $q = 1-(1-\alpha)^N$
s_0, g	start and goal; t_i, b_i top- and bottom-route positions
ρ_{rock}	binary rock flag; β peer hold-probability (E5)

Table 1: Notation.

B Extended Related Work

Table 2: **From established tools to structural lifetimes.** The analysis combines conditional-probability control with peer-induced variation to quantify survival and performance consequences.

Literature	Established object	Role in the structural-lifetime analysis
Planning landmarks	Necessary facts and orderings within a fixed planning task (Hoffmann et al., 2004; Richter & Westphal, 2010).	Quantitative coverage of abstract structure among stochastic successes under a changing trajectory law.
Frequent sequence mining	Discovery of constrained subsequences above an empirical support threshold (Girgin & Polat, 2005; Beedkar et al., 2019).	A transition- and peer-drift bound on conditional support loss, yielding a certified survival time.
Total-variation conditioning	Sharp conditional-probability envelopes in imprecise-probability models (Pelessoni & Vicig, 2022).	The established factor $\frac{1}{p_0}$ converts trajectory drift into coverage loss; accumulated peer drift then yields a structural lifetime.
Non-stationary MDPs	Dynamic-regret guarantees from reward and transition variation budgets (Even-Dar et al., 2009; Cheung et al., 2020; Mao et al., 2021).	Variation controls the survival of reusable structure and, with calibration, its policy-value and transfer consequences.
MARL non-stationarity	Focal-agent or best-response induced MDPs, plus algorithms compensating for co-learning and replay shift (Littman, 1994; Shoham & Leyton-Brown, 2008; Busoniu et al., 2008; Foerster et al., 2017; 2018).	A peer-update budget combined with sharp conditional coverage yields an explicit structural lifetime and downstream decision guarantees.
Continual RL	Formalizations and taxonomies of ongoing adaptation under non-stationarity (Khetarpal et al., 2022; Elelimy et al., 2025).	A success-conditioned structural witness whose coverage, lifetime, and value consequences are controlled under the continual sequence.

The conditional inequality in Lemma 3.2 specializes the total-variation model of Pelessoni & Vicig (2022, Section 2.1 and Proposition 3.2); its short proof is retained for self-containment. The contribution developed here is the structural-lifetime analysis: cumulative peer updates control coverage erosion, positive margins yield survival horizons, effective conflict gives a matching first-exit order, and calibrated execution connects survival to policy value and transfer. The focal induced-MDP representation identifies the source of variation, while established conditioning control supplies the probabilistic link in this chain.

Planning landmarks, RL bottlenecks, probabilistic landmarks, and sequence search supply useful candidate structures (McGovern & Barto, 2001; Şimşek & Barto, 2009; Chan et al., 2025). State abstraction (Dean & Givan, 1997; Li et al., 2006; Abel et al., 2016) shrinks the alphabet, while options (Sutton et al., 1999; Konidaris & Barto, 2009) supply executors. Our stability analysis transports these candidates across episodes; executability remains explicit in Appendix C.2.

Work on co-adapting agents addresses non-stationarity through replay correction, opponent modeling, and recursive reasoning (Hernandez-Leal et al., 2017; Papoudakis et al., 2019; He et al., 2016; Raileanu et al., 2018). Zero-shot coordination and ad hoc teamwork seek partner-robust behavior (Hu et al., 2020; Lupu et al., 2021; Stone et al., 2010); the robust core (Definition C.1) identifies structure shared across success-enabling partners. Inequality ($\star$) adds a finite-time survival guarantee for that structure under peer learning.

C Deferred Structural and Statistical Results

C.1 Existence and non-triviality

For the graph characterization below, call a successful trajectory *feasible* when every displayed action has positive transition probability along it. The probe has *full conditional support* when every feasible success has positive μ-mass. Define the *action-labelled time-unrolled success graph* G_{S} as follows: its vertices are (s,t), and it has an edge $(s,t) \xrightarrow{a} (s',t+1)$ whenever $P(s' \mid s,a) > 0$ and that step lies on a μ-positive success; the edge carries label $\phi(s,a)$. Because time and action are retained, every s_0–$\mathcal{G}$ path in G_{S} is a feasible successful trajectory, and full conditional support gives it positive μ-mass. Conversely, every μ-positive success traces such a path. A set of labelled edges is a *step-cut* when it intersects every one of these paths.

Theorem C.1 (Existence and non-triviality)**.** *Fix μ with $\mu(\mathsf{S}) > 0$.*

(a) ***(Existence)*** *For every $\delta \in [0,1)$, some frontier element of $\mathrm{Core}^{\delta}_{\phi}(\mu)$ contains g_Σ; in particular $\mathrm{Core}^{\delta}_{\phi}(\mu) \neq \varnothing$.*

(b) ***(Non-triviality)*** *A non-terminal symbol σ with $c_\mu(\sigma) = 1$ is a* ϕ-bottleneck*; some frontier element contains (σ, g_Σ) as a subsequence and therefore strictly extends (g_Σ). Under full conditional support, σ is a bottleneck iff its labelled steps cut every start–goal path in the time-unrolled success graph.*

(c) A list of non-terminal bottleneck occurrences $\sigma_1, \ldots, \sigma_m$ is order-consistent *if every μ-positive success τ has positions $i_1 < \cdots < i_m$ with $\phi(\tau)_{i_j} = \sigma_j$. Symbols may repeat, but their occurrence positions must be distinct. Some element of $\mathrm{Core}^{0}_{\phi}(\mu)$ then contains $(\sigma_1, \ldots, \sigma_m, g_\Sigma)$ and has length at least $m+1$.*

The goal symbol makes existence immediate; the content is the step-cut criterion. In key–door tasks, for example, "find key", "reach door", and "open door" can form ordered bottlenecks (Chevalier-Boisvert et al., 2018; Hung et al., 2019; Sun et al., 2023).

C.2 Sound pruning and execution requirements

Proposition C.1 (Soundness (admissible pruning))**.** *Suppose μ has full conditional support on the optimal successes (every optimal $\tau^\star$ has $\mu(\tau^\star) > 0$). For any $u \in \mathcal{H}^0_\phi(\mu)$ and any $\tau \in \mathsf{S}$ with $\mu(\tau) > 0$, $u \preccurlyeq \phi(\tau)$; in particular every optimal successful trajectory contains u, so restricting search to $\{\tau : \forall u \in \mathrm{Core}^0_\phi(\mu),\ u \preccurlyeq \phi(\tau)\}$ preserves* every *optimal successful trajectory (in particular an optimum). Without the support hypothesis the guarantee is probe-relative: pruning preserves an optimum only among μ-positive successes.*

This proposition supplies admissible trajectory-level pruning. Expected-value policy optimization additionally requires an executor and an ordering of the frontier:

Proposition C.2 (Non-orderability)**.** *Let a branching finite-horizon MDP have exactly two positive-mass successful abstract strings, (x, y, g_Σ) and (y, x, g_Σ). Their maximal common subsequences are (x, g_Σ) and (y, g_Σ), so $\mathrm{Core}^0_\phi(\mu) = \{(x, g_\Sigma), (y, g_\Sigma)\}$ while neither (x, y) nor (y, x) is generated. Hence no concatenation of core symbols is forced to be a valid plan, and any executor must supply an order the core does not determine.*

Model each symbol as an option $o_\sigma = (I_\sigma, \pi_\sigma, \beta_\sigma)$ on time-augmented states (s,t) (Sutton et al., 1999); its initiation set, policy $\pi_\sigma(\cdot \mid s,t)$, and termination rule may depend on the remaining horizon.

Assumption C.1 (Execution)**.** (A1) the first option can start at s_0; (A2) each option emits its symbol on termination; (A3) consecutive options are chainable; (A4) the chain terminates almost surely within $H-1$ transitions; and (A5) its final termination lies in $\mathcal{G}$.

Proposition C.3 (Ordered-core Bellman decomposition)**.** *Let $(\sigma_1, \ldots, \sigma_r, g_\Sigma)$ be one ordered frontier string. Under Assumption C.1, its option chain reaches $\mathcal{G}$. For additive costs, define $V^\star(s,t)$ as the infimum expected remaining cost among policies that reach $\mathcal{G}$ almost surely by time $H-1$.*

Suppose each σ_i labels a step entering one fixed state g_i, and every success enters $g_1, \dots, g_r$ in this order. Set $g_0 = s_0$, and let the final g_Σ-labelled step enter a fixed goal state $g_{r+1} \in \mathcal{G}$. For each $i = 1, \dots, r+1$, let $\mathcal{P}_i(t)$ contain the segment policies that start from (g_{i-1}, t) and terminate at g_i almost surely within the remaining horizon. Let (C_i, T_i) denote a segment option's cost and terminal time. Suppose each supplied segment option belongs to $\mathcal{P}_i(t)$ and attains

$$V^\star(g_{i-1}, t) = \inf_{\pi_i \in \mathcal{P}_i(t)} \mathbb{E}[C_i + V^\star(g_i, T_i) \mid g_{i-1}, t] \qquad (i = 1, \dots, r+1),$$

where $V^\star(g_{r+1}, \cdot) = 0$. Then the supplied option chain attains the goal-reaching comparator $V^\star$. The entered-state landmarks separate the time-unrolled graph, which makes this recursion Bellman-optimal.

Proposition C.3 gives a Bellman decomposition for supplied options attaining the segment infima. Entered-state landmarks provide the separators needed for optimal chaining, while the abstraction and option library jointly determine executability.

C.3 Complexity of core computation

For the support core, computing *one longest* frontier element is a longest-common-subsequence problem. For k successful strings, exact product-space dynamic programming visits

$$O\left(\prod_{i=1}^{k}(H_i + 1)\right) = O\big(H^k\big),$$

and generalized LCS is NP-hard (Maier, 1978). The product-space dynamic program returns a longest frontier witness exactly. A pointwise abstraction shrinks the alphabet, while temporal abstraction shortens the strings. For $\delta > 0$, support-constrained sequence-search procedures such as DESQ provide candidate-discovery machinery (Beedkar et al., 2019). Given a discovered candidate set $\mathcal{U}$, a two-pointer subsequence scan evaluates all empirical coverages in $O(N|\mathcal{U}|H)$ time; the implemented length-1 and length-2 monitors instantiate this evaluator.

C.3.1 Statistical cost: fixed-candidate membership testing

For N i.i.d. probe rollouts, let N_{S} be the number of successes and

$$\widehat{c}_N(u) = \frac{1}{N_{\mathsf{S}}} \sum_{i:\tau_i \in \mathsf{S}} \mathbf{1}[u \preccurlyeq \phi(\tau_i)] \qquad (N_{\mathsf{S}} > 0).$$

Set $\widehat{c}_N(u) = 0$ when $N_{\mathsf{S}} = 0$.

Theorem C.2 (Fixed-candidate membership testing)**.** *Let $p_0 = \mu(\mathsf{S}) > 0$, $t \in (0, 1]$, $\gamma > 0$, and $\rho \in (0, 1)$.*

(a) If $N \geq \frac{1}{p_0 t^2} \log \frac{4}{\rho} + \frac{8}{p_0} \log \frac{2}{\rho}$, then with probability at least $1 - \rho$, $|\widehat{c}_N(u) - c_\mu(u)| \leq t$.

(b) For M pre-specified candidates each γ-separated from $1 - \delta$, $N = O(\frac{\log(\frac{M}{\rho})}{p_0 \gamma^2})$ recovers every membership with probability at least $1 - \rho$.

(c) At a fixed interior threshold, one candidate and constant error require $N = \Omega(\frac{1}{p_0 \gamma^2})$; the constant is threshold-dependent near 0 and 1.

Part (a) combines two requirements. The second term ensures that the sample contains enough successful trajectories; conditional on those successes, the first term controls the coverage estimate. Parts (b) and (c) give the corresponding uniform upper rate and matching lower rate.

Remark C.1 (Estimation and drift are interchangeable budgets). With known p_0, sampling slack t and drift slack $\frac{\varepsilon_e}{p_0}$ consume the same margin: if $\widehat{c}_N(u) - (1 - \bar{\delta}) \geq t + \frac{\varepsilon_e}{p_0}$, then Theorems 3.1 and C.2 certify $u \in \mathcal{H}_\phi^{\bar{\delta}}(\mu_e)$ with probability at least $1 - \rho$. For unknown success mass, use the lower confidence bound in Proposition E.1; an empirical success fraction alone need not be conservative.

C.4 THE ROBUST CORE

Definition C.1 (Robust (partner-universal) core)**.** For success-enabling peer policies Π_2^+, define $\mathcal{H}_{\mathrm{rob}} = \bigcap_{\pi_2 \in \Pi_2^+} \mathcal{H}_\phi^0(\mu(\pi_2))$ and $\mathrm{Core}_{\mathrm{rob}} = \max_{\preccurlyeq} \mathcal{H}_{\mathrm{rob}}$. The intersection is taken over families, not frontiers.

For any peer policy π_2, write $\mathrm{Supp}_{\mathsf{S}}(\pi_2) := \{\tau \in \mathsf{S} : \mu(\pi_2)(\tau) > 0\}$.

Proposition C.4 (Overlap collapses to the robust core)**.** *The intersection of any two success-enabling episode families contains $\mathcal{H}_{\mathrm{rob}}$. For a visited peer sequence $\{\pi_2^e\}_{e \geq 0} \subseteq \Pi_2^+$, if*

$$\bigcup_{e \geq 0} \mathrm{Supp}_{\mathsf{S}}(\pi_2^e) = \bigcup_{\pi_2 \in \Pi_2^+} \mathrm{Supp}_{\mathsf{S}}(\pi_2),$$

then $\bigcap_{e \geq 0} \mathcal{H}_\phi^0(\mu_e) = \mathcal{H}_{\mathrm{rob}}$. In SRC, top and bottom routes share only (s_0, g), so $\mathrm{Core}_{\mathrm{rob}} = \{(s_0, g)\}$.

C.5 AN ANALYTIC EFFECTIVE-CONFLICT CLASS

Proposition C.5 (Non-vacuity: exact policy gradient in a two-route game)**.** *In a two-route game, a peer chooses* keep*, entering the state labelled u, or* divert*, entering a bypass; either route then reaches the goal. Parameterize*

$$\pi_\theta(\textit{divert}) = \varsigma(\theta), \qquad c(\theta) = \pi_\theta(\textit{keep}) = \varsigma(-\theta), \qquad \varsigma(x) = (1 + e^{-x})^{-1}.$$

Let the divert reward exceed the keep reward by $\Delta > 0$ and use exact gradient ascent on expected reward. For target coverage $\ell := 1 - \bar{\delta}$ with $0 < \ell < c_0 := c(\theta_0) < 1$, put

$$r_\star := \min_{x \in [\ell, c_0]} x(1 - x) > 0, \qquad m := c_0 - \ell.$$

Then (P2) holds termwise with $\lambda = \Delta r_\star^2$; also $G = \frac{\Delta}{4}$, $L_c = \frac{1}{4}$, and $\mathrm{TV}(\pi_{\theta_{e+1}}, \pi_{\theta_e}) \leq \frac{\Delta}{16}\eta$. For $\eta \leq \frac{64 r_\star^2}{\Delta}$,

$$\left\lfloor \frac{8m}{\Delta \eta} \right\rfloor + 1 \;\leq\; E_{\mathrm{exit}} \;\leq\; \left\lfloor \frac{2m}{\Delta r_\star^2 \eta} \right\rfloor + 1,$$

so the two-sided theorem is non-vacuous.

D FULL PROOFS

D.1 STANDARD TOOLS AND CONVENTIONS

These facts are standard background tools rather than assumptions on the Markov game.

Total variation and Markov kernels. For probability measures p, q on a finite or countable space,

$$\mathrm{TV}(p,q) := \sup_A |p(A) - q(A)| = \frac{1}{2}\|p - q\|_1 .$$

Thus $|p(A) - q(A)| \le \mathrm{TV}(p,q)$ for every event A. A Markov kernel is an ℓ_1 contraction:

$$\|pK - qK\|_1 \le \|p - q\|_1 .$$

Indeed, summing the absolute signed mass after applying K and then using that each row of K sums to one gives the right-hand side. Combining contraction with the triangle inequality yields the one-step comparison

$$\|pK - qL\|_1 \le \|p - q\|_1 + \sup_x \|K(x,\cdot) - L(x,\cdot)\|_1 . \tag{1}$$

For controlled kernels we use the rowwise norm

$$\|P - Q\|_{1,\infty} := \sup_{s,a} \sum_{s'} |P(s' \mid s,a) - Q(s' \mid s,a)|.$$

Equation (1) is the elementary telescoping step behind the factor $H - 1$ in trajectory simulation.

Independent products. Tensoring with a common probability factor leaves total variation unchanged, so

$$\mathrm{TV}(p \otimes r, q \otimes r) = \mathrm{TV}(p,q).$$

Replacing independent factors one at a time and applying the triangle inequality therefore gives

$$\mathrm{TV}\left(\bigotimes_{j=1}^m p_j, \bigotimes_{j=1}^m q_j\right) \le \sum_{j=1}^m \mathrm{TV}(p_j, q_j). \tag{2}$$

Concentration and simultaneous events. If $N_{\mathsf{S}} \sim \mathrm{Bin}(N, p_0)$, the multiplicative Chernoff bound at half the mean is

$$\Pr\left[N_{\mathsf{S}} < \frac{Np_0}{2}\right] \le \exp\left(-\frac{Np_0}{8}\right). \tag{3}$$

If $\widehat{c}$ is the mean of n independent Bernoulli variables with mean c, Hoeffding's inequality gives

$$\Pr[|\widehat{c} - c| > t] \le 2\exp(-2nt^2). \tag{4}$$

For events $E_1, \dots, E_M$, the union bound $\Pr(\bigcup_i E_i) \le \sum_i \Pr(E_i)$ converts these pointwise statements into the simultaneous guarantees used for fixed candidate families and policy libraries. No independence across candidates or checkpoints is required for this last step.

Two-point testing. For $x, y \in (0,1)$, Bernoulli relative entropy satisfies

$$\mathrm{KL}(\mathrm{Bern}(x) \| \mathrm{Bern}(y)) \le \frac{(x-y)^2}{y(1-y)}. \tag{5}$$

For product observations, relative entropy tensorizes and Pinsker's inequality converts it to total variation:

$$\begin{aligned} \mathrm{KL}(P^{\otimes N} \| Q^{\otimes N}) &= N\,\mathrm{KL}(P\|Q), \\ \mathrm{TV}(P,Q) &\le \sqrt{\frac{1}{2}\mathrm{KL}(P\|Q)}. \end{aligned} \tag{6}$$

Finally, every binary test ψ obeys the two-point bound

$$\max\{P(\psi = 1), Q(\psi = 0)\} \ge \frac{1 - \mathrm{TV}(P,Q)}{2}. \tag{7}$$

These facts turn small information distance between two threshold-separated models into a necessary sample size.

Smoothness and logistic policies. If f has an L-Lipschitz gradient, integration along the line segment from x to $x+h$ gives the descent lemma

$$f(x+h) \le f(x) + \langle \nabla f(x), h \rangle + \frac{L}{2} \|h\|_2^2 . \tag{8}$$

For the logistic map $\varsigma(x) = (1+e^{-x})^{-1}$,

$$\varsigma'(x) = \varsigma(x)(1-\varsigma(x)), \qquad |\varsigma'(x)| \le \frac{1}{4}, \tag{9}$$

so ς is $\frac{1}{4}$-Lipschitz. Equation (8) controls the second-order cost of a policy update, while Equation (9) converts parameter movement into policy total variation in the analytic example.

The sharp conditioning relation used in Lemma 3.2 is also established background: it specializes the total-variation model of Pelessoni & Vicig (2022, Section 2.1 and Proposition 3.2). The proof below makes the specialization and its equality case explicit.

D.2 Proofs

Proof of Lemma 2.1. If $u \preceq w \preceq \phi(\tau)$ then $u \preceq \phi(\tau)$, so the occurrence event for w is contained in that for u and $c_\mu(u) \ge c_\mu(w)$. Thus $\mathcal{H}_\phi^\delta(\mu)$ is hereditary. For any u in this family, choose among its finitely many high-coverage extensions one of maximum length. It is maximal, and heredity places everything below every such maximal element back in the family. □

Proof of Theorem C.1. Step 1: establish non-emptiness. By the goal-symbol convention every $\tau \in \mathsf{S}$ satisfies $g_\Sigma \preceq \phi(\tau)$, so $c_\mu(g_\Sigma) = 1 \ge 1-\delta$. Thus (g_Σ) belongs to the finite high-coverage family, which consequently has a $\preceq$-maximal element. This proves part (a).

Step 2: extend a bottleneck to the frontier. Let σ be a ϕ-bottleneck. Since $\sigma \ne g_\Sigma$ and g_Σ is the unique terminal label, every positive-mass success contains σ before g_Σ. Hence (σ, g_Σ) is common. Extending this common sequence within the finite hereditary family yields a maximal element $u^\star$ with $(\sigma, g_\Sigma) \preceq u^\star$. In particular, $u^\star$ strictly extends (g_Σ).

Step 3: identify the cut condition. Work in the action-labelled time-unrolled success graph G_{S} defined before the theorem. A path records its state, time, and action at every edge. Each edge has positive transition probability, so the Markov property makes the product of its transition probabilities positive; an s_0–$\mathcal{G}$ path is therefore a feasible success. Full conditional support assigns every such success positive μ-mass. Conversely, every positive-mass success traces one of these paths. If the σ-labelled edges are not a step-cut, a path avoids them and gives a positive-mass success without σ, hence $c_\mu(\sigma) < 1$. Conversely, a step-cut forces every success to take a σ-labelled step. This proves the equivalence in part (b).

Step 4: assemble ordered bottlenecks. If $\sigma_1, \dots, \sigma_m$ are order-consistent, each positive-mass success has indices $i_1 < \cdots < i_m$ carrying them. All these labels precede the unique terminal label, so $(\sigma_1, \dots, \sigma_m, g_\Sigma)$ is common. Any maximal common extension has length at least $m+1$. This proves part (c). □

Why time indexing is necessary. The un-indexed union of success edges can splice transitions from different successes into an infeasible walk. With $H = 4$ and successes $s_0 \to a \to z \to g$, $s_0 \to a \to b \to g$, $s_0 \to b \to c \to g$, and $s_0 \to c \to g$, label every success so that it contains σ. The spliced walk $s_0 \to a \to b \to c \to g$ can avoid σ, but it uses four transitions and exceeds the budget $H - 1 = 3$. Such walks are absent from G_{S} because time is part of each vertex.

Proof of Proposition C.1. By Lemma 2.1, $u \in \mathcal{H}_\phi^0(\mu)$ means $c_\mu(u) = 1$, i.e. $\mu(\{u \preceq \phi(\tau)\} \mid \mathsf{S}) = 1$, so every μ-positive success contains u. Under the full-support hypothesis every optimal $\tau^\star$ has $\mu(\tau^\star) > 0$, hence $u \preceq \phi(\tau^\star)$. Restricting the feasible set to trajectories containing all core elements therefore removes no optimal successful trajectory. Thus the core restriction is admissible for optimal successful-trajectory search; expected-return guarantees follow after executability calibration, as formalized in Theorem 3.3. □

Proof of Proposition C.3. Step 1: establish goal-reaching. Assumption (A1) starts the first option, (A2) realizes every σ_i, (A3) chains consecutive options, and (A4)–(A5) terminate the chain in $\mathcal{G}$ within $H-1$ transitions. Thus the chain reaches the goal almost surely.

Step 2: expose the separating states. Use the comparator $V^\star$ and admissible segment classes $\mathcal{P}_i(t)$ defined in the proposition. Index each landmark by the state its labelled step *enters*: let g_i be the unique entered state for σ_i, put $g_0 = s_0$, and let $g_{r+1} \in \mathcal{G}$ be entered by the final g_Σ-labelled step. By hypothesis, every success enters $g_1, \dots, g_r$ in order, so each $(g_i, \cdot)$ separates s_0 from $\mathcal{G}$ in the time-unrolled success graph. Consequently every policy admissible for $V^\star(g_{i-1}, t)$ has a first arrival at g_i and decomposes into an admissible segment in $\mathcal{P}_i(t)$ followed by a continuation from (g_i, T_i).

Step 3: write the Bellman decomposition. Let T_i be the arrival time at g_i and C_i the cost of segment i, including the final goal-labelled transition in segment $r+1$. Backward induction on the time-augmented state gives

$$V^\star(g_{r+1}, t) = 0, \qquad V^\star(g_{i-1}, t) = \inf_{\pi_i \in \mathcal{P}_i(t)} \mathbb{E}\big[\, C_i + V^\star(g_i, T_i) \,\big|\, g_{i-1}, t \,\big] \quad (i = r{+}1, \dots, 1).$$

The separators prevent probability mass from bypassing a segment, so this is exactly the Bellman equation on the ordered-landmark structure. By the proposition's attainment premise, each supplied option realizes its segment infimum; composing them therefore realizes $V^\star(s_0, 0)$. The downstream value appears once, inside the recursion. For deterministic segment durations the recursion reduces to $\sum_i d^\star(g_{i-1}, g_i)$. □

Why entered-state singletons are needed. Indexing by the state a labelled step leaves would place the terminal symbol outside $\mathcal{G}$ and omit the final transition cost. Likewise, if $\phi^{-1}(\sigma_i)$ can enter two states g_i, g_i' with different continuation values, a locally cheapest preceding option can terminate at the state with the larger downstream cost. The resulting greedy chain is strictly suboptimal, with a gap that can be $\Omega(H)$.

Proof of Proposition C.2. Construct a start state with two actions leading to two deterministic branches. Label the first branch in order by (x, y, g_Σ) and the second by (y, x, g_Σ), and let the probe select both start actions with positive probability. These are exactly the positive-mass successes. Their maximal common subsequences are (x, g_Σ) and (y, g_Σ), which form an antichain; neither linear ordering of x, y is common. □

Proof of Proposition 3.1. Step 1: marginalize one transition. Fix episode e and a focal history ending at $(s_t, a_{1,t})$. The peers sample their joint action from the stated Markov kernel; persistent randomness is part of the state. Hence

$$\Pr(s_{t+1} = s' \mid s_t, a_{1,t}) = \sum_{a_{-1}} P(s' \mid s_t, a_{1,t}, a_{-1}) \pi^e_{-1}(a_{-1} \mid s_t) = P_e(s' \mid s_t, a_{1,t}).$$

Step 2: lift the identity to complete trajectories. The initial laws agree. If the two models agree on every length-t focal history, then the focal policy gives the same next-action law and Step 1 gives the same next-state law; therefore they agree on every length-$(t+1)$ history. Induction proves equality of the complete focal trajectory laws. The same marginalization gives conditional stage reward $R_e(s_t, a_{1,t})$, so expected additive returns agree. Success, abstraction, subsequence occurrence, coverage, and the core are functions of this law and therefore transport exactly.

Step 3: translate peer drift into MDP drift. For every (s, a_1),

$$\begin{aligned}\sum_{s'} |P_e(s' \mid s, a_1) - P_{e-1}(s' \mid s, a_1)| &\le \sum_{a_{-1}} |\pi^e_{-1}(a_{-1} \mid s) - \pi^{e-1}_{-1}(a_{-1} \mid s)| \\ &= 2\mathrm{TV}(\pi^e_{-1}(\cdot \mid s), \pi^{e-1}_{-1}(\cdot \mid s)).\end{aligned}$$

Taking the supremum proves the kernel bound. Similarly,

$$|R_e(s, a_1) - R_{e-1}(s, a_1)| \le R_{\max} \sum_{a_{-1}} |\pi^e_{-1}(a_{-1} \mid s) - \pi^{e-1}_{-1}(a_{-1} \mid s)| \le 2R_{\max} D_e.$$

Step 4: separate independent peers. If the peers act conditionally independently, their joint law is a product. Replacing one factor at a time and using the product bound equation 2 gives $\mathrm{TV}(\bigotimes_{j=2}^{n}\pi_j^e, \bigotimes_{j=2}^{n}\pi_j^{e-1}) \le \sum_{j=2}^{n}\mathrm{TV}(\pi_j^e, \pi_j^{e-1})$ pointwise in s, which proves the final claim. □

Proof of Proposition 3.2. The Switch–Rock family (Appendix F) has, at $\alpha = 0$, exactly the top-route successes, whose focal-position string is (s_0, t_1, t_2, g); so $\mathrm{Core}_\phi^0(\mu_0) = \{(s_0, t_1, t_2, g)\}$ and t_1 is generated ($t_1 \preceq (s_0, t_1, t_2, g)$, a ϕ-bottleneck by Theorem C.1(b)). For $0 < \alpha \le \frac{1}{2}$ the peer clears the rock with positive probability, opening the bottom route, while holding also has positive probability, preserving top-route successes. Thus both route strings occur with positive mass, $c_{\mu_\alpha}(t_1) < 1$, and the frontier collapses to $\{(s_0, g)\}$. The kernel bound $\|P_\alpha - P_0\|_{1,\infty} \le 2\sup_s \mathrm{TV}(\pi_2^\alpha, \pi_2^0) = 2\alpha$ follows from the definition of the induced kernel and the identity $\|\sum_{a_2} P(\cdot \mid s, a_1, a_2)\Delta\pi_2(a_2)\|_1 \le \sum_{a_2}|\Delta\pi_2(a_2)| = 2\mathrm{TV}$. Since the indicator jumps from 1 to 0 at $\alpha = 0^+$ while the drift $\to 0$, no modulus of continuity exists. □

Proof of Lemma 3.1. Step 1: place both processes on one history space. Pad terminated trajectories with an absorbing state. Let ν_t^e, ν_t^0 be the resulting history laws after t transitions, put $d_t := \|\nu_t^e - \nu_t^0\|_1$, and write $\Delta := \|P_e - P_0\|_{1,\infty}$. Because the focal probe is fixed, the two history-extension kernels K_e, K_0 differ by at most Δ in rowwise ℓ_1.

Step 2: telescope the one-step discrepancy. Applying the one-step kernel comparison equation 1 gives

$$\begin{aligned} d_{t+1} &= \|\nu_t^e K_e - \nu_t^0 K_0\|_1 \\ &\le \|\nu_t^e K_e - \nu_t^0 K_e\|_1 + \|\nu_t^0 K_e - \nu_t^0 K_0\|_1 \le d_t + \Delta. \end{aligned}$$

Since $d_0 = 0$, induction yields $d_{H-1} \le (H-1)\Delta$.

Step 3: convert ℓ_1 distance to total variation. Using the total-variation identity stated in Appendix D.1 proves $\|\mu_e - \mu_0\|_{\mathrm{TV}} \le \frac{1}{2}(H-1)\Delta = \varepsilon_e$. □

Proof of Lemma 3.2. *Connection to conditioning literature.* In the finite trajectory space, the probability laws in the total-variation model with reference $P_0 = \mu$ and parameters $b = 1$, $a = -\varepsilon$ are exactly those with $\mathrm{TV}(\nu, \mu) \le \varepsilon$. For $0 < \varepsilon < \mu(A)$ and $\mathsf{S} \setminus A \ne \varnothing$, Proposition 3.2 of Pelessoni & Vicig (2022), with conditioning event $B = \mathsf{S}$, gives the lower conditional envelope

$$\inf_{\nu:\,\mathrm{TV}(\nu,\mu)\le\varepsilon} \nu(A \mid \mathsf{S}) = \frac{\mu(A) - \varepsilon}{\mu(\mathsf{S})} = \mu(A \mid \mathsf{S}) - \frac{\varepsilon}{p}.$$

When $A = \mathsf{S}$, the conditional probability is one; when $\varepsilon = 0$, the laws coincide. For self-containment, the following direct proof covers all cases of the stated inequality.

Put $C := \mathsf{S} \setminus A$, $a := \mu(A)$, and $c := \mu(C)$, so $a + c = p$.

Step 1: bound the two success cells. Total variation gives

$$\nu(A) \ge a - \varepsilon, \qquad \nu(C) \le c + \varepsilon.$$

The assumption $\varepsilon < a$ implies $\nu(A) > 0$, and hence $\nu(\mathsf{S}) > 0$.

Step 2: normalize only after the extremal move. The map $(x, y) \mapsto \frac{x}{x+y}$ is increasing in x and decreasing in y. Therefore

$$\nu(A \mid \mathsf{S}) = \frac{\nu(A)}{\nu(A) + \nu(C)} \ge \frac{a - \varepsilon}{(a - \varepsilon) + (c + \varepsilon)} = \frac{a}{p} - \frac{\varepsilon}{p} = \mu(A \mid \mathsf{S}) - \frac{\varepsilon}{p}.$$

Step 3: show sharpness. On a three-cell space (A, C, S^c), choose $\mu(A) > \varepsilon$ and obtain ν by moving exactly ε mass from A to C. The total-variation distance is ε, the success mass remains p, and equality holds above. Thus no smaller universal coefficient than $\frac{1}{p}$ is possible. □

Proof of Theorem 3.1. Step 1: move from kernels to trajectories. By Lemma 3.1, $\mathrm{TV}(\mu_e, \mu_0) \le \varepsilon_e$.

Step 2: apply the sharp conditioning lemma. Let $A := \{u \preceq \phi(\tau)\} \cap \mathsf{S}$. Then $\mu_0(A) = p_0 c_{\mu_0}(u) \ge (1-\delta)p_0 > \varepsilon_e$. Applying Lemma 3.2 with $(\mu, \nu) = (\mu_0, \mu_e)$ gives

$$c_{\mu_e}(u) \ge c_{\mu_0}(u) - \frac{\varepsilon_e}{p_0} \ge (1-\delta) - \frac{\varepsilon_e}{p_0}.$$

Step 3: return to the core family. Put $\delta' := 1 - c_{\mu_e}(u)$. The last display gives $\delta' \le \delta + \frac{\varepsilon_e}{p_0} < 1$, so $u \in \mathcal{H}_\phi^{\delta + \frac{\varepsilon_e}{p_0}}(\mu_e)$; Lemma 2.1 then places u below a frontier element of the same family. Tightness follows from the mass transfer in Step 3 of Lemma 3.2. □

Remark on the hypothesis. The condition $\varepsilon_e < (1-\delta)p_0$ is exact for the displayed non-vacuous conclusion. The weaker requirement $\varepsilon_e < p_0$ does not suffice: with $p_0 = 1$, $\delta = 0.9$, and $\varepsilon_e = 0.1$, coverage can fall to zero.

Proof of Proposition 3.3. Let $\mathcal{K}_e = \mathcal{K}_e^{\delta_{\mathrm{ref}} \to \bar{\delta}}$ as defined before the proposition. If $u \preceq v$, then the occurrence event for v is contained in that for u, hence $c_{\mu_e}(u) \ge c_{\mu_e}(v)$ at every episode. Therefore $v \in \mathcal{K}_e$ and $u \preceq v$ imply $u \in \mathcal{K}_e$, so $\mathcal{K}_e$ is hereditary and has a finite maximal frontier. The same inequality implies that whenever v remains above the target threshold, so does u; hence $E_{\mathrm{exit}}^{\bar{\delta}}(u) \ge E_{\mathrm{exit}}^{\bar{\delta}}(v)$. This argument is episode-wise and permits later re-entry. □

Proof of Corollary 3.1. Step 1: accumulate peer movement. The kernel triangle inequality and Proposition 3.1 give

$$\|P_e - P_0\|_{1,\infty} \le V_e \le 2 \sum_{j=1}^{e} \sup_s \mathrm{TV}(\pi_2^j, \pi_2^{j-1}) \le 2c\eta e.$$

Hence $\varepsilon_e \le c\eta(H-1)e$.

Step 2: spend the coverage margin. The initial slack is δ_0 and the target slack is $\bar{\delta}$, so the available margin is $m = \bar{\delta} - \delta_0$. By Theorem 3.1, membership in $\mathcal{H}^{\bar{\delta}}$ is certified whenever

$$\frac{\varepsilon_e}{p_0} \le m, \qquad \text{i.e.} \qquad e \le \frac{m p_0}{c\eta(H-1)}.$$

Because $m < 1 - \delta_0$, the premise $\varepsilon_e < (1-\delta_0)p_0$ also holds throughout this range.

Step 3: respect episode integrality. Taking the integer part gives the displayed $E^\star = \Omega(\frac{1}{\eta})$. Membership holds through episode $E^\star$, so $E_{\mathrm{exit}}^{\bar{\delta}}(u) \ge E^\star + 1$. □

Proof of Proposition C.5. Step 1: compute the exact update direction. Since $c'(\theta) = -c(\theta)(1-c(\theta))$ and the divert reward exceeds the keep reward by Δ,

$$d_e = J'(\theta_e) = -\Delta c'(\theta_e) = \Delta c(\theta_e)(1 - c(\theta_e)) > 0.$$

Step 2: verify effective conflict. Before exit, $c(\theta_e) \in [\ell, c_0]$, and therefore

$$-c'(\theta_e) d_e = \Delta [c(\theta_e)(1-c(\theta_e))]^2 \ge \Delta r_\star^2.$$

Thus (P2) holds termwise with $\lambda = \Delta r_\star^2$.

Step 3: verify boundedness and smoothness. We have $|d_e| \le \frac{\Delta}{4}$ and $|c''(\theta)| = c(\theta)(1 - c(\theta))|1 - 2c(\theta)| \le \frac{1}{4}$, establishing (P1) and (P3). By equation 9, the logistic map is $\frac{1}{4}$-Lipschitz, so $\mathrm{TV}(\pi_{\theta_{e+1}}, \pi_{\theta_e}) \le \frac{\eta |d_e|}{4} \le \frac{\Delta \eta}{16}$.

Step 4: invoke the lifetime theorem. Apply Theorem 3.2 with $p_0 = 1$, $H - 1 = 2$, $c = \frac{\Delta}{16}$, and $\lambda = \Delta r_\star^2$. Its step-size condition becomes $\eta \le \frac{64 r_\star^2}{\Delta}$, and substitution gives the displayed bounds. □

Proof of Lemma 3.3. Step 1: control one update. The smoothness inequality equation 8 with $h = \eta d_e$ gives $c(\theta_{e+1}) \le c(\theta_e) + \eta\langle \nabla c(\theta_e), d_e\rangle + \frac{L_c}{2}\eta^2\|d_e\|_2^2$.

Step 2: sum before first exit. Summing to E and applying (P1)–(P2) yields

$$c(\theta_E) \le c(\theta_0) - \eta E\left(\lambda - \frac{L_c G^2}{2}\eta\right) \le c(\theta_0) - \frac{\lambda}{2}\eta E.$$

□

Proof of Theorem 3.2. Step 1: certify the lower lifetime. Corollary 3.1 gives $E_{\text{exit}} \ge \lfloor \frac{mp_0}{c\eta(H-1)} \rfloor + 1$. Also $c_{\mu_0}(u) = 1 - \bar{\delta} + m > 1 - \bar{\delta}$, so the pattern is present initially.

Step 2: force an exit by cumulative erosion. Put $K = \lfloor \frac{2m}{\lambda\eta} \rfloor + 1$. If $K < E_{\text{exit}}$, then Lemma 3.3 gives $c_{\mu_K}(u) \le c_{\mu_0}(u) - \frac{\lambda}{2}\eta K < c_{\mu_0}(u) - m = 1 - \bar{\delta}$. This contradicts membership before the first exit. Hence $E_{\text{exit}} \le K$ and the exit is finite.

Step 3: check the boundary episode. If $E_{\text{exit}} > 1$, the last pre-exit episode satisfies

$$1 - \bar{\delta} \le c_{\mu_{E_{\text{exit}}-1}}(u) \le c_{\mu_0}(u) - \frac{\lambda}{2}\eta(E_{\text{exit}} - 1),$$

so $E_{\text{exit}} - 1 \le \frac{2m}{\lambda\eta}$. The floor-plus-one form records integrality and retention at threshold equality. □

Proof of Proposition 3.4. Membership drift is realized by the SRC clearing construction: any positive clearing probability gives the bottom route positive successful mass, so the top-only pattern leaves the support family. For executability drift, let the peer clear with probability zero and hold the top gate independently with probability $\beta > 0$ at each attempt. Every successful trajectory still follows the top convention and contains the same minimal abstract string $(s_0, t_1, t_2, g_\Sigma)$; because a one-attempt success has positive probability for every $\beta > 0$, the full support core and its frontier are unchanged. Now fix an option calibrated at $\beta = 1$ that repeatedly attempts the gate until it opens, terminating after k failed attempts. Independence gives its probability of reaching the next initiation set as

$$1 - (1 - \beta)^k,$$

which tends to zero as $\beta \downarrow 0$. Thus the supplied chain can fail (A3)–(A5) with probability arbitrarily close to one while exact core membership and the frontier remain unchanged. □

Proof of Theorem 3.3. Fix (e, k) and abbreviate $A = A_k$, $\mu = \mu_{e,k}$, $\nu = \nu_{e,k}$, $G = G_{e,k}$, $r = r_k$, and $\Gamma = \Gamma_{e,k}$.

Step 1: decompose value around compatible success mass. Splitting on A gives the exact identity

$$V_{e,k} - r\Gamma = \mathbb{E}_\nu[(G - r)\mathbf{1}[A]] + \mathbb{E}_\nu[G\mathbf{1}[A^c]] + r\big(\nu(A) - \mu(A)\big).$$

The three terms are bounded respectively by $\xi_{e,k}$, $b_{e,k}$, and $r_k\chi_{e,k}$. The triangle inequality yields $|V_{e,k} - r_k\Gamma_{e,k}| \le \zeta_{e,k}$.

Step 2: obtain the value floor. If $u_k \in \mathcal{H}_\phi^\delta(\mu_{e,k})$ and $p_{e,k} \ge \underline{p}_k$, then $\Gamma_{e,k} = p_{e,k}c_{\mu_{e,k}}(u_k) \ge \underline{p}_k(1 - \delta)$. The lower side of Step 1 gives the displayed floor.

Step 3: construct simultaneous value intervals. Since $\mathbf{1}[A_k]$ is Bernoulli with mean $\Gamma_{e,k}$, equation 4 and a union bound over $\mathcal{I}$ imply, with probability at least $1 - \rho$, that $|\widehat{\Gamma}_{e,k} - \Gamma_{e,k}| \le \beta_{e,k}$ for every $(e, k) \in \mathcal{I}$. Work on this event and put $q_{e,k} := r_k\beta_{e,k} + \zeta_{e,k}$. The value approximation implies

$$L_{e,k} = r_k\widehat{\Gamma}_{e,k} - q_{e,k} \le V_{e,k} \le r_k\widehat{\Gamma}_{e,k} + q_{e,k} = L_{e,k} + 2q_{e,k}.$$

Step 4: compare the selected library element. Since $\widehat{k}_e$ maximizes $L_{e,k}$, every comparator k satisfies

$$V_{e,k} \le L_{e,k} + 2q_{e,k} \le L_{e,\widehat{k}_e} + 2q_{e,k} \le V_{e,\widehat{k}_e} + 2q_{e,k}.$$

Rearranging proves the comparator-wise guarantee. Taking $k = k_e^\star$ and summing proves cumulative transfer regret. The simultaneous event, rather than checkpoint independence, drives this argument; only the monitoring rollouts within each fixed (e, k) evaluation need the stated law. □

Proof of Corollary 3.2. Corollary 3.1 gives $u_k \in \mathcal{H}_\phi^{\bar{\delta}}(\mu_{e,k})$ for every $e \le E^\star$. Apply the value floor in Theorem 3.3 with $\delta = \bar{\delta}$, $p_{e,k} \ge \underline{p}_k$, and $\zeta_{e,k} \le \bar{\zeta}_k$. For the endogenous floor, total variation gives $p_{e,k} \ge p_{0,k} - \varepsilon_{e,k}$; the certified horizon ensures $\varepsilon_{e,k} \le m p_{0,k}$, and $0 < m < 1$. □

Proof of Proposition C.4. If $u \in \mathcal{H}_{\mathrm{rob}}$ then $c_{\mu(\pi_2)}(u) = 1$ for every $\pi_2 \in \Pi_2^+$, in particular for success-enabling π_2^e, π_2^{e+1}, so $u \in \mathcal{H}_\phi^0(\mu_e) \cap \mathcal{H}_\phi^0(\mu_{e+1})$; this gives the stated inclusion. A pattern has coverage one under every visited peer exactly when it occurs in every trajectory in $\bigcup_e \mathrm{Supp}_{\mathsf{S}}(\pi_2^e)$. Under the displayed union-of-supports equality in the proposition, this is equivalent to occurring in every trajectory in $\bigcup_{\pi_2 \in \Pi_2^+} \mathrm{Supp}_{\mathsf{S}}(\pi_2)$, which is equivalent to coverage one under every $\pi_2 \in \Pi_2^+$. Therefore $\bigcap_e \mathcal{H}_\phi^0(\mu_e) = \mathcal{H}_{\mathrm{rob}}$. Taking $\max_\preceq$ of both sides gives $\mathrm{Core}_{\mathrm{rob}}$. □

Proof of Theorem C.2(a)–(b) (upper bounds). Step 1: guarantee enough successful probes. Since $N_{\mathsf{S}} \sim \mathrm{Bin}(N, p_0)$, equation 3 gives $\Pr[N_{\mathsf{S}} < \frac{Np_0}{2}] \le e^{-\frac{Np_0}{8}} \le \frac{\rho}{2}$ under the theorem's second summand. Conditional on $N_{\mathsf{S}} = n \ge \frac{Np_0}{2}$, the occurrence indicators on successful trajectories are i.i.d. $\mathrm{Bernoulli}(c_\mu(u))$. Hoeffding therefore gives

$$\Pr[|\widehat{c}_N(u) - c_\mu(u)| > t \mid N_{\mathsf{S}} = n] \le 2e^{-2nt^2} \le 2e^{-Np_0t^2} \le \frac{\rho}{2}.$$

Step 2: combine the two deviations. A union bound proves (a).

Step 3: lift to M fixed candidates. Apply part (a) with tolerance $\frac{\gamma}{2}$ and failure probability $\frac{\rho}{M}$ to each candidate, then union-bound over candidates. Every estimate is then on the same side of the threshold as its population coverage, proving (b). □

Proof of Theorem C.2(c) (lower bound). Step 1: construct two threshold-separated laws. Fix $a \in (0, \frac{1}{2}]$, assume $1 - \delta \in [a, 1 - a]$ and $0 < \gamma \le \frac{a}{2}$, and put $c_- = 1 - \delta - \gamma$, $c_+ = 1 - \delta + \gamma$. Under single-probe laws P, Q, let a trajectory be unsuccessful with probability $1 - p_0$ and otherwise let $\mathbf{1}[u \preceq \phi(\tau)] \sim \mathrm{Bernoulli}(c_\pm)$. These laws require opposite membership decisions.

Step 2: bound their information distance. Only the successful branch differs, so

$$\mathrm{KL}(P\|Q) = p_0\,\mathrm{KL}\big(\mathrm{Bern}(c_-)\|\mathrm{Bern}(c_+)\big) \le p_0\,\frac{(c_- - c_+)^2}{c_+(1-c_+)} = \frac{4p_0\gamma^2}{(1-\delta+\gamma)(\delta-\gamma)},$$

using equation 5. The interior conditions give $c_+(1-c_+) \ge \kappa_a := \frac{a}{2}(1-\frac{a}{2})$, hence $\mathrm{KL}(P\|Q) \le \frac{4p_0\gamma^2}{\kappa_a}$.

Step 3: reduce estimation to testing. The tools in equation 6 yield $\mathrm{TV}(P^{\otimes N}, Q^{\otimes N}) \le \sqrt{\frac{2Np_0\gamma^2}{\kappa_a}}$. By equation 7, error below $\frac{1}{3}$ requires $\mathrm{TV} \ge \frac{1}{3}$. Therefore $N \ge \frac{\kappa_a}{18p_0\gamma^2} = \Omega_a(\frac{1}{p_0\gamma^2})$. □

E Monitoring Certificates and Windowed Erosion

E.1 A certificate with estimated success mass

A small reference success probability amplifies drift, so replacing p_0 by an upward-biased estimate can invalidate a certificate. The following construction budgets success-mass and conditional-coverage uncertainty together.

Proposition E.1 (Finite-sample survival certificate)**.** *Fix a candidate u, reference law μ_0 with $p_0 = \mu_0(\mathsf{S}) > 0$, and $N \ge 1$ independent reference rollouts. Let n be their number of successes, $\widehat{p} = \frac{n}{N}$,*

and $\widehat{c}$ their conditional coverage estimate. Choose $\rho_p, \rho_c > 0$ with $\rho_p + \rho_c < 1$ and define

$$\underline{p} = \max\left\{0, \widehat{p} - \sqrt{\frac{\log(\frac{1}{\rho_p})}{2N}}\right\}, \qquad \underline{c} = \begin{cases} \max\left\{0, \widehat{c} - \sqrt{\frac{\log(\frac{2}{\rho_c})}{2n}}\right\}, & n > 0, \\ 0, & n = 0. \end{cases}$$

Let $\mathrm{TV}(\mu_e, \mu_0) \le \varepsilon_e$ be a valid drift bound on a common trajectory space with fixed success event and abstraction. With probability at least $1 - \rho_p - \rho_c$, simultaneously at every episode with such a bound,

$$\underline{p} > 0, \qquad \underline{c} - \frac{\varepsilon_e}{\underline{p}} \ge 1 - \bar{\delta} \quad \Longrightarrow \quad \mu_e(\mathsf{S}) > 0, \quad u \in \mathcal{H}_\phi^{\bar{\delta}}(\mu_e), \qquad \bar{\delta} < 1.$$

If $\underline{p} = 0$, the rule issues no certificate.

Proof. **Step 1: control the two estimates.** One-sided Hoeffding gives $\underline{p} \le p_0$ except on an event of probability ρ_p. Conditional on $n > 0$, the successful-rollout indicators are independent Bernoulli variables with mean $c_{\mu_0}(u)$; Hoeffding gives $\underline{c} \le c_{\mu_0}(u)$ except with probability at most ρ_c. For $n = 0$ this inequality is automatic. A union bound controls both estimates, without assuming that they are independent.

Step 2: verify the conditioning lemma's domain. On this event, whenever the rule certifies,

$$\varepsilon_e \le \underline{p}(\underline{c} - 1 + \bar{\delta}) < \underline{p}\,\underline{c} \le p_0 c_{\mu_0}(u).$$

Thus Lemma 3.2 applies and gives $c_{\mu_e}(u) \ge c_{\mu_0}(u) - \frac{\varepsilon_e}{p_0} \ge \underline{c} - \frac{\varepsilon_e}{\underline{p}} \ge 1 - \bar{\delta}$. The same estimation event works for every episode whose drift bound is valid. □

For M candidates fixed before evaluation, allocate ρ_c across the M conditional-coverage bounds; the success-mass bound is shared. Data-dependent candidate discovery requires separate evaluation data or a uniform bound over its search class. An estimated drift bound requires its own confidence allocation; a union bound then adds its failure probability.

E.2 Windowed erosion and forward prediction

Write $[x]_+ := \max\{0, x\}$. The experimental erosion statistic is a positive coverage drop across a window. Its relation to cumulative directional progress is explicit.

Proposition E.2 (Windowed erosion)**.** *Let $\theta_{j+1} = \theta_j + \eta d_j$, $\|d_j\|_2 \le G$, and let c have L_c-Lipschitz gradient on each update segment for $q \le j < t$, where $q = t - w$. Put $a_j = -\langle \nabla c(\theta_j), d_j \rangle$ and $B_w = \frac{L_c G^2}{2} \eta^2 w$. Then*

$$\left| c(\theta_q) - c(\theta_t) - \eta \sum_{j=q}^{t-1} a_j \right| \le B_w, \qquad \left| [c(\theta_q) - c(\theta_t)]_+ - \left[\eta \sum_{j=q}^{t-1} a_j \right]_+ \right| \le B_w.$$

If this window has $\frac{1}{w}\sum_{j=q}^{t-1} a_j \ge \lambda > 0$ and $\eta \le \frac{\lambda}{L_c G^2}$, its erosion is at least $\frac{\lambda}{2}\eta w$. For monitored values C_q, C_t satisfying $|C_j - c(\theta_j)| \le r_j$ at the two endpoints,

$$|[C_q - C_t]_+ - [c(\theta_q) - c(\theta_t)]_+| \le r_q + r_t.$$

In particular, on these events the observed score is at least $\frac{\lambda}{2}\eta w - r_q - r_t$ under the window-progress premise.

Proof. **Step 1: isolate each update's remainder.** The two-sided smoothness bound gives $c(\theta_{j+1}) = c(\theta_j) - \eta a_j + R_j$ with $|R_j| \le \frac{L_c}{2}\eta^2 G^2$. Telescoping from q to t proves the first inequality.

Step 2: pass to the monitored statistic. The positive-part map is 1-Lipschitz. Applying it to the first inequality proves the second; applying it to the two endpoint errors proves the observation bound. Window progress and the step-size restriction imply $\eta \sum a_j - B_w \ge \lambda\eta w - \frac{\lambda}{2}\eta w$, which proves the lower bound. □

For confidence intervals $[L_j, U_j]$ containing $c(\theta_j)$, one can take $C_j = L_j$ and $r_j = U_j - L_j$. A lower bound alone, without control of its width, does not control changes in that bound. This applies to confidence intervals with their stated joint coverage; the Wilson intervals used in the experiments are nominal rather than a replacement for the finite-sample construction above.

The prefix condition (P2) starts at episode zero and does not imply positive progress on every later window. Moreover, an observed past drop by itself does not guarantee a future drop. If the progress condition instead holds on every future prefix starting at t until exit, the cumulative-erosion proof restarts at t: with current margin $m_t = c(\theta_t) - (1 - \bar{\delta}) > 0$, first exit occurs by $t + \left\lfloor \frac{2m_t}{\lambda\eta} \right\rfloor + 1$. Thus the empirical prediction tests examine whether past erosion is informative about subsequent performance; the deterministic exit bound additionally requires future progress.

E.3 Event-specific calibration and performance loss

The value decomposition uses only $\nu_{e,k}(A_k) - \mu_{e,k}(A_k)$, rather than the discrepancy on every trajectory event. If independent monitoring and deployment samples of sizes N_μ, N_ν give event frequencies $\widehat{a}_\mu, \widehat{a}_\nu$, then with probability at least $1 - \rho_\mu - \rho_\nu$,

$$|\nu(A_k) - \mu(A_k)| \le |\widehat{a}_\nu - \widehat{a}_\mu| + \sqrt{\frac{\log(\frac{2}{\rho_\nu})}{2N_\nu}} + \sqrt{\frac{\log(\frac{2}{\rho_\mu})}{2N_\mu}}.$$

This is two applications of Hoeffding and the triangle inequality. The right-hand side, clipped at one, supplies $\chi_{e,k}$. To use estimated calibration bounds in Theorem 3.3, budget their simultaneous failure probability in addition to the theorem's monitoring error. The execution and off-event reward bounds $\xi_{e,k}, b_{e,k}$ remain separate requirements.

Corollary E.1 (Coverage loss and value loss)**.** *Fix one policy, candidate, and reward scale $r \ge 0$ at two episodes q, t. Under Assumption 3.2, abbreviate their coverages by c_q, c_t, success masses by p_q, p_t, and radii by ζ_q, ζ_t. Then*

$$|(V_q - V_t) - r(p_q c_q - p_t c_t)| \le \zeta_q + \zeta_t.$$

In particular,

$$V_q - V_t \ge r p_q (c_q - c_t) - r|p_q - p_t| - \zeta_q - \zeta_t.$$

Proof. Subtract the two value approximations in Theorem 3.3 and apply the triangle inequality. Next use $p_q c_q - p_t c_t = p_q(c_q - c_t) + (p_q - p_t)c_t$ and $0 \le c_t \le 1$. □

A coverage drop therefore certifies a value drop when its compatible-return contribution exceeds the success-mass change and calibration errors. This identifies the additional measurements needed to turn an erosion signal into a policy-performance certificate.

F Experimental Details

F.1 Formal correspondence of the measured objects

The four suites use the same conditional-event quantity at different levels.

(i) **SRC (E1–E9):** μ_e is the fixed focal-probe trajectory law under peer checkpoint e, S is goal reachability, and $A = \mathsf{S} \cap \{u \preccurlyeq \phi(\tau)\}$. Thus the plotted population quantity is exactly $c_{\mu_e}(u) = \mu_e(A \mid \mathsf{S})$ from Definition 2.1.

(ii) **Continual control (E10):** draw X_0 uniformly from the fixed reference interval X and evaluate the next state under the two extreme disturbances. The candidate event A_{ctrl} occurs when both next states remain in X. Because the entire reference population is designated successful, its conditional coverage is exactly the normalized size of the estimated viability set: $\mu_t(A_{\mathrm{ctrl}} \mid \mathsf{S}) = \frac{\mathrm{Leb}(\widehat{K}_t)}{\mathrm{Leb}(X)} = \widehat{c}_t$. Emitting one symbol on A_{ctrl} makes this a length-one coverage monitor with $p_t = 1$.

(iii) **Cue-MNIST (E12):** draw index I uniformly from the fixed probe set and let μ_t be the empirical law of $Z_t = (y_I, f_t(x_I, q_{\mathrm{ref}}(y_I)), (f_t(x_I, c))_{c\in U_3})$, where $q_{\mathrm{ref}}(y) = (y+1) \bmod 10$, $U_3 = \{0, \dots, 9\}$, and $f_t(x, c)$ predicts from the image rendered with cue c. On this common coordinate space define the fixed events $\mathsf{S}_{\mathrm{cue}} = \{Z_{\mathrm{ref}} = Z_y\}$ and $A_{\mathrm{cue}} = \mathsf{S}_{\mathrm{cue}} \cap \{Z_c = Z_y \ \forall c \in U_3\}$. Then $\widehat{c}_t = \mu_t(A_{\mathrm{cue}} \mid \mathsf{S}_{\mathrm{cue}})$ exactly; the reported C_t is its Wilson lower bound.

(iv) **LBF (E11):** μ_t is again a fixed focal-probe trajectory law under a learned peer checkpoint, S is successful coordination, u is the discovered left-route sequence, and L_t lower-bounds $c_{\mu_t}(u)$. This peer-induced instantiation uses exploratory erosion in place of the registered absolute-risk predictor; the three-checkpoint window matches the already-frozen failure horizon.

Consequently E1–E9 test the structural laws directly, E10 and E12 test the downstream prediction-and-intervention use of the same coverage functional, and E11 traces the full peer-update–coverage–failure chain under learned partners. Detailed constructions appear in Appendices F.3 and F.5 to F.7.

Switch–Rock Corridor (SRC). The joint state is $z = (f, \rho_{\mathrm{r}})$ with focal position $f \in \{s_0, t_1, t_2, b_1, b_2, g\}$ and a binary rock flag $\rho_{\mathrm{r}} \in \{0, 1\}$ (12 states). Focal actions are $\{\mathrm{up}, \mathrm{down}, \mathrm{fwd}, \mathrm{wait}\}$; peer actions are $\{\mathrm{hold}, \mathrm{clear}, \mathrm{idle}\}$. Let $\mathrm{hold} = \mathbf{1}[a_2{=}\mathrm{hold}]$, $\mathrm{clear} = \mathbf{1}[a_2{=}\mathrm{clear}]$, and $\rho'_{\mathrm{r}} = \rho_{\mathrm{r}} \vee \mathrm{clear}$. Transitions are deterministic given (a_1, a_2):

- $f{=}s_0$: up$\to t_1$, down$\to b_1$, else stay.
- $f{=}t_1$: fwd $\wedge$ hold $\wedge$ $\rho_{\mathrm{r}}{=}0 \to t_2$, else stay. *(top gate; needs the rock intact)*
- $f{=}t_2$: fwd$\to g$.
- $f{=}b_1$: fwd $\wedge$ $\rho_{\mathrm{r}}{=}1 \to b_2$, else stay. *(bottom gate; needs the rock cleared)*
- $f{=}b_2$: fwd$\to g$; $f{=}g$: absorbing.

Because reaching t_2 requires $\rho_{\mathrm{r}}{=}0$ while reaching b_2 requires $\rho_{\mathrm{r}}{=}1$, and ρ_{r} is monotone, the two routes are mutually exclusive: a single episode yields a top *or* a bottom success, never both. The abstraction is $\phi(z, \cdot) = f$, so $\Sigma = \{s_0, t_1, t_2, b_1, b_2, g\}$; the robust core (Definition C.1) is generated by (s_0, g), i.e. only the start and goal lie on every success under every success-enabling peer.

Generalized domain SRC-L-N. The corridor length generalizes to L (top positions $t_1, \dots, t_L$, bottom positions $b_1, \dots, b_L$), and the single peer generalizes to N i.i.d. peers each clearing with probability α. The N peers enter the focal-induced kernel only through the scalar $q = 1 - (1-\alpha)^N$ (the probability that at least one peer clears), so the induced kernel remains exact and of size independent of N. This yields $\|P_q - P_{q'}\|_{1,\infty} = 2|q - q'|$ and, at $q{=}0$, a top core generated by the single frontier string $(s_0, t_1, \dots, t_L, g)$, whose high-coverage family contains $\Theta(L)$ length-1 patterns. The base SRC is recovered at $L{=}2, N{=}1$, and the reported $\frac{1}{N}$ law follows analytically from this scalar coupling.

Probe policy and core estimation. The base studies use $H = 12$ labels (at most 11 transitions); scaling uses $H = 3L + 4$. Unless stated otherwise the focal *probe* splits equally at s_0 and advances with probability 0.9 elsewhere; this fixed measurement policy makes observed drift attributable to the peer (Section 2.1). Coverage of a length-1 pattern σ is the fraction of successful probe rollouts whose focal-position string contains σ; the estimated high-coverage family is $\{\sigma : c \geq 1 - \delta\}$, with length-2 patterns evaluated analogously for Proposition C.2. This is the bounded-length evaluator for the hereditary family $\mathcal{H}^\delta$ generated by the core (Lemma 2.1), and Theorem C.2 gives its sample complexity. Estimates use $N \geq 5{,}000$ rollouts (E1: 4×10^4; E2: 4×10^4; E3: 7×10^3 per measurement; E4: 5×10^3).

E4 additionally computes population coverage *exactly*, conditional on each realized peer-parameter path. Let $x_t(z, h)$ be the probability of SRC state z after t transitions with hit bit $h \in \{0, 1\}$ indicating whether t_1 has appeared as a source-state label on a completed decision step. The update marks the state being departed, matching $\phi(z, a_1) = f(z)$. Starting from $x_0((s_0, 0), 0) = 1$, the forward recursion is

$$x_{t+1}(z', h') = \sum_{z, h, a_1, a_2} x_t(z, h)\pi_1(a_1 \mid z)\pi_2^e(a_2 \mid z)P(z' \mid z, a_1, a_2)\,\mathbf{1}[h' = h \vee \mathbf{1}[f(z){=}t_1]]\,.$$

After $H-1$ transitions,

$$c_{\mu_e}(t_1) = \frac{\sum_{z:f(z)=g} x_{H-1}(z,1)}{\sum_{z:f(z)=g} \sum_{h=0}^{1} x_{H-1}(z,h)}.$$

Population coverage is evaluated by this exact finite-state recursion.

Peer families. E1–E2 use $\pi_2^\alpha = \alpha\cdot\text{clear} + (1-\alpha)\cdot\text{hold}$ at every state, so α is a scalar drift knob and $\|P_\alpha - P_0\|_{1,\infty} = 2\alpha$. E3 trains two independent tabular Q-learners (ε=0.2, step 0.2, γ=1, terminal reward 1, focal step penalty 0.02, *no* private peer reward) for 8,000 episodes; the peer's convention is thus determined only by symmetry breaking. E4 uses a softmax policy-gradient peer (REINFORCE with a mean baseline, batch 192) initialized to hold, with step sizes $\eta \in \{0.25, 0.5, 1, 2, 4\}$. Finite horizon, bounded return, and bounded softmax scores give a deterministic bound on each batch gradient, hence the per-update condition $\sup_s \mathrm{TV}(\pi_2^e, \pi_2^{e-1}) \le c\eta$ (Remark 3.2). Population exit is evaluated exactly after every update; the finite-rollout exit is checked every 15 updates using an independent evaluation RNG, so evaluation does not alter the subsequent training path. E5 uses a hold-probability β peer that never clears (both routes' membership therefore unchanged) and a fixed option chain calibrated at β=1. E6 (single-agent control) fixes the peer to hold and Q-learns only the focal agent. The co-adaptation experiments (E3, E4) and all scaling experiments average over 10 seeds.

Experimental suites.

Suite	Exp.	Evidence
Structural degradation and lifetime	E1–E6	Exact membership discontinuity, zero stability-bound violations across 61 drift levels, persistent robust structure under convention churn, $R^2 = 0.999986$ inverse-rate population lifetime, executability separation, and stable single-agent control.
Rate and sharpness	E7–E9	$N^\star p_0 \gamma^2$ stays within $1.21\times$ across 15 cells; the trajectory-law ratio reaches 0.996, and the kernel-family limit reaches $\frac{255}{256} = 0.9961$.
Registered downstream confirmation	E10, E12	Continual-control AP advantage 0.237 and reward advantage 0.537; cue-MNIST AP and replay-reward advantages 0.039 and 0.0117.
Exploratory peer-induced analysis	E11	Post-hoc learned-peer LBF erosion AP advantage 0.228 across eight development pairings.

Reproducibility. The supplement contains `src.py` (base domain and exact population evaluator), `src_scaling.py` (generalized SRC-L-N with exact induced kernels), and fixed-seed experiment scripts. The command `python3 code/regenerate_all_figs.py` runs the SRC evaluations and regenerates the statistical figures, including E10–E12 from their supplied summary records. Task illustrations are supplied as fixed image assets. The supplement records per-seed measurements and the threshold conventions for the reported crossings.

F.2 GENERALIZED DOMAIN AND FIGURES

Figure 3 shows the environment and its SRC-L-N generalization. Experiments E1–E6 (Figures 4 to 8) test the single-instance predictions; the core-effect experiments (Figure 9) test soundness; the $L \times N$ scaling (Figures 11 to 13) and core-computation cost (Figure 10) test the scaling and complexity laws.

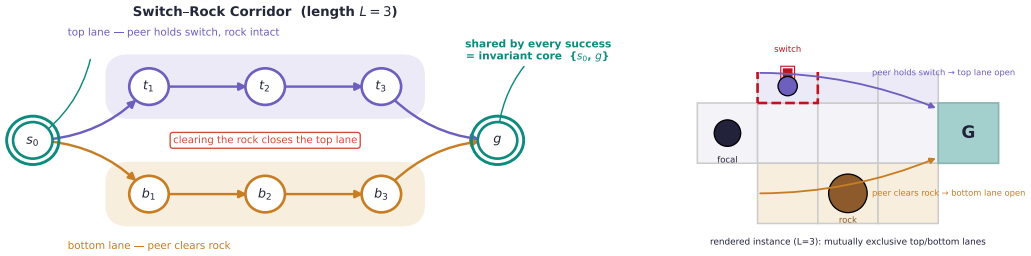

Figure 3: **The Switch–Rock Corridor.** *Left:* transition-structure schematic—two mutually exclusive routes (top needs the switch held and rock intact; bottom needs the rock cleared) sharing only s_0 and g; clearing the rock disables the top gate. *Right:* a rendered gridworld instance (L=3).

E1: impossibility and repair. Sweeping the peer's clear-probability α from 0, the support mass of t_1 decays continuously, but coverage-1 membership of t_1 (equivalently, whether t_1 is generated by the δ=0 core) is lost at the first $\alpha > 0$ tested ($\alpha = 2 \times 10^{-4}$), exactly as Proposition 3.2 predicts, while δ-families lose it only once the drift crosses δ-ordered thresholds ($\alpha \approx 2.3 \times 10^{-3}, 1.3 \times 10^{-2}, 6.0 \times 10^{-2}$ for $\delta = 0.02, 0.1, 0.3$). The robust core (s_0, g) is generated throughout.

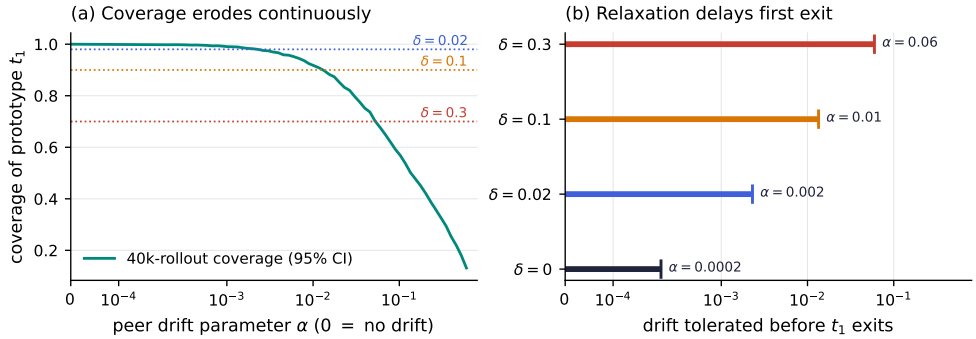

Figure 4: **E1.** *(a)* Prototype mass decays continuously in α; the band is an approximate 95% conditional-binomial interval from 40,000 rollouts. *(b)* Directly labeled first-exit drift: coverage-1 membership is lost at any $\alpha > 0$, whereas δ-families persist to δ-ordered thresholds (Proposition 3.2 and Theorem 3.1).

E2: the stability bound holds. Across 61 drift levels the empirical coverage of t_1 never falls below the Theorem 3.1 guarantee $1 - \delta - \frac{\varepsilon_e}{p_0}$ (0 violations), with median slack 0.132. The slack is smallest at α=0, where guarantee and coverage coincide at 1.

E3: robust core is invariant while conventions churn. Two independent Q-learners co-adapt; the peer has no private reward, so convention choice is pure symmetry breaking. Over 8,000 episodes the preferred convention flips ≈ 126 times on average (10 seeds), the prototypes t_1, b_1 oscillate, yet the robust core (s_0, g) is generated in 100% of measurements across all seeds (Proposition C.4). The single-agent control shows the estimated family equal to the truth with 0 changes (Section 2.1).

E4: exact population survival tracks $\frac{1}{\eta}$. A softmax policy-gradient peer with step size η erodes t_1. Using the exact recursion above after every update, the mean population first-exit updates are

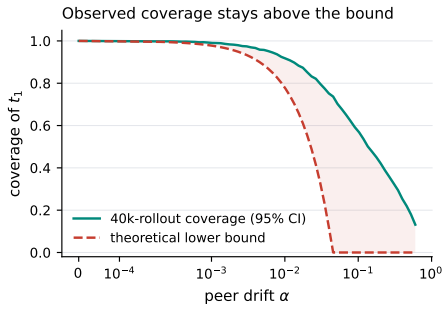

Figure 5: **E2.** Empirical coverage (approximate 95% conditional-binomial band, 40,000 rollouts) stays above the Theorem 3.1 guarantee $1 - \delta - \frac{\varepsilon_e}{p_0}$ at all 61 drift levels (0 violations); the guarantee is clipped at 0 where it goes vacuous.

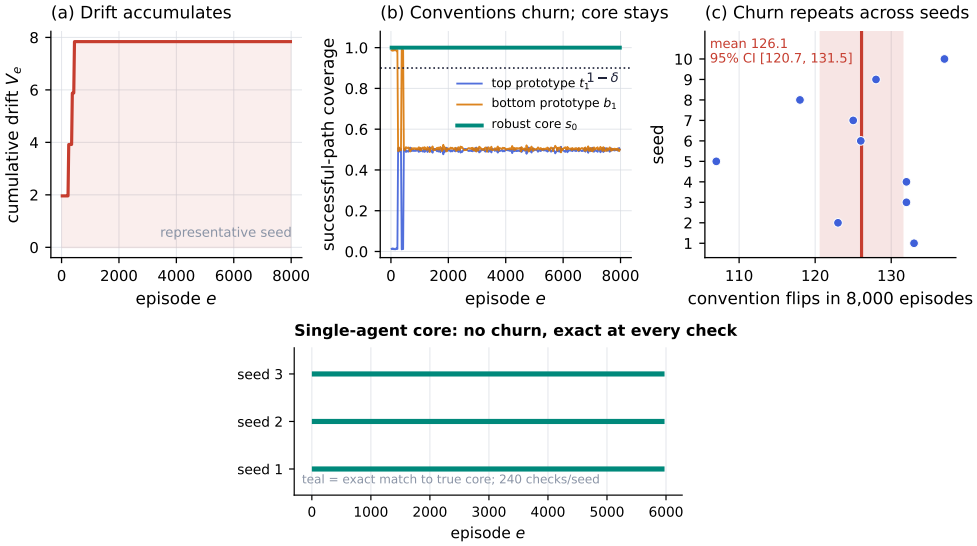

Figure 6: **E3.** *Top:* a representative seed shows cumulative kernel drift (a) and conventional-prototype churn while the robust core stays at mass 1 (b); seed-level flip counts with a 95% CI show that churn repeats across all 10 seeds (c). *Bottom:* in the single-agent control, every one of 240 checks in each of three seeds exactly matches the true core.

$127.9, 64.9, 33.0, 17.1, 9.6$ for $\eta = 0.25, 0.5, 1, 2, 4$ (10 seeds). They are linear in $\frac{1}{\eta}$ with $R^2 = 0.999986$; the products ηE_{exit} range from 32.0 to 38.4, with the largest-step deviation attributable in part to integer-time resolution. This is an exact population statistic evaluated along each stochastic training path.

At the scheduled 15-update evaluations, the 5,000-rollout coverage estimate has RMSE 0.0068 against exact coverage, and its first-exit update matches the exact grid first exit in all 50 seed–step-size runs. Its mean absolute exit-time delay relative to the update-wise population exit is 8.9 updates, exactly equal here to the delay from checking only on the 15-update grid: there are no additional Monte-Carlo crossing errors. Thus E4 isolates inverse-step-size learning from finite-evaluation resolution and realizes the population lifetime law in the smooth SRC peer family.

E5: membership vs. executability. Lowering the peer's hold-probability β, the prototypes t_1, t_2 remain coverage-1 in the δ=0 family for all $\beta > 0$ (membership intact), yet a fixed option chain calibrated at β=1 drops from success probability 1.0 to 0.10 as $\beta \to 0.05$ (Proposition 3.4).

Effect of the core. Figure 9*(a)* measures the fraction of the length-H string space surviving the soundness constraint; since the frontier string has length $\Theta(L)$ it shrinks exponentially in L, never

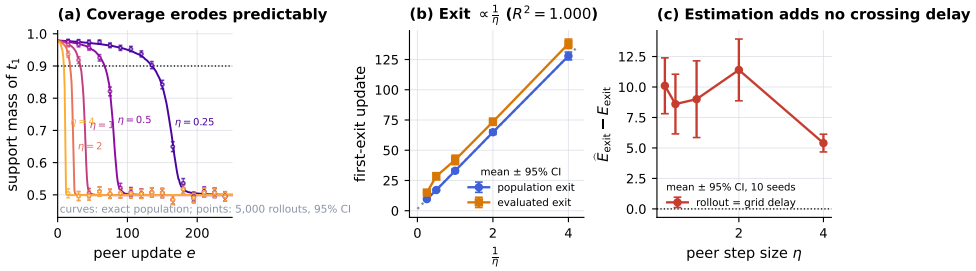

Figure 7: **E4.** *(a)* Exact population coverage after every update (curves) and 5,000-rollout estimates with approximate 95% conditional-binomial intervals every 15 updates (points), first seed. *(b)* Population and finite-evaluation first exits against $\frac{1}{\eta}$ (mean and 95% CI across 10 training seeds); the population relation has $R^2 = 0.999986$. *(c)* The rollout-estimated and exact-grid delays coincide in all runs, so one curve displays their common mean and 95% CI.

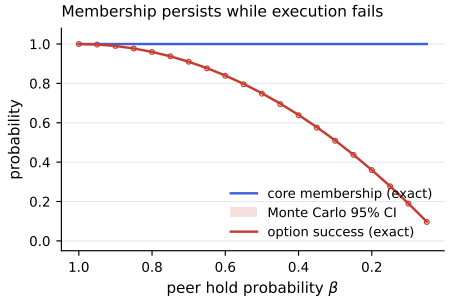

Figure 8: **E5.** Exact membership of t_1, t_2 is retained for all $\beta > 0$ while a fixed option chain's exact success collapses—executability drift without membership drift. Open points and the 95% Monte-Carlo band use 40,000 trials.

excluding an optimum under the full-support probe (Proposition C.1). Figure 9*(b)* plots that frontier-string length, and Figure 9*(c)* contrasts a core-guided option chain (100% success) with a core-blind random-admissible controller (∼17–28%).

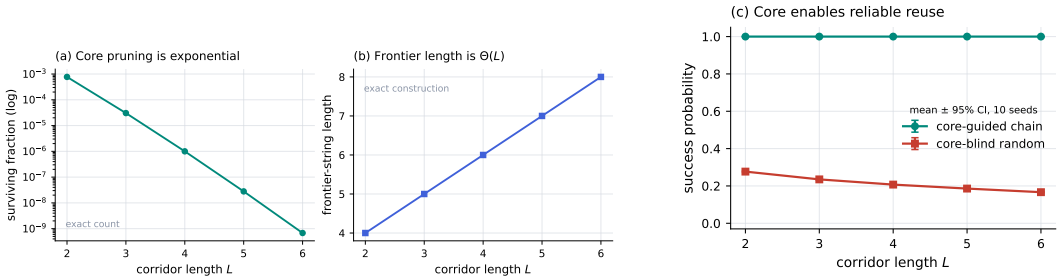

Figure 9: **Effect of the core.** *(a)* The exact soundness count leaves an exponentially shrinking fraction of trajectory strings as L grows. *(b)* The frontier here is a *single* exact string, so $|\mathrm{Core}^0_\phi(\mu)| = 1$; its *length* grows as $\Theta(L)$. *(c)* Core-guided (100%) vs. core-blind (∼17–28%), with means and 95% CIs across 10 seeds.

Core-computation cost. The exact multi-sequence LCS cost (DP cells) grows as $O(H^k)$ in the horizon for each number of successes k, with the fitted exponent tracking k (Appendix C.3).

Scaling: the effect grows with N and L. Define core fragility as $\frac{10^3}{E_{\text{exit}}}$, where E_{exit} is the first update at which no length-1 top-route prototype remains in the estimated δ=0.05 family. Both

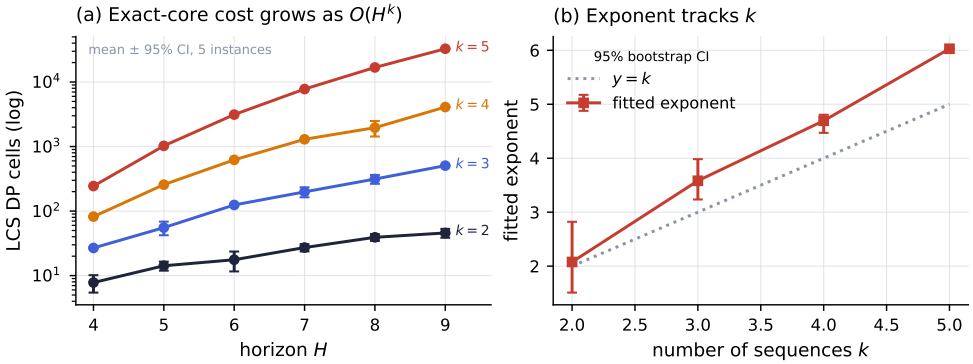

Figure 10: **Core-computation cost.** *(a)* Exact multi-string LCS cost (DP cells) grows as $O(H^k)$ (log scale); bars are 95% CIs over five random instances per cell. *(b)* The fitted exponent tracks k, with 95% instance-bootstrap intervals, verifying the product-index DP profile of Appendix C.3 for the support core (δ=0).

fragility and aggregate boundary drift increase monotonically with L and N (Figure 11). The slices in Figure 12 show survival $\propto \frac{1}{N}$ ($R^2 = 0.999$), as predicted by the scalar coupling $q = 1-(1-\alpha)^N$ of SRC-L-N. They also show that full-core survival decreases with L.

The bound is never violated at any tested (L, N). Figure 13 reports the minimum slack over drift levels where the guarantee is non-vacuous. It ranges from 0.009 at $(L, N) = (2, 1)$ to 0.289 at $(8, 8)$. The guarantee becomes more conservative as drift accumulates and becomes vacuous once $\varepsilon_e \geq (1 - \delta)p_0$.

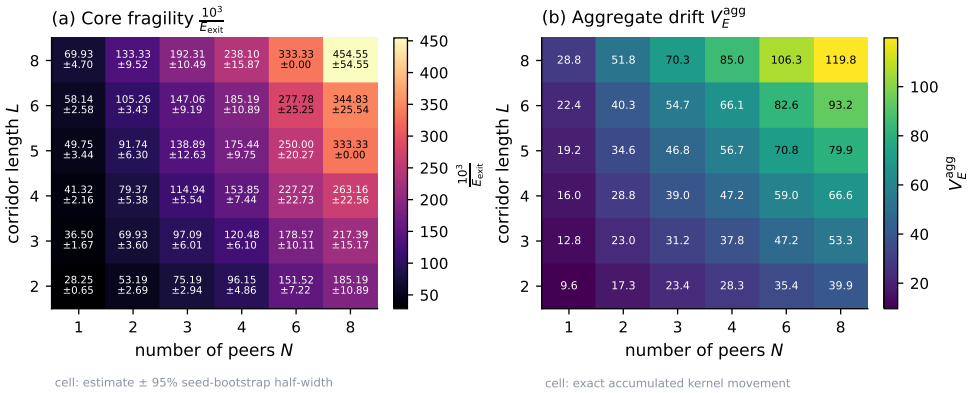

Figure 11: **Scaling grid ($L \times N$).** *(a)* Core fragility $\frac{10^3}{E_{\text{exit}}}$ grows in both N and L; cells report the estimate and 95% seed-bootstrap half-width over 10 seeds. *(b)* Exact aggregate boundary drift grows in both.

F.3 Direct quantitative validations of theorem-level predictions (E7–E9)

E7 tests the predicted sampling rate, while E8 attains the factor one at the trajectory-law level and E9 shows that the kernel-parameterized coefficient is unimprovable uniformly over horizons.

E7: the $\frac{1}{p_0\gamma^2}$ rate. The plug-in estimator uses only two counts: the number of successes and the number of successes containing u. Under i.i.d. rollouts, $N_{\mathsf{S}} \sim \text{Bin}(N, p_0)$, and the occurrence count conditional on N_{S} follows $\text{Bin}(N_{\mathsf{S}}, c_\mu(u))$. SRC rollouts verify this reduction (means 0.7128 vs 0.7132, standard deviations 0.0092 vs 0.0094 at N=4000).

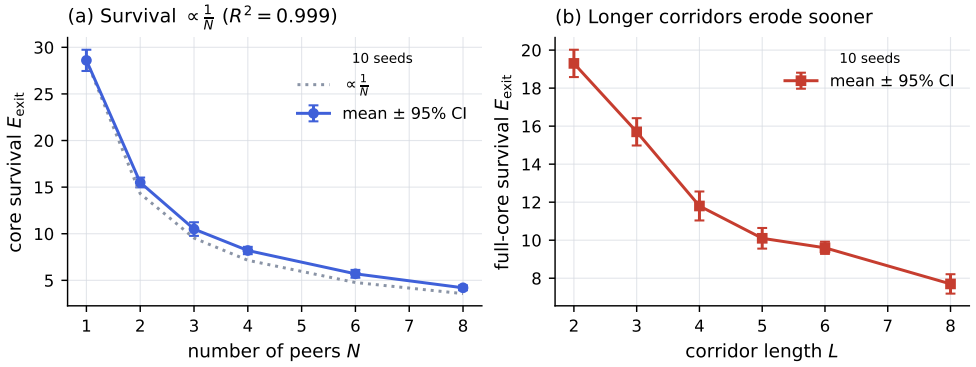

Figure 12: **Scaling slices (10 seeds).** Points and bars are means and 95% CIs. *(a)* $E_{\text{exit}} \propto \frac{1}{N}$ ($R^2 = 0.999$). *(b)* Full-core survival decreases as the corridor lengthens.

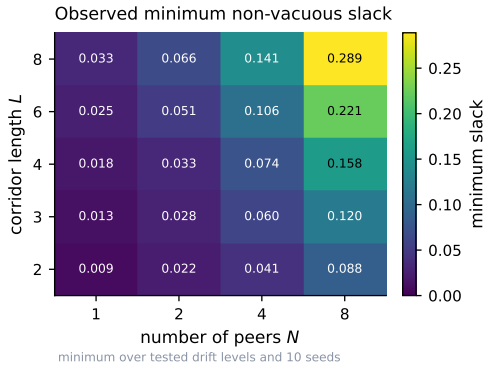

Figure 13: **Bound validity across the grid.** Observed minimum slack (empirical coverage minus the Theorem 3.1 guarantee) over tested non-vacuous drift levels and 10 seeds. There are no violations; slack grows with L and N. Every displayed cell has at least one tested non-vacuous drift level.

We then sweep $p_0 \in [0.05, 0.8]$ and $\gamma \in [0.03, 0.08]$. Varying the probe alone moves p_0 by under 10%, so the direct sweep is needed to exercise the $\frac{1}{p_0}$ factor. For each cell, $N^\star$ is the smallest sample size at which the plug-in rule misclassifies with probability at most 5% over 4000 repetitions. Across 15 cells, $N^\star$ spans 126 to $13{,}934$, while $N^\star p_0 \gamma^2$ remains in $[0.574, 0.692]$, a spread of $1.21\times$ (Figure 14). Direct simulation of the exact sampling law therefore confirms the predicted rate and its stable constant over the tested grid.

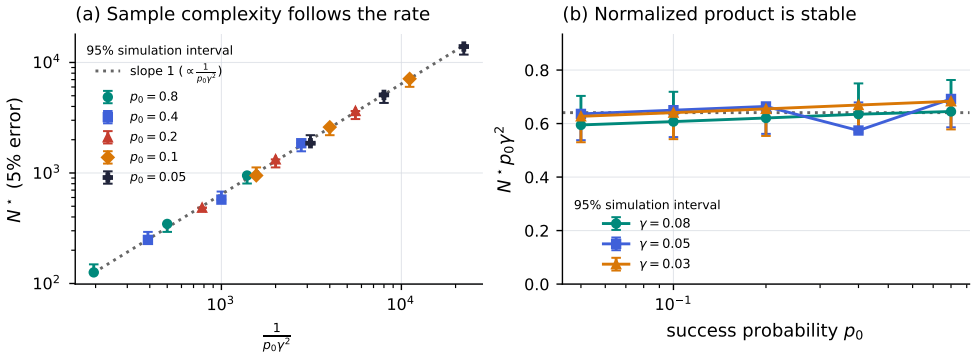

Figure 14: **E7.** *(a)* $N^\star$ against $\frac{1}{p_0\gamma^2}$ on log axes, with a slope-one reference. *(b)* $N^\star p_0 \gamma^2$ across all 15 cells, constant to within $1.21\times$. Bars bracket the 95% Wilson uncertainty in the simulated 5% error crossing using $4{,}000$ repetitions per tested N.

E8: worst-case tightness of Theorem 3.1. Tightness is measured at the *trajectory-law* level using the actual $\mathrm{TV}(\mu_\alpha, \mu_0)$ over the abstract outcome classes rather than the kernel bound of Lemma 3.1. SRC moves success mass precisely from "u occurs" to "u absent" while leaving total success mass nearly fixed, which is the equality case. The realized coverage drop never exceeds the guaranteed maximum $\frac{\mathrm{TV}}{p_0}$ at any of the 14 drift levels, and the ratio reaches 0.996 at the smallest drift, decaying to ≈ 0.66 as the drift grows (Figure 15). The factor one is attained in the vanishing-drift limit.

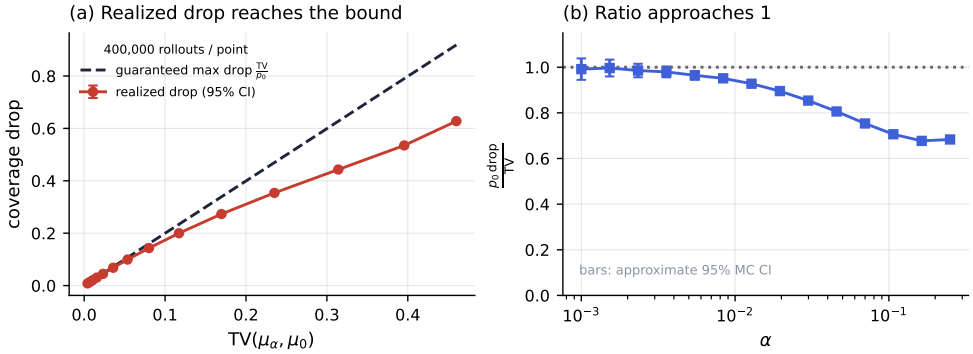

Figure 15: **E8.** *(a)* Realized coverage drop against the guaranteed maximum $\frac{\mathrm{TV}}{p_0}$. *(b)* Their ratio approaches 1 as drift vanishes and never exceeds it. Bars are approximate 95% Monte-Carlo intervals from 400,000 rollouts per point.

E9: tightness in the kernel parameterization. E8 establishes tightness at the trajectory-law level, where ε_e is the actual $\mathrm{TV}(\mu_e, \mu_0)$. The kernel-parameterized guarantee, including the telescoping in Lemma 3.1, is tested by an explicit depth-indexed family. In the *cascade* $\mathrm{Casc}(n)$ the focal agent has a single action and positions $t_0, \ldots, t_n$ (top), $b_1, \ldots, b_n$ (bypass), and g. At $t_0, \ldots, t_{n-1}$ the peer either holds (continue on top) or, with probability α, diverts to the bypass; t_n then enters g. Both routes reach g, so $p_0 = 1$, and diverted mass lands on successes lacking $u = (t_1, \ldots, t_n)$. A top success has labels $(t_0, \ldots, t_n, g_\Sigma)$, hence $H = n + 2$, $\|P_\alpha - P_0\|_{1,\infty} = 2\alpha$, and the theorem gives $\varepsilon_\alpha = (n+1)\alpha$. The realized drop is exactly $1 - (1-\alpha)^n$, so Bernoulli's inequality keeps it below the guarantee and

$$\lim_{\alpha \downarrow 0} \frac{1 - (1-\alpha)^n}{(n+1)\alpha} = \frac{n}{n+1}.$$

For fixed $n = 11$ this limit is $\frac{11}{12}$; across the family it tends to one as $n \to \infty$, reaching $\frac{255}{256} = 0.9961$ at $n = 255$ (Figure 16). Thus no horizon-uniform coefficient smaller than one can replace the kernel factor in Theorem 3.1.

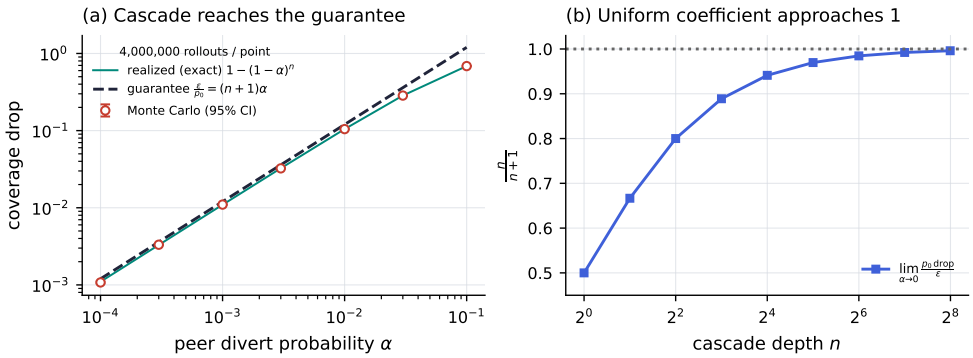

Figure 16: **E9.** *(a)* For $n = 11$, realized coverage drop (exact and Monte Carlo with 95% binomial intervals from 4,000,000 rollouts) against the corrected kernel guarantee $\frac{\varepsilon}{p_0} = (n+1)\alpha$. *(b)* The infinitesimal-drift ratio is $\frac{n}{n+1}$ and approaches one with cascade depth, proving that the horizon-uniform coefficient is sharp.

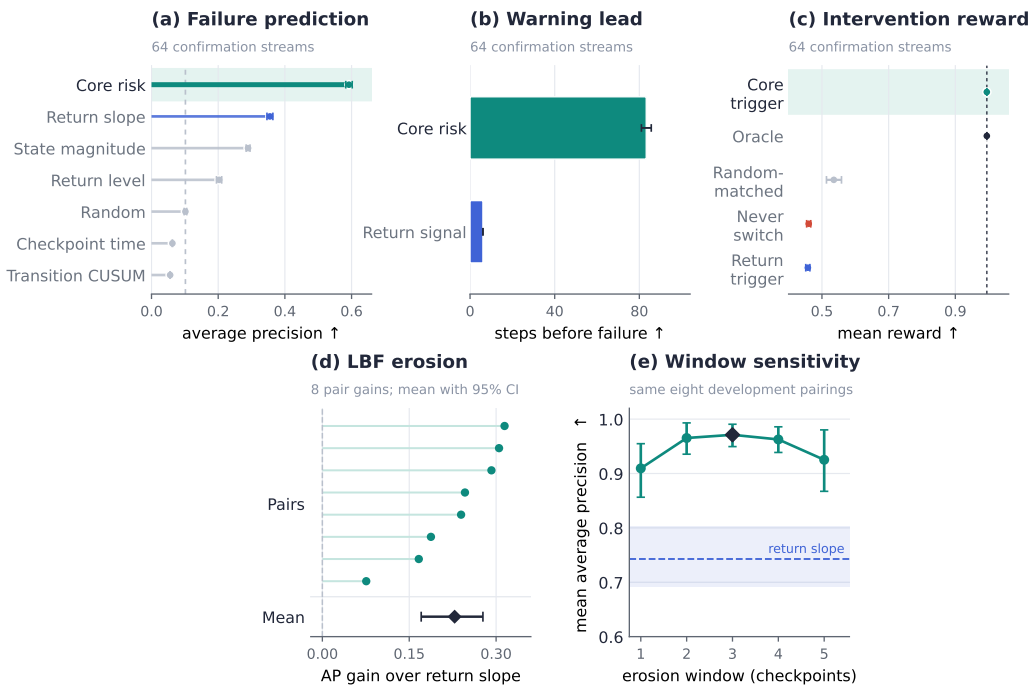

Figure 17: **Extended prediction and intervention results.** *(a)–(c), E10 confirmation:* Failure prediction against all baselines, warning lead, and intervention reward across 64 complete streams. Dashed vertical lines mark random AP in (a) and oracle reward in (c). *(d)–(e), E11 exploratory reanalysis:* Gains over return slope and erosion-window sensitivity on eight development pairings. Each unlabelled row in (d) is one pairing, ordered by gain; the diamond is the mean with its 95% interval. Whiskers and bands resample complete streams or pairings.

F.4 Recorded configurations and comparison inputs

The tables collect the frozen configuration values used for E10–E12. Each independent unit is a complete stream or peer/focal pairing; checkpoint trajectories remain intact under resampling.

E10: continual control	
Plant and reference set	$a = 1.08$, $R = 0.75$, state limit 2.5, action limit 0.45, process-noise standard deviation 0.006.
Controllers and probe	Nominal gain 0.50, backup gain 2.00, probing amplitude 0.035, robust disturbance radius $w_{\max} = 0.080$.
Costs	Action-cost coefficient 0.015; backup cost 0.010 per step.
Gain schedule	High gain 1.0; low gain in $[0.045, 0.075]$; ramp lengths 300–500, low plateaus 650–950, stable phases 650–1050 steps; initial stable phase 700–1100.
RLS	Initial gain estimate 1.0, initial covariance 8.0, forgetting factor 0.985, estimate clipped to $[0, 1.25]$.
Monitoring	Every 20 steps; reward window 100; healthy/failure reward thresholds $0.90/0.70$; prospective horizon 240 steps.
Switching hysteresis	Coverage enter/exit $0.95/0.985$; return enter/exit $0.82/0.94$; minimum backup duration 80 steps.
Inference	64 confirmation streams; 10,000 paired bootstrap resamples, seed 884219.
E11: LBF coordination fork	
Peer parameterization	Feed-forward logistic route policy; initial route bias -4.6, cue weight 1.5, style logit -1.0; initialization standard deviation 0.04.

Peer updates	Sampled REINFORCE with batch size 512, route learning rate 0.30, style learning rate 0.35, 32 updates per checkpoint; initial anchoring for 16 updates.
Coordination mechanism	Neutral wait-then-follow probe; probe and robust-policy success parameters 0.995 and 0.970; mismatch-recovery probability 0.03; style latency cost 0.015.
Discovery	Four separate pairings, 2000 episodes each; minimum candidate coverage 0.95, maximum length 4, peer-action symbols.
Evaluation	Eight development pairings, 72 checkpoints; separate batches of 500 probe, 500 focal-monitoring, and 500 performance episodes per checkpoint.
Thresholds	Nominal confidence 0.95; coverage membership/recovery $0.90/0.94$; performance healthy/failure $0.90/0.70$; prediction horizon 3 checkpoints.
Inference	2000 complete-pairing bootstrap resamples, seed 20260831; erosion window chosen post hoc on these eight pairings.
E12: continual cue-MNIST	
Stream and splits	300 checkpoints, four events; 55,000 training examples; disjoint core/shift/performance probes of $2000/500/1000$ examples.
Network and optimizer	$784 \to 64$ tanh$\to 10$ logits; 50,890 parameters. Adam $(0.9, 0.999)$, $\epsilon = 10^{-8}$; weight-matrix L2 regularization.
Pretraining	3500 supervised updates; batches of 64 base images with two cue views each; learning rate 0.0015, L2 coefficient 10^{-5}.
Continual updates	299 intervals of 12 updates, batch size 128; learning rate 0.001, L2 coefficient 0.001; checkpoint zero is pretrained.
Cue and image parameters	10 spatial 3×3 cues, intensity 1.0; $q_{\mathrm{ref}}(y) = (y+1) \bmod 10$; reference/harmful contrast 0.30, nuisance contrast 0.27. Harmful cue-label mixture probability 0.98; benign cues uniform.
Nested cue sets	$U_1 = \{0,1\}$, $U_2 = \{0,\dots,4\}$, $U_3 = \{0,\dots,9\}$; the primary coverage requires correctness under all ten cues.
Monitoring	Three-checkpoint erosion; healthy/failure/recovery accuracy $0.92/0.88/0.92$; future horizon 7 checkpoints.
Replay	Coverage enter/exit $0.90/0.94$; replay fraction 0.50; intervention cost 0.002 per checkpoint.
Inference	64 confirmation streams; 10,000 paired bootstrap resamples, seed 20261211.

Table 3: Recorded E10–E12 configuration values.

Information used by the comparisons. In E10, the structural monitor combines observed transitions and an RLS estimate of the unknown actuator gain with the specified plant coefficient a, nominal feedback gain, reference interval, and disturbance set. The return monitors use reward history, the state-magnitude baseline uses the current state, and transition-surprise CUSUM uses transition residuals. The reported comparison therefore concerns these task-specific monitors with their stated inputs; it does not isolate the contribution of structural coverage from model knowledge. The operational monitor receives neither the true actuator gain nor future drift. The oracle substitutes the true gain in the same viability formula.

In E12, the structural monitor queries counterfactual cues on the core probe, cue association uses the separate shift-monitor split, and the accuracy trigger uses reference-cue accuracy on the same core probe; held-out performance is evaluated separately. Replay-trigger comparisons share the registered replay procedure and differ in intervention timing. These comparisons evaluate trigger choice within a fixed learning setup. Replay-based continual-learning algorithms such as CLEAR (Rolnick et al., 2019), Bayesian change-point inference (Adams & MacKay, 2007), and partner-population training such as fictitious co-play (Strouse et al., 2021) address complementary algorithmic questions. They are literature context, not evaluated baselines in these studies.

Which theoretical quantities are tested. SRC evaluates conditional coverage, trajectory drift, and first exits directly, including rare-success amplification. E10 instantiates a length-one coverage functional with success mass one; E12 uses a length-one cue-robust event conditional on reference

correctness, with success mass $\frac{|B_t|}{2000}$. Their stream experiments test prediction and intervention under exogenous drift. E11 traces peer-induced coverage erosion in a controlled coordination fork. Monitoring these fixed candidates uses event evaluation rather than recomputation of the maximal subsequence frontier.

Artifact scope. The supplied source runs the SRC studies and regenerates E10–E12 statistics figures from frozen aggregate and per-unit records. The cue-MNIST training and analysis source is also supplied, together with runnable configurations and source-verification checks. The compact E10 and E11 bundles contain protocols, configurations, and records; their original training drivers and large per-step arrays are not included. E12 implementation details are specified in Appendix F.7.

F.5 E10: REGISTERED SINGLE-AGENT CONTINUAL-CONTROL CONFIRMATION

E10 tests a downstream consequence of the focal continual-RL setting represented by Proposition 3.1: degradation of reusable structure should warn before performance fails, and acting on that warning should preserve performance (Figure 2(a)–(b); Figure 19, left). Each seed runs uninterrupted for $18{,}000$ steps:

$$x_{t+1} = ax_t + b_t u_t + w_t,$$

where actuator gain b_t undergoes five seed-randomized continuous high-to-low-to-high drift–recovery events. Physical state, recursive-least-squares (RLS) gain estimate, rolling monitors, controller state, and intervention state are never reset. The action is clipped; a nominal controller is inexpensive but loses stability at low gain, while a costed backup remains stable throughout the registered range. Each treatment receives the same drift schedule, process noise, and probing sequence within a seed but evolves its own unbroken state and RLS estimate.

On the monitored reference interval $X = [-R, R]$, clipping is inactive for the nominal controller even at maximal probing amplitude: the registered constants satisfy $k_{\mathrm{nom}} R + a_{\mathrm{probe}} = 0.410 < 0.45 = u_{\max}$. The disturbance radius also satisfies $0 \le w_{\max} = 0.080 < R = 0.75$. Action saturation therefore does not alter the nominal feedback law on X, and the viability set below is nonempty whenever its displayed radius is positive; it is the one-step viability core of the monitor's estimated linear model.

Structural quantity and prospective target. For the fixed reference interval $X = [-R, R]$ and disturbance set $W = [-w_{\max}, w_{\max}]$, define the estimated one-step viability core of the nominal feedback controller by

$$\widehat{K}_t := \left\{ x \in X : (a - \widehat{b}_t k_{\mathrm{nom}})x + w \in X \text{ for every } w \in W \right\}.$$

Its normalized coverage is

$$\widehat{c}_t := \frac{\mathrm{Leb}(\widehat{K}_t)}{\mathrm{Leb}(X)} = \begin{cases} 1, & a - \widehat{b}_t k_{\mathrm{nom}} = 0, \\ \min\left\{1, \frac{R - w_{\max}}{R\,|a - \widehat{b}_t k_{\mathrm{nom}}|}\right\}, & \text{otherwise.} \end{cases}$$

Thus $\widehat{c}_t$ is exactly the fraction of X that remains viable under every registered disturbance; operational core risk is $1 - \widehat{c}_t$. The construction below represents this quantity exactly as the coverage of a length-one candidate. The latent b_t appears only in an oracle diagnostic. At a checkpoint with healthy current rolling reward, the positive prospective label is a new downward crossing of the registered failure threshold within the next 240 steps. Existing failure plateaus and incomplete tail windows are censored, and recovery re-enters the risk set.

Exact correspondence to conditional-event coverage. Set $q_t = a - \widehat{b}_t k_{\mathrm{nom}}$, draw $X_0 \sim \mathrm{Unif}(X)$, and let μ_t be the law of

$$Z_t = (X_0, q_t X_0 - w_{\max}, q_t X_0 + w_{\max}).$$

All μ_t live on the same three-coordinate space. Take the success event to be that whole space, so $p_t = 1$, and define the fixed structural event $A_{\mathrm{ctrl}} := \{Z : Z_2, Z_3 \in X\}$. Because an affine image of the disturbance interval attains its extrema at $\pm w_{\max}$,

$$X_0 \in \widehat{K}_t \iff |q_t X_0| + w_{\max} \le R \iff Z_t \in A_{\mathrm{ctrl}}.$$

Therefore $\mu_t(A_{\text{ctrl}} \mid \mathsf{S}) = \Pr[X_0 \in \widehat{K}_t] = \frac{\text{Leb}(\widehat{K}_t)}{\text{Leb}(X)} = \widehat{c}_t$. Equivalently, form a one-step probe MDP with initial-state distribution μ_t. Its only action is a dummy action, its abstraction emits SAFE exactly on A_{ctrl} (and OTHER otherwise), and every terminal state is successful. Then $\widehat{c}_t$ is exactly the Definition 2.1 coverage of the length-one candidate (SAFE). E10 uses this exact coverage as a monitor and tests its downstream predictive and control consequence; it does not estimate a peer-policy drift budget.

Separation, inference, and interventions. Thirty-two full development streams (seeds 5000–5031) selected return slope as the strongest of six simple baselines. The plant, drift distribution, estimator, structural quantity, prediction target, thresholds, controllers, and intervention costs were then frozen before running 64 confirmation streams (seeds 0–63). The two registered co-primary contrasts are structural-risk minus return-slope average precision and certificate-trigger minus return-trigger whole-stream mean reward. Their 95% percentile-bootstrap intervals use 10,000 resamples of the 64 complete streams. Checkpoints, the 320 drift incidents, and time steps are not treated as independent replicates. Besides the return trigger, controls are never switching, an intervention-count-matched random cyclic shift of the certificate schedule, and a latent-gain oracle.

Confirmation results. Core risk has average precision 0.5922, versus 0.3551 for the frozen return-slope baseline; the paired advantage is 0.2370 (95% CI $[0.2233, 0.2509]$). Mean warning lead is 83.24 steps for core risk and 6.07 for the return signal. Core-triggered control has mean reward 0.99444, compared with 0.45778 for the return trigger, 0.46035 for never switching, 0.53558 for random-matched intervention, and 0.99447 for the oracle. The paired core-triggered advantages are respectively 0.53666 ($[0.53058, 0.54281]$), 0.53409 ($[0.52790, 0.54036]$), and 0.45886 ($[0.43605, 0.48086]$). Both registered co-primary lower endpoints exceed zero, so the pre-specified intersection–union rule passes. The core-triggered arm records no failure steps across the 64 streams.

Protocol and reproducibility. The experiment validates the downstream core mechanism in a genuinely continual stream: erosion of a monitored structural core anticipates failure, and acting on that erosion preserves performance. The imported compact artifact at `code/continual_control_toy_v1_confirmation_stats/` contains both frozen protocols and configurations, aggregate summaries, every per-stream statistic, integrity decisions, and a SHA-256 manifest; all files pass that manifest. The large per-step arrays remain on the originating cluster. Figure regeneration uses `code/plot_continual_control_confirmation.py`; the illustrative task image and its generation prompt are stored beside the figures.

F.6 E11: Exploratory peer-induced core erosion in Level-Based Foraging

E11 is an exploratory reanalysis of eight development pairings from a controlled Level-Based Foraging (LBF) coordination-fork experiment (Figure 2(c); Figure 19, right). The joint task and rewards remain stationary while a feed-forward logistic route peer is updated by sampled REINFORCE through two right-learning and left-recovery cycles. A fixed neutral focal probe waits for the peer's route and follows it, whereas a separately frozen old-left focal emits its original cue. Their monitoring and held-out performance rollouts are disjoint.

The independently discovered peer sequence is ENTER-LEFT-GATE, LEFT-READY, LOAD-LEFT. Four discovery pairings use 2,000 episodes each, and all discovery and baseline-coverage gates pass. Each learned-peer pairing contains 72 peer checkpoints, with 500 probe, 500 focal-monitoring, and 500 performance episodes per checkpoint. An additional style-learning block changes action and occupancy statistics without changing the route convention, providing a nuisance-drift control.

Registered result and exploratory reanalysis. Let L_t denote the Wilson lower confidence bound on candidate coverage. The registered V3 primary predictor was absolute structural risk, $1 - L_t$. It obtained mean AP 0.5260, below the one-checkpoint monitoring-return slope baseline at 0.7428; the registered paired contrast was -0.2168 with a 95% pairing-bootstrap interval $[-0.2884, -0.1520]$, so the registered prediction gate failed.

Here the correspondence to Definition 2.1 is direct: at peer checkpoint t, μ_t is the neutral focal's fixed-probe trajectory law, S is successful coordination, and u =

(ENTER-LEFT-GATE, LEFT-READY, LOAD-LEFT). The monitored population target is $c_{\mu_t}(u)$ and L_t is its finite-rollout lower confidence bound. Peer learning changes μ_t while the focal probe, abstraction, success event, and candidate remain fixed.

After examining the same eight pairings, the reanalysis replaced the registered absolute-risk predictor with active erosion. Its three-checkpoint window matches the already-frozen prospective failure horizon:

$$s_t^{(3)} = \max\{0, L_{t-3} - L_t\}\,.$$

The prospective label is a new downward crossing of old-focal held-out success below 0.70 within the next three checkpoints, restricted to currently eligible checkpoints. The erosion score measures a windowed coverage change; Proposition E.2 gives its directional-progress and estimation-error decomposition. It attains mean AP 0.9712, versus 0.7428 for return slope, and exceeds that baseline in all eight pairings. The paired advantage is 0.2284 (pairing-bootstrap 95% CI $[0.1707, 0.2773]$). Windows of one through five checkpoints give mean AP $0.9095, 0.9653, 0.9712, 0.9626, 0.9252$, respectively, all above return slope. The neighboring-window results are sensitivity diagnostics. The pairing-bootstrap intervals condition on the erosion predictor and do not account for the post-hoc change of predictor. Because erosion was evaluated on the same development pairings that informed this change, these results are exploratory and motivate confirmation on new pairings.

Protocol and reproducibility. The exploratory pattern contains the full downstream chain in a multi-agent environment: peer learning changes the induced focal problem, a fixed-probe trajectory core erodes, and old-partner performance later fails. The imported artifact at `code/lbf_v3_core_erosion_exploratory_stats/` contains the frozen V3 protocol and configuration, all eight raw trajectory summaries, the registered result, the post-hoc aggregate and per-pairing reanalysis statistics, and SHA-256 checksums. Figure regeneration uses `code/plot_lbf_core_erosion_exploratory.py`; the illustrative task image and its complete generation/edit prompt set are stored beside the figures. The reported APs and pairing-bootstrap intervals were independently recomputed from the supplied summaries.

F.7 E12: REGISTERED CONTINUAL CUE-MNIST CONFIRMATION

E12 trains a persistent classifier through four randomized shortcut-learning and recovery cycles. The reported confirmation comprises 64 supervised streams (seeds 8200–8263); model and optimizer state persist across phases.

Images, cues, and network. Each normalized 28×28 MNIST image is multiplied by contrast 0.30, then a single 3×3 patch of intensity one is overwritten at one of ten cue locations. Cue identities 0–4 use zero-based top-left coordinates $(1,1), (1,7), (1,13), (1,19), (1,25)$; identities 5–9 use the same columns in row 25. Nuisance phases reduce contrast to 0.27. The input remains a flattened 784-pixel image. The classifier has architecture $784 \to 64\,(\tanh) \to 10$ logits, with 50,890 trainable parameters and argmax evaluation. Entries of W_1 and W_2 are initialized independently as $\mathcal{N}(0, \frac{2}{784})$ and $\mathcal{N}(0, \frac{2}{64})$, respectively; biases are zero and computations use float32. During harmful phases, the cue equals the digit label with mixture probability 0.98 and is otherwise uniform; hence cue–label agreement has probability 0.982. Benign phases use independent uniform cues. Digit labels never change.

Training and disjoint data. The first $55{,}000$ MNIST training images supply learning batches. A seed-specific permutation of the remaining 5000 provides disjoint fixed core and shift probes of 2000 and 500 images; a separate fixed 1000-image subset of the official test set measures performance. Supervised pretraining uses 3500 updates with batches of 128 rendered images: 64 base images are drawn with replacement, each receiving two distinct cues j and $(j+1) \bmod 10$. Pretraining minimizes mean cross-entropy at learning rate 0.0015 with weight-matrix L2 coefficient 10^{-5}. Continual learning uses the same loss, batch size 128, learning rate 0.001, and L2 coefficient 0.001; biases are unregularized. Adam uses $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and bias correction. Each treatment copies both pretrained parameters and Adam state. Checkpoint zero records the pretrained model; the remaining 299 intervals contain 12 updates each (3588 continual updates).

Success-conditioned cue coverage. Write $f_t(x, j)$ for the predicted label after rendering image x with cue j at reference contrast, and set $q_{\mathrm{ref}}(y) = (y+1) \bmod 10$. Reference-cue accuracy therefore

uses a label-dependent cue, not an unmodified image. The nested probe sets are $U_1 = \{0, 1\}$, $U_2 = \{0, \ldots, 4\}$, and $U_3 = \{0, \ldots, 9\}$. At each checkpoint, recompute

$$B_t := \{i : f_t(x_i, q_{\text{ref}}(y_i)) = y_i\}, \qquad \widehat{c}_t := \frac{1}{|B_t|} \sum_{i \in B_t} \mathbf{1}[\forall j \in U_3,\ f_t(x_i, j) = y_i]\,.$$

Thus $\widehat{c}_t$ measures correctness under *all ten* cues among currently reference-correct examples. For $n_t = |B_t| > 0$, the operational score is the lower endpoint of a nominal two-sided 95% Wilson interval:

$$C_t = \frac{\widehat{c}_t + \frac{z^2}{2n_t} - z\sqrt{\frac{\widehat{c}_t(1-\widehat{c}_t)}{n_t} + \frac{z^2}{4n_t^2}}}{1 + \frac{z^2}{n_t}}, \qquad z = 1.959963984540054.$$

If $B_t = \varnothing$, conditional coverage is undefined (NaN in the implementation) and the operational fallback is $C_t = 0$. This case never occurs in the 19,200 recorded untreated confirmation checkpoints: $|B_t| \geq 1040$. Since $q_{\text{ref}}(y) \in U_3$, unconditional compatible mass is $\frac{|B_t|}{2000}\widehat{c}_t$. Each checkpoint evaluates 2000 reference renderings and 20,000 cue variants, representing 2000 base examples; repeated renderings are not independent observations. The fixed probe supplies repeated monitoring, while inference on prediction and intervention resamples complete streams.

Exact correspondence to conditional-event coverage. Draw an index I uniformly from the fixed 2,000-example probe and define

$$Z_t = (y_I, f_t(x_I, q_{\text{ref}}(y_I)), (f_t(x_I, c))_{c \in U_3})\,.$$

Let μ_t be the empirical law of Z_t on this common coordinate space. Define the fixed events

$$\mathsf{S}_{\text{cue}} := \{Z_{\text{ref}} = Z_y\}, \qquad A_{\text{cue}} := \mathsf{S}_{\text{cue}} \cap \{Z_c = Z_y \text{ for every } c \in U_3\}.$$

Then $\frac{|B_t|}{2000} = \mu_t(\mathsf{S}_{\text{cue}})$ and the displayed estimator satisfies

$$\widehat{c}_t = \mu_t(A_{\text{cue}} \mid \mathsf{S}_{\text{cue}}).$$

Form a one-step probe MDP with initial-state distribution μ_t. Its only action is a dummy action, its goal set is S_{cue}, and its abstraction emits CUE-ROBUST exactly on A_{cue} (and OTHER otherwise). Its Definition 2.1 coverage for the length-one candidate (CUE-ROBUST) is exactly $\widehat{c}_t$. The network update changes μ_t while the probe indices, cues, coordinate events, success set, abstraction, and candidate remain fixed. E12 tests whether erosion of this coverage predicts failure and whether intervention triggered by its lower bound preserves performance.

Prospective prediction and intervention. The frozen prediction score is $s_t = \max\{0, C_{t-3} - C_t\}$, with $s_t = 0$ for $t < 3$, as analyzed in Proposition E.2. Its target is a reference-cue accuracy downcrossing below 0.88 within seven checkpoints, evaluated from healthy checkpoints at or above 0.92; the frozen comparator is cue–label association. Replay instead uses the *level* C_t: it activates below 0.90 and deactivates above 0.94. The accuracy comparator uses reference-cue accuracy on the same core probe, with enter/exit thresholds 0.88/0.92. The decision at checkpoint t controls the next training interval. Active replay replaces the first 64 renderings of each 128-image batch with independently random-cued renderings of those same base images at reference contrast. On the separate performance split, let $A_{\text{ref},t}$ and $A_{\text{current},t}$ denote argmax accuracies under reference and current cues. The reported balanced reward is the whole-stream mean of

$$R_t = \frac{A_{\text{ref},t} + A_{\text{current},t}}{2} - 0.002\,\mathbf{1}\{\text{replay active in the interval ending at } t\}.$$

The configuration, score, comparator, thresholds, seed list, treatment rules, and analysis were frozen before confirmation. Paired 95% intervals use 10,000 nonparametric bootstrap resamples of the 64 complete streams. Checkpoints, incidents, and images remain within-stream observations.

Confirmation results. Core erosion has mean AP 0.3842, versus 0.3450 for cue–label association, for a paired advantage of 0.0392 (95% CI $[0.0224, 0.0555]$). Mean structural warning lead is 4.68 checkpoints. Core-triggered replay has mean balanced reward 0.94995, versus 0.93824 for the accuracy trigger, 0.92836 for matched-random replay, and 0.89602 for never intervening; the oracle obtains 0.95019. The registered core-over-accuracy advantage is 0.01171 ($[0.01131, 0.01212]$). Both co-primary lower endpoints exceed zero, the core raises no nuisance alarms, and every registered supervised integrity and premise gate passes (Figure 18).

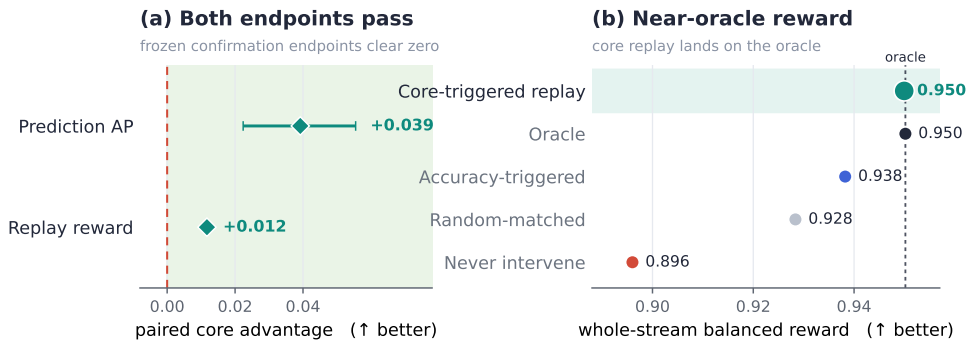

Figure 18: **E12: continual cue-MNIST confirmation (64 complete streams).** *(a)* Registered paired advantages with 95% intervals from 10,000 complete-stream bootstrap resamples. *(b)* Core-triggered replay reaches near-oracle balanced reward and exceeds the accuracy-triggered, matched-random, and never-intervene controls.

Protocol and reproducibility. The compact artifact at `code/continual_cue_mnist_v4_confirmation_stats/` contains the original supervised protocol, configuration snapshot, aggregate summary, and 64 per-stream records with SHA-256 provenance. The implementation at `code/continual_mnist/`, runnable configurations at `code/continual_mnist_configs/`, and accompanying implementation note specify the network and rendering procedure. The note corrects the compact protocol's *three-cue* wording to the third nested set, containing ten cues, and identifies the stored `clean_accuracy` field as reference-cue accuracy. The original protocol and numerical records are retained. Source fingerprints match the freeze manifest and its documented numerical addendum; the latter repairs categorical sampling only in the nonprimary policy-gradient arm and leaves the 64 supervised streams unchanged. Figure regeneration uses `code/plot_cue_mnist_confirmation.py`; aggregate means were independently recomputed from the packaged records.

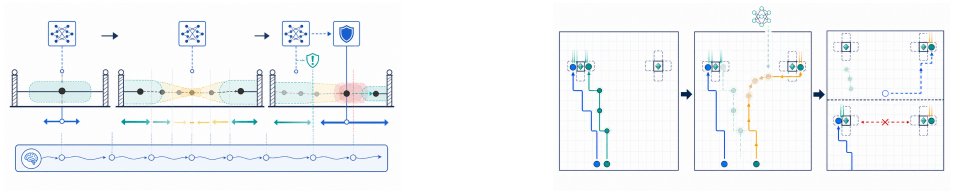

Figure 19: **Task schematics.** *(a)* E10 persistent continual control: actuator drift erodes the nominal controller's viability core, which drives intervention. *(b)* E11 LBF coordination fork: peer learning moves behavior between the left and right conventions while the focal policy remains fixed. Illustrations are explanatory; quantitative results appear in Figure 2.

G LLM USAGE STATEMENT

Language models assisted with developing and editing mathematical proofs by proposing proof outlines and intermediate lemmas, expanding implicit derivations into explicit steps, examining assumptions and edge cases, and providing adversarial review of theorem statements and arguments. They also assisted with copy-editing, LaTeX formatting, figure scripting, terminology lookup, literature discovery across neighboring fields, and non-data task schematics. The authors directed, challenged, and revised this assistance, steered the mathematical development, selected and refined the final arguments, and remain responsible for the manuscript's claims, citations, proofs, empirical results, and contents.