Certified Policy Optimisation for Nested Causal Bandits via PAC-Bayes Risk

Tim Woydt
ProdAxon
Department of Computer Science
Technical University of Darmstadt

Paul-David Zuercher
ProdAxon
Institute for Manufacturing
Department of Engineering
University of Cambridge

aegis+{paul, tim}@prodaxon.com

Abstract

Critical sequential decisions are rarely single-timescale: a strategic decision causally shapes the context in which every subsequent tactical choice is made; standard bandit and reinforcement-learning theory does not capture this causal coupling between timescales. We formalise the problem class as *Nested Contextual Causal Bandits* (NCCBs), a hierarchical SCM where each level's action sets the next level's context distribution, and propose *Nested Causal Thompson Sampling* (NCTS), which draws one mechanism-factorised belief per episode and acts recursively under it. Our main theoretical result is a causal PAC-Bayesian value bound that certifies any posterior-sampled deployment policy from historical data alone, off-policy and anytime, answering the deployment question: *can we trust this agent here, and at what risk?* Experiments on a hierarchical SCM show that, against a matched RFF-GP joint regression on the same function class, the factorised SCM-mechanism posterior transfers *significantly* better zero-shot under large exogenous distribution shifts, the recursive meta-to-inner commit *significantly* dominates the joint-commit alternative in distribution, and the certificate *significantly* contracts as offline data accumulates. Combining these results, we establish *progressive certified handover*, a safe-deployment method: each timescale flips from a legacy controller to NCTS when gains can be certified, re-certified after each earlier flip.[1]

1 Introduction

Critical sequential decision-making, treatment titration in healthcare, input allocation in agronomy, batch configuration in process engineering, is still entrusted to humans because the learned-policy guarantees on offer answer the wrong question. Frequentist regret bounds hold worst-case over all environments; Bayesian regret bounds hold on average over a prior; neither speaks to *this* realised environment, with *this* confounding, under *this* legacy controller. These domains are also naturally hierarchical: a strategic decision generates the context for a tactical one, which generates the context for an operational one. Progressive transfer of control must therefore proceed level by level, gated at each level by an empirical bound meeting a practitioner-specified risk threshold. What is needed is a $(1-\delta)$-bound that holds for the realised environment, simultaneously over candidate policies, across both the offline warm-start phase and the online deployment phase, and gates handover timescale by timescale; no framework we know provides guarantees that are simultaneously timescale-resolved, off-policy valid, and context-specific.

[1] Code: `https://github.com/PaulDavidZuercher/AEGIS`.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

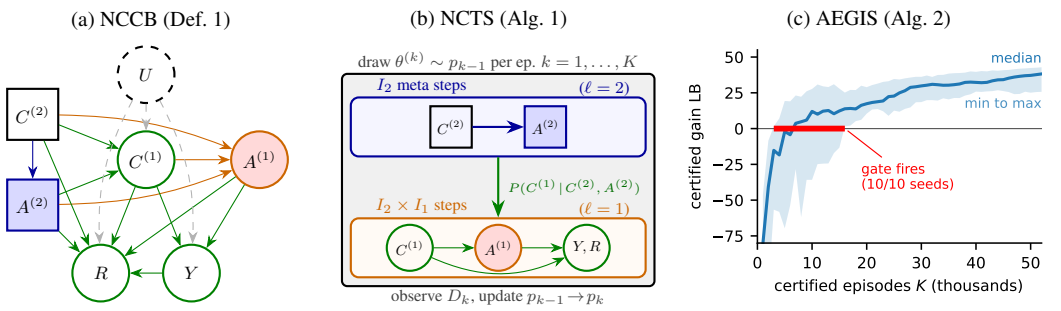


Figure 1: **The NCCB framework at a glance** ($L = 2$). (a) Hierarchical SCM problem class: meta-context $C^{(2)}$ drives meta-action $A^{(2)}$ ($\ell = 2$); together they set the inner context $C^{(1)}$ for the inner action $A^{(1)}$, producing outcome Y and reward R. (b) NCTS (Algorithm 1) draws one belief per episode and acts recursively across levels, with I_2 meta steps and $I_2 \times I_1$ inner steps; after the episode, the posterior is updated. (c) AEGIS deploys NCTS only after the per-level PRISM certificate (Theorem 1) flips, realising the *progressive certified handover* the framework targets. On the two-level Crossing bench (10 seeds, $\delta = 0.05$, posterior still learning, gate suppressed), the leaf-level certified-gain lower bound (LB; median; band: min to max over seeds) first crosses 0, where the gate would fire, between $K = 3{,}000$ and $16{,}000$ (red bar: spread of first crossings over seeds) and reaches $+25.7$ to $+42.9$ by $K = 52{,}000$; on 4 of the 10 seeds it dips below 0 at least once before staying positive (App. H.5). Visual encoding in (a, b): blue/orange, policy decisions; green, learnable mechanisms; grey dashed, latent confounder.

We introduce the *Nested Contextual Causal Bandit* (NCCB), an L-level hierarchical SCM in which each level's decision causally generates the context for the level below; the causal graph is given, only the mechanism parameters are unknown. For this class we propose *Nested Causal Thompson Sampling* (NCTS), which draws one belief per episode and acts on it at every nested level, and *AEGIS* (Anytime Episodic Gain-Certified Inference Schedule), a host-agnostic deployment-time wrapper that runs progressive per-level handover off any host with a Bayesian posterior and a reward predictor. The analysis rests on a single causal PAC-Bayesian value bound (Theorem 1), which certifies the realised environment, off-policy and anytime, with a regulariser that decomposes along the causal mechanisms.

Contributions. *Problem class and algorithm.* We formalise the NCCB problem class (Section 4) and propose NCTS (Section 5), a recursive Thompson rule that draws one belief per episode and acts level by level under it. *Theoretical guarantee.* PRISM (Theorem 1) is a causal PAC-Bayes value bound that, for any intervention scope $\mathcal{I} \subseteq [L]$, factorises both the KL regulariser (over learnable mechanisms, per-mechanism dimension) and the importance weight (over controlled levels). With logarithmic smoothing [29, 8] it is off-policy and anytime via Ville's inequality, with informative offline-data priors; Corollary 5 gives excess risk $\widetilde{O}(1/\sqrt{K})$ under its conditions. AEGIS (Algorithm 2) deploys PRISM as a host-agnostic recipe that hands over level by level. *Empirical evidence.* A hierarchical-SCM benchmark (Section 7) corroborates transfer, nesting and contraction at $p < 0.05$ or better with the function class held fixed (RFF-GP), and two public logs we did not simulate (HeartSteps V1, Open Bandit) run the certificate on real logged data, with zero coverage violations where coverage is checkable (Section 7.3). The certified leaf-level gain becomes and stays positive on all 10 seeds with the posterior still learning. With a frozen posterior, all 120 per-level certificates at depths $L = 3$, 4 and 5 are positive as well (App. H.5, App. K).

2 Related Work

Causal MDPs and Causal Bandits. Lu et al. [17] introduced C-MDPs and C-UCBVI, proving regret $\widetilde{O}(HS\sqrt{ZKH})$ with known graph, where Z compresses the action space via key variables; this compression requires A to affect R only via intermediate variables, so it is vacuous in our NCCB where A directly parents R (Definition 1). Mutti et al. [18] used a partial causal graph prior with posterior sampling (C-PSRL) on factored MDPs, not directly comparable to our nested-bandit setting. Lee and Bareinboim [14, 15] introduced structural causal bandits with see/do separation and

characterised optimal interventions; we extend these primitives hierarchically and add PAC-Bayes off-policy guarantees absent from their graphical analysis.

Posterior-sampling RL and hierarchical bandits. PSRL [22, 20, 21] addresses posterior sampling for tabular and factored MDPs; NCTS is not directly comparable, operating over continuous-parameter SCM posteriors with two-scope (meta/inner) actions rather than tabular factorisations. Hierarchical-bandit work [11, 10, 6] and dyadic RL [16] do not include causal mechanism structure.

PAC-Bayesian learning. Seldin et al. [30] established PAC-Bayes bounds for martingales; Haddouche and Guedj [7] developed the online PAC-Bayes framework. Haddouche and Sakhi [8] extended the logarithmic smoothing of Sakhi et al. [29] to adaptive off-policy PAC-Bayes, which we adapt to between-episode aggregation (Appendix C, Remark 5). Aouali et al. [2] provide a PAC-Bayesian pessimism analysis for offline contextual bandits with regularised IS; closest in technique to Theorem 1, but single-decision and without causal mechanism structure or decision timescales.

Safe policy improvement and OPE. Thomas et al. [32] introduced high-confidence off-policy evaluation; Laroche et al. [13] studied safe policy improvement (SPI) with baseline bootstrapping, which falls back on the baseline policy where uncertainty is high, the closest line in spirit to AEGIS's per-level handover gate. SPI is single-policy and value-estimate-based; AEGIS is per-level and mechanism-factorised through Theorem 1 (anytime-valid via Ville's inequality), so the certificate compounds along the SCM hierarchy rather than collapsing into one policy-level guarantee. Adapting SPI / Aouali-style bounds to NCTS is non-trivial: they assume a single flat policy and an unstructured logged dataset, whereas NCCB requires a per-level certificate over an SCM-factorised policy with mechanism-wise KL decomposition. We leave a tailored SPI-for-NCTS comparison to future work.

3 Preliminaries

Causal models and interventions. A structural causal model (SCM) over endogenous variables $\mathbf{X}$ specifies, for each $X \in \mathbf{X}$, a deterministic equation $X = f_X(\mathrm{Pa}(X), \epsilon_X; \theta^*_X)$ with parent set $\mathrm{Pa}(X) \subseteq \mathbf{X}$ and exogenous noise ϵ_X; the parent sets induce a DAG $\mathcal{G}$, treated as known. Pearl's hard intervention $\mathbf{do}(X = x)$ replaces f_X by $X := x$; a *soft intervention* $\mathbf{do}(X := h)$ generalises this by replacing f_X with a new mechanism h whose parents are non-descendants of X in $\mathcal{G}$. A *policy intervention* on an action A is the soft intervention $\mathbf{do}(A := \pi(\cdot \mid X))$ for an observable context X excluding descendants of A [23].

Contextual bandits. A contextual bandit is an SCM over context C, action A, reward R with $R = f_R(C, A, \epsilon_R; \theta^*_R)$. A policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ induces a policy intervention $\mathbf{do}(A := \pi(\cdot \mid C))$, with value $V(\pi) := \mathbb{E}_{C \sim P_C,\, A \sim \pi(\cdot|C)}[R]$. In the off-policy setting, data are collected under a behaviour policy $\mu \neq \pi$, and $V(\pi)$ is inferred via importance weighting $w_t = \pi(a_t \mid c_t)/\mu(a_t \mid c_t)$ or related estimators.

Thompson sampling. Thompson sampling maintains a posterior $p_k(\theta)$ over the unknown parameters: at episode k, draw one *belief* $\theta^{(k)} \sim p_{k-1}$, act greedily under it, observe $\mathcal{D}_k$, and Bayes-update to p_k. Drawing (rather than maximising) provides exploration with well-understood Bayes-regret guarantees [27, 22].

PAC-Bayes for dependent data. Our off-policy analysis builds on the online PAC-Bayes framework of Haddouche and Guedj [7]: for losses adapted to a filtration $(\mathcal{F}_t)_{t \geq 0}$ and an *online predictive sequence* of priors (P_t) (each P_t being $\mathcal{F}_{t-1}$-measurable, P_1 data-free), a change-of-measure argument with an exponential moment bound yields concentration at a fixed horizon T, uniformly over posterior sequences (Q_t), in the line of the martingale bounds of Seldin et al. [30]. Theorem 1 instead adapts the logarithmic-smoothing bound [8, Thm. 3.2], whose martingale and Ville's inequality give anytime validity; see Appendix C.

Notation. A summary of the main symbols used throughout the paper is provided in Appendix A.

4 Nested Contextual Causal Bandits

We formalise the decision hierarchy informally described in Section 1: a slow outer decision causally shapes the context of every faster inner decision. The SCM template (Definition 1) together with the three assumptions stated below fix the problem class on which the certificate of Theorem 1 operates.

Definition 1 (Nested Contextual Causal Bandit (NCCB))**.** *An L-level NCCB is a hierarchical SCM with endogenous variable set $\mathbf{X} = \{C^{(\ell)}, A^{(\ell)}\}_{\ell=1}^{L} \cup \{Y, R\}$. At level ℓ, the agent observes the context $C^{(\ell)}$ and chooses the action $A^{(\ell)}$ via a stochastic policy $\pi^{(\ell)}$, which we treat as a* soft intervention *on the SCM that replaces the natural (or legacy-logged) mechanism for $A^{(\ell)}$. Y is the* outcome *of the innermost decision and R the reward. Levels are indexed inside-out: $\ell = 1$ is the fastest (innermost), $\ell = L$ the slowest (outermost). The level-ℓ decision causally generates the context $C^{(\ell-1)}$ at the next-faster level via the SCM. The outermost context $C^{(L)}$ is exogenous (no parents in $\mathcal{G}$); the action variables $A^{(\ell)}$ are policy-set; every remaining $X \in \mathbf{X}$ is generated by a learnable mechanism with parameters θ^*_X, giving the canonical learnable-mechanism set $\Phi = \{C^{(\ell)} : \ell < L\} \cup \{Y, R\}$. At each episode k, level ℓ runs I_ℓ decisions: I_L at the outermost level, I_ℓ per parent-level decision for $\ell < L$; the innermost level generates the per-step rewards. Total samples per episode: $N_L := \prod_{\ell=1}^{L} I_\ell$. The reward has one context plus one action per level as parents, plus the outcome Y, so $|\mathrm{Pa}_{\mathcal{G}}(R)| = 2L + 1$ in the scalar case.*

Definition 2 (Intervention scope)**.** *An* intervention scope $\mathcal{I} \subseteq \{1, \ldots, L\}$ *specifies the levels at which the agent acts under a new policy. For $\mathcal{I} \neq \emptyset$, let $\ell^\star_{\mathcal{I}} := \max(\mathcal{I})$ be the outermost new-policy level; only the contexts below $\ell^\star_{\mathcal{I}}$ are relevant for estimating the effect of actions, namely $\Phi_{\mathcal{I}} = \{C^{(\ell)} : \ell < \ell^\star_{\mathcal{I}}\} \cup \{Y, R\}$. The extremes $\ell^\star_{\mathcal{I}} = 1$ and $\ell^\star_{\mathcal{I}} = L$ give the minimal scope $\Phi_{\mathcal{I}} = \{Y, R\}$ and the full scope $\Phi_{\mathcal{I}} = \Phi$. Also, $\Phi_\emptyset := \emptyset$.*

Assumption 1 (Causal Markov + exogenous noise)**.** *The SCM satisfies the Causal Markov condition w.r.t. $\mathbf{X} \cup \mathbf{U}$, where $\mathbf{U} = \{U^{(\ell)}\}_{\ell=1}^{L}$ is a family of exogenous within-level latent variables. Each $U^{(\ell)}$ is i.i.d. across episodes and within-episode samples, mutually independent across ℓ, and independent of the i.i.d. base noise terms ε_X of the mechanisms in Φ (which combine with $U^{(\ell)}$ via additive loadings to form the residual noise of f_X). Soft interventions on $A^{(\ell)}$ replace its natural mechanism without altering any other mechanism in Φ.*

Assumption 2 (Per-level i.i.d. contexts)**.** *For each $\ell < L$, the level-ℓ contexts $C^{(\ell)}$ generated within one parent-level decision are i.i.d. given that decision and the relevant mechanism parameters, and independent of past actions taken inside the same parent-level decision. The outermost contexts are i.i.d. across episodes and across the I_L outer decisions of an episode.*

Assumption 3 (Overlap + backdoor admissibility)**.** *For each level ℓ, any new policy $\pi^{(\ell)}$ is absolutely continuous w.r.t. the legacy policy $\mu^{(\ell)}$: $\pi^{(\ell)}(a \mid \cdot) > 0 \Rightarrow \mu^{(\ell)}(a \mid \cdot) > 0$. The policy context $\mathrm{Pa}_{\mathcal{G}}(A^{(\ell)})$ observable at level-ℓ decision time satisfies the backdoor criterion [23] w.r.t. $(A^{(\ell)}, R)$ in $\mathcal{G}$, so the value of any new π is identifiable from logged data under $\mu = (\mu^{(1)}, \ldots, \mu^{(L)})$.*

Remark 1 (Exogenous confounding)**.** *The framework allows for exogenous within-level confounding at the innermost level ($L \geq 2$): a latent factor $U^{(1)}$ shared across the mechanisms $\{C^{(1)}, Y, R\}$, e.g. a per-step disturbance affecting context, outcome and reward, is absorbed into the residual noise structure under Assumption 1. $V(\pi; \theta^*)$ remains identifiable because $U^{(1)}$ is no parent of $A^{(1)}$, so Assumption 3 holds. Outer levels carry at most one learnable mechanism each, so within-level sharing is moot there. Cross-level confounding (a single latent acting on variables at multiple levels) lies outside this model class; if it also drives an action, handling it requires non-backdoor identification (e.g. instrumental or proximal variables), which we leave to future work.*

Two roles of Assumption 2. The certificate needs only that per-episode conditional means are constant and that episodes are independent given $\mathcal{H}_0$ (App. A; within-episode dependence is permitted, Remark 4); its within-level content (no state transitions inside a level) is needed by the planning recursion (2) and the conditional-mean target of Lemma 2. Within a level the certificate does require *exogeneity*: the context process must not be driven by the within-level actions. Appendix J measures both sides (exogenous AR(1) state at $\rho = 0.9$: 0 of 60 false certifications; endogenous state, engineered candidate, AR(1) coefficient 0.95: 40 of 40); extending the certificate to within-level dynamics is left to future work.

5 Nested Causal Thompson Sampling

NCTS is the recursive Thompson-sampling rule on the L-level NCCB, parametrised by an intervention scope $\mathcal{I} \subseteq \{1, \ldots, L\}$ identifying the levels at which the agent plans (Definition 2). Algorithm 1 summarises one episode; §5.1 gives the formal scope-aware policy and recursive value function, §5.2 the per-mechanism Bayesian update. The $L = 2$ specialisation matching the Sec. 7 experiments

Algorithm 1: NCTS

Input: $\mathcal{G}$ (known); prior p_0; scope $\mathcal{I}$; legacy $(\mu^{(\ell)})_{\ell\notin\mathcal{I}}$; K; $(I_\ell)_{\ell=1}^{L}$.
Output: Posteriors (p_k); within-episode policies $\{\pi_k^{(\ell)}\}$.
Procedure `Traverse(`ℓ, $\tilde{c}^{(\ell+1)}$`)`**:**
  for $i_\ell = 1, \dots, I_\ell$ **do**
    Sample $C^{(\ell)} \sim P_{C^{(\ell)}}(\cdot \mid \tilde{c}^{(\ell+1)}; \theta^*)$; extend $\tilde{c}^{(\ell)} \leftarrow (\tilde{c}^{(\ell+1)}, C^{(\ell)})$;
    if $\ell \in \mathcal{I}$ **then** $a^{(\ell)} \leftarrow \pi_k^{(\ell)}(\tilde{c}^{(\ell)})$ via (1) under $(\mathcal{G}, \theta^{(k)})$;
    else $a^{(\ell)} \sim \mu^{(\ell)}(\cdot \mid \tilde{c}^{(\ell)})$;
    $\tilde{c}^{(\ell)} \leftarrow (\tilde{c}^{(\ell)}, a^{(\ell)})$; **if** $\ell > 1$ **then** `Traverse(`$\ell - 1$, $\tilde{c}^{(\ell)}$`)`;
    else observe Y, R; append $(\tilde{c}^{(\ell)}, Y, R)$ to $\mathcal{D}_k$;
for $k = 1, \dots, K$ **do**
  Draw $\theta^{(k)} \sim p_{k-1}$ (held fixed for episode k); `Traverse(`L, $()$`)`; update p_k from $\mathcal{D}_k$ (per-mechanism: NIG / online Gibbs).

(innermost-level closed form, outer-level Thompson rule) is summarised at the end of this section. An expanded version, with the D-optimal alternative, is in Appendix B; instances up to $L = 5$ are in Appendix K.

5.1 Within-Episode Bandit

Recursive value function. Fix an intervention scope $\mathcal{I} \subseteq \{1, \dots, L\}$. At the start of episode k, draw one belief $\theta^{(k)} \sim p_{k-1}(\theta)$ and hold it fixed throughout the episode. The agent traverses the L levels from $\ell = L$ down to $\ell = 1$; at each level ℓ, after observing the level-ℓ context $C^{(\ell)}$, the level-ℓ action $a^{(\ell)} \in \mathcal{A}^{(\ell)}$ is set by the scope-aware policy

$$\pi_k^{(\ell)}\big(\tilde{c}^{(\ell)}\big) = \begin{cases} \arg\max_{a \in \mathcal{A}^{(\ell)}} \hat{R}_{\mathcal{I}}^{(\ell)}\big(a,\, \tilde{c}^{(\ell)};\, \theta^{(k)}\big), & \ell \in \mathcal{I}, \\ a \sim \mu^{(\ell)}\big(\,\cdot \mid \tilde{c}^{(\ell)}\big), & \ell \notin \mathcal{I}, \end{cases} \tag{1}$$

where the level-ℓ *augmented context* $\tilde{c}^{(\ell)} := (C^{(L)}, a^{(L)}, \dots, a^{(\ell+1)}, C^{(\ell)})$ collects every higher-level context and action together with the current level-ℓ context, $\tilde{c}^{(\ell-1)} := (\tilde{c}^{(\ell)}, a, C^{(\ell-1)})$ extends $\tilde{c}^{(\ell)}$ with the level-ℓ choice a and the freshly drawn level-$(\ell-1)$ context, and $\hat{R}_{\mathcal{I}}^{(\ell)}(a, \tilde{c}^{(\ell)}; \theta)$ is the expected reward over all descendant levels of ℓ under belief θ when controlled levels follow $\pi_k^{(\ell')}$ and uncontrolled levels follow the legacy $\mu^{(\ell')}$:

$$\hat{R}_{\mathcal{I}}^{(\ell)}\big(a,\, \tilde{c}^{(\ell)};\, \theta\big) = \begin{cases} \mathbb{E}\big[R \mid A^{(1)}{=}a,\, \tilde{c}^{(1)};\, \theta\big], & \ell = 1, \\ \mathbb{E}\Big[\max\limits_{a' \in \mathcal{A}^{(\ell-1)}} \hat{R}_{\mathcal{I}}^{(\ell-1)}\big(a',\, \tilde{c}^{(\ell-1)};\, \theta\big) \;\Big|\; A^{(\ell)}{=}a,\, \tilde{c}^{(\ell)};\, \theta\Big], & \ell{-}1 \in \mathcal{I}, \\ \mathbb{E}\big[\,\mathbb{E}_{a' \sim \mu^{(\ell-1)}}\big[\hat{R}_{\mathcal{I}}^{(\ell-1)}\big(a',\, \tilde{c}^{(\ell-1)};\, \theta\big)\big] \mid A^{(\ell)}{=}a,\, \tilde{c}^{(\ell)};\, \theta\big], & \ell{-}1 \notin \mathcal{I}. \end{cases} \tag{2}$$

Equation (2) computes the level-ℓ expected reward recursively over the hierarchy: by Assumption 2, no state transitions occur within a level, so the level-ℓ value is the expected level-$(\ell{-}1)$ optimum (or legacy mean) over the freshly drawn descendant context, evaluated in closed form for linear Gaussian ANMs and by Monte Carlo otherwise (Appendix B).

5.2 Between-Episode Bayesian Update

The learnable mechanisms are $\Phi = \{C^{(\ell)} : \ell < L\} \cup \{Y, R\}$ (Definition 1). We adopt a *mechanism-factorised working model* for prior and likelihood: $p_0(\theta) = \prod_{X \in \Phi} p_0(\theta_X)$ and $q(\mathcal{D}_k \mid \theta, \mathcal{G}) = \prod_{X \in \Phi} p\big(\mathcal{D}_k^{(X)} \mid \theta_X, \mathrm{Pa}(X)\big)$, yielding the per-mechanism update $p_k(\theta_X) \propto p\big(\mathcal{D}_k^{(X)} \mid \theta_X, \mathrm{Pa}(X)\big) \cdot p_{k-1}(\theta_X)$, with $\mathcal{D}_k = \big\{\big(C_{k,\mathbf{i}}^{(\ell)}, a_{k,\mathbf{i}}^{(\ell)}\big)_{\ell=1}^{L},\, Y_{k,\mathbf{i}},\, R_{k,\mathbf{i}}\big\}$ indexed by the within-episode multi-index $\mathbf{i} = (i_1, \dots, i_L)$. For linear ANMs with Gaussian noise, each factor is a closed-form NIG-conjugate update, no MCMC required. For nonlinear ANMs, we use the online Gibbs posterior of Haddouche and Guedj [7], which is PAC-Bayes optimal and absorbs any approximation error into the KL regulariser of Theorem 1 without an additive penalty.

Remark 2 (Working model vs. true likelihood)**.** *Under no shared latent confounding (i.e. at most one $X \in \Phi$ has $\delta_{UX} \neq 0$ in Example 1) and with correctly specified mechanisms, the working model $q(\mathcal{D}_k \mid \theta, \mathcal{G})$ is the true likelihood by the Causal Markov Condition [23], and p_k the exact posterior. Under shared $U^{(1)}$ across $\{C^{(1)}, Y, R\}$ (Remark 1), the marginal likelihood does not factorise mechanism-wise, so p_k is misspecified relative to it. PAC-Bayes absorbs this misspecification: Theorem 1 certifies $V(\pi_Q^{\mathcal{I}}; \theta^*)$ via the IS estimator $\widehat{V}_{\mathrm{pes}}^{\mathrm{LS}}$, which depends on the data only through the level-wise importance weights and the realised reward, not on the working likelihood. Identifiability of V from logged data is supplied by Assumption 3, which holds regardless of the working-model factorisation; what the misspecification can do is widen the certificate (Effect 1, Appendix D), not invalidate it.*

Remark 3 (Known vs. unknown)**.** *The agent is given the functional forms f_X of the learnable mechanisms and an* I-map *$\mathcal{G}$ of the data-generating process: a DAG over $\mathbf{X}$ that contains every true causal edge, possibly with additional spurious ones. This is strictly weaker than knowing the minimal causal graph; every d-separation read off $\mathcal{G}$ holds in the true distribution, so the backdoor criterion (Assumption 3) and the mechanism factorisation (Definition 1) remain sound while the practitioner is freed from resolving internal structure of multi-dimensional variables. The agent is also given the legacy logging policies $\mu = (\mu^{(\ell)})_{\ell=1}^{L}$ that produced the offline data and that act at every level $\ell \notin \mathcal{I}$ during deployment; outside $\mathcal{I}$, $\mu^{(\ell)}$ is treated as a known component of the environment when planning. The only unknown is $\theta^* = (\theta_X^*)_{X \in \Phi}$, the parameters of the learnable mechanisms. In the motivating domains this superset structure is documented and does not need to be discovered (App. J).*

Worked instantiation at $L = 2$. Sec. 7 experiments run at $L = 2$ unless stated otherwise, with $T = C^{(2)}$, $M = A^{(2)}$, $C = C^{(1)}$, $A = A^{(1)}$, $J = I_2$, $I = I_1$ (Figure 1a), so $\tilde{c}^{(2)} = T$ and $\tilde{c}^{(1)} = (T, M, C) =: x$. The inner recursion (2) reduces to a contextual-bandit argmax and the meta-rule integrates over the freshly drawn inner context; Appendix B gives the formulas, the confounder absorption and a D-optimal meta-rule.

6 Causal PAC-Bayesian Bound and Per-Level Certification

The causal factorisation of the posterior (Section 5.2) and the level-factored intervention scope $\mathcal{I} \subseteq \{1, \ldots, L\}$ (Definition 2) together enable a PAC-Bayesian value bound for any level-wise mixture of legacy and new policies. The bound features two parallel factorisations: the KL regulariser sums over the post-intervention learnable mechanisms $\Phi_{\mathcal{I}}$, and the importance weight is a product across the controlled levels $\ell \in \mathcal{I}$. It accommodates linear and nonlinear ANM mechanisms and does not require the reward to decompose additively. We use Q to denote a generic factorised distribution over θ (the certified "posterior" in the PAC-Bayes sense); the bound holds simultaneously over all such Q, with the running Bayesian posterior p_k from Section 5.2 as the natural choice.

Assumption 4 (Bounded rewards and per-level bounded importance weights)**.** (i) Rewards. *The reward is bounded: $R_t \in [0, R_{\max}]$ almost surely, with $0 < R_{\max} < \infty$ known to the practitioner. Raw rewards in a known range $[R_{\mathrm{lo}}, R_{\mathrm{hi}}]$ enter Eq.* (5) *after the affine map $\tilde{R}_t = (R_t - R_{\mathrm{lo}})/(R_{\mathrm{hi}} - R_{\mathrm{lo}})$ to $[0, 1]$; an unbounded reward is first clipped to that range, so that (i) holds by construction for the clipped reward.* (ii) Importance weights. *For each level $\ell \in \mathcal{I}$, there exists $B_w^{(\ell)} < \infty$ such that $\sup_{\theta,\, a^{(\ell)},\, \tilde{c}^{(\ell)}} \pi_\theta^{(\ell)}(a^{(\ell)} \mid \tilde{c}^{(\ell)}) / \mu^{(\ell)}(a^{(\ell)} \mid \tilde{c}^{(\ell)}) \leq B_w^{(\ell)}$.* (iii) Loss range. *Combining (i) and (ii), the per-step loss $\mathcal{L}_t^{\mathcal{I}}(\pi_\theta^{\mathcal{I}}) = w_t^{\mathcal{I}}(\theta) \cdot R_t$ satisfies $\mathcal{L}_t^{\mathcal{I}} \in [0, B^{\mathcal{I}}]$ with $B^{\mathcal{I}} := R_{\max} \prod_{\ell \in \mathcal{I}} B_w^{(\ell)}$, a constant fixed before the data. In the on-policy case ($\pi_\theta^{(\ell)} = \mu^{(\ell)}$ for all $\ell \in \mathcal{I}$), $B_w^{(\ell)} = 1$ and $B^{\mathcal{I}} = R_{\max}$.*

Theorem 1 (PRISM: Per-mechanism Risk Inequality with Scope Mixing)**.** *Fix any intervention scope $\mathcal{I} \subseteq \{1, \ldots, L\}$. The hybrid policy $\pi_\theta^{\mathcal{I}}$ has $\pi_\theta^{(\ell)} = \mu^{(\ell)}$ at $\ell \notin \mathcal{I}$ (Eq.* (1)*); let $\pi_Q^{\mathcal{I}}$ denote the randomised version that draws $\theta \sim Q$ once per episode, so that $V(\pi_Q^{\mathcal{I}}; \theta^*) = \mathbb{E}_{\theta \sim Q}[V(\pi_\theta^{\mathcal{I}}; \theta^*)]$. Index the certification dataset $\mathcal{D}_{\mathrm{cert}}$ by episode $k = 1, \ldots, K$ and within-episode multi-index $\mathbf{i} \in [N_L]$. The level-factored importance weight is the product over all L levels of (deployment policy)/(behaviour policy); when $\mathcal{D}_{\mathrm{cert}}$ was collected under the legacy $\mu = (\mu^{(\ell)})_{\ell=1}^{L}$, the legacy/legacy factors at $\ell \notin \mathcal{I}$ cancel, leaving*

$$w_t^{\mathcal{I}}(\theta) := \prod_{\ell=1}^{L} \frac{\pi_\theta^{(\ell)}\big(a_t^{(\ell)} \mid \tilde{c}_t^{(\ell)}\big)}{\mu^{(\ell)}\big(a_t^{(\ell)} \mid \tilde{c}_t^{(\ell)}\big)} \;=\; \prod_{\ell \in \mathcal{I}} \frac{\pi_\theta^{(\ell)}\big(a_t^{(\ell)} \mid \tilde{c}_t^{(\ell)}\big)}{\mu^{(\ell)}\big(a_t^{(\ell)} \mid \tilde{c}_t^{(\ell)}\big)}. \tag{3}$$

Define the per-sample hybrid loss $\mathcal{L}^{\mathcal{I}}_{k,\mathbf{i}}(\pi^{\mathcal{I}}_\theta) := w^{\mathcal{I}}_{k,\mathbf{i}}(\theta) \cdot R_{k,\mathbf{i}}$ *and the* episode-mean loss $\bar{\mathcal{L}}^{\mathcal{I}}_k(\pi^{\mathcal{I}}_\theta) := \frac{1}{N_L}\sum_{\mathbf{i}\in[N_L]} \mathcal{L}^{\mathcal{I}}_{k,\mathbf{i}}(\pi^{\mathcal{I}}_\theta) \in [0, B^{\mathcal{I}}]$. *Under Assumptions 1–4, for any data-free prior* $p_0 = \prod_{X\in\Phi} p_{0,X}$, *any* $\delta \in (0,1)$ *and any fixed* $\lambda \in (0, 1/B^{\mathcal{I}})$, *in the off-policy phase (legacy* μ *fixed across episodes), with probability at least* $1-\delta$ *simultaneously over all factorised* $Q = \prod_{X\in\Phi_{\mathcal{I}}} Q_X$ *and all* $K \geq 1$*:*

$$V(\pi^{\mathcal{I}}_Q;\, \theta^*) \geq \widehat{V}^{\mathrm{LS}}_{\mathrm{pes}}(\pi^{\mathcal{I}}_Q) - \frac{\sum_{X\in\Phi_{\mathcal{I}}} \mathrm{KL}(Q_X \| p_{0,X}) + \log(1/\delta)}{\lambda K}, \tag{4}$$

where the pessimistic Logarithmic Smoothing value estimator is

$$\widehat{V}^{\mathrm{LS}}_{\mathrm{pes}}(\pi^{\mathcal{I}}_Q) := B^{\mathcal{I}} + \frac{1}{\lambda K}\sum_{k=1}^{K} \mathbb{E}_{\theta\sim Q}\Big[\log\Big(1 - \lambda\left(B^{\mathcal{I}} - \bar{\mathcal{L}}^{\mathcal{I}}_k(\pi^{\mathcal{I}}_\theta)\right)\Big)\Big]. \tag{5}$$

The certified value is the expected reward of $\pi^{\mathcal{I}}_Q$ *under the true environment* θ^**; the KL regulariser sums over* $\Phi_{\mathcal{I}}$ *only, mirroring the cancellation of the IS weight at* $\ell \notin \mathcal{I}$*. For ANMs whose mechanism posteriors satisfy* $\mathrm{KL}(Q_X\|p_{0,X}) = O(d_X(\sum_{\beta\in\mathrm{Pa}(X)} d_\beta + 1)\log K)$ *(a regularity condition met by Gaussian-conjugate updates), the total regulariser scales with the per-mechanism parameter dimension rather than the full joint space. Under further conditions, the excess-risk rate of* $\widetilde{O}(1/\sqrt{K})$ *in episodes is Corollary 5; the full proof is in Appendix C.*

Proof sketch. (1) Causal KL decomposition: the mechanism-wise factorisation makes Q and p_0 product distributions over Φ; the chain rule gives $\mathrm{KL}(Q\|p_0) = \sum_X \mathrm{KL}(Q_X\|p_{0,X})$, and for $X \notin \Phi_{\mathcal{I}}$ the convention $Q_X = p_{0,X}$ collapses the sum to $\sum_{X\in\Phi_{\mathcal{I}}}$. *(2) Hybrid IS unbiasedness*: actions are drawn level by level, so the chain rule along the level-ordering yields the level-factored Radon–Nikodym derivative $dP_{\mathrm{hybrid}}/dP_{\mathrm{legacy}} = w^{\mathcal{I}}$ per sample; together with the level-wise backdoor admissibility of Assumption 3 on $\mathcal{I}$, every per-sample loss has constant conditional mean $\mathbb{E}[\mathcal{L}^{\mathcal{I}}_{k,\mathbf{i}} \mid \mathcal{H}_{k-1}] = V(\pi^{\mathcal{I}}_\theta;\theta^*)$, and so does the episode mean $\mathbb{E}[\bar{\mathcal{L}}^{\mathcal{I}}_k \mid \mathcal{H}_{k-1}] = V(\pi^{\mathcal{I}}_\theta;\theta^*)$. *(3) Episode-level LS supermartingale*: applying the logarithmic-smoothing construction of Haddouche and Sakhi [8, Thm. 3.2] to the episode-mean gap $\Delta\bar{\mathcal{L}}^{\mathcal{I}}_k := B^{\mathcal{I}} - \bar{\mathcal{L}}^{\mathcal{I}}_k \in [0, B^{\mathcal{I}}]$ across the K off-policy episodes, which are conditionally independent given $\mathcal{H}_0$ (Assumption 2), and combining with Ville's inequality and change of measure (Appendix C, Remark 5), yields the bound. □

Interpreting the LS estimator, and what is certified. Since $-\log(1-u) \geq u$ for $u < 1$, the log-tempered mean (5) lies below the IPS mean, and (4) is a high-probability lower bound. As $\lambda \to 0$ the estimator recovers plain IPS; $\lambda > 0$ adds pessimism that grows as the loss falls below $B^{\mathcal{I}}$. The logarithm makes the underlying process a supermartingale (Lemma 4), hence anytime validity. Unlike Thompson-sampling regret analyses, which control the exploration cost of a sampling rule, Theorem 1 certifies a given posterior from logged data. For the full scope, Gaussian priors and a value Lipschitz in θ, Corollary 5 bounds the excess risk of the certificate-maximising posterior by $O(p_\Phi \log(K+1)/\sqrt{K})$, with a constant affine in $B^{\mathcal{I}}$. Cumulated over K episodes this has the $\sqrt{K}$ shape of Thompson-sampling analyses, with p_Φ in place of the joint dimension, but for a posterior fitted to logs, not the regret of NCTS.

Scope and extensions. Corollary 6 extends PRISM to $\mathcal{H}_{k-1}$-measurable behaviour (μ_k) with overlap.

6.1 AEGIS: Anytime Episodic Gain-Certified Inference Schedule

Per-level certification and safe-deployment selection. The bound (4) promotes the deployment scope $\mathcal{I}$ from an all-or-nothing decision to a design choice the practitioner optimises against the certificate. Since $\widehat{V}^{\mathrm{LB}}_{\mathcal{I}}$, the right-hand side of (4), is a high-probability lower bound on the realised value of deploying $\pi^{\mathcal{I}}_Q$, the safest scope maximises the guarantee:

$$\mathcal{I}^\star \;=\; \arg\max_{\mathcal{I}\in\mathfrak{J}} \widehat{V}^{\mathrm{LB}}_{\mathcal{I}}(\pi^{\mathcal{I}}_Q), \tag{6}$$

not necessarily the scope with the highest expected reward, since the multiplicative compounding $B^{\mathcal{I}} \leq R_{\max}\prod_{\ell\in\mathcal{I}} B^{(\ell)}_w$ in $\lambda < 1/B^{\mathcal{I}}$ can loosen the certificate faster than the mean improves.

Single-level scopes $\mathcal{I} = \{\ell\}$ are the natural building blocks of the progressive workflow advertised in Section 1: deploy at the safest level, collect fresh data under the hybrid behaviour, re-apply Theorem 1

for the next handover candidate. Note that the per-level certificates need to be recomputed on fresh data after the deployment of any intervention-scoped policy.

Computing the per-scope certificates. Every candidate scope $\mathcal{I} \in \mathfrak{I}$ is certified on the *same* $\mathcal{D}_{\text{cert}}$, recomputing only $w^{\mathcal{I}}$ (3), $B^{\mathcal{I}}$ and the KL sum over $\Phi_{\mathcal{I}}$. The temperature is $\lambda^{\star}_{\mathcal{I}} = \left(2\left(\sum_{X\in\Phi_{\mathcal{I}}} \mathrm{KL}(Q_X\|p_{0,X}) + \log(1/\delta)\right)/(K\,\widehat{S}^2_{\mathcal{I}})\right)^{1/2}$ with $\widehat{S}^2_{\mathcal{I}}$ the second moment of the episode-mean gap, clipped inside $\lambda < 1/B^{\mathcal{I}}$ and computed on the prior split (Remark 6); all scopes together cost a $\log|\mathfrak{I}|$ union-bound term and linear enumeration (Appendix C.8).

The AEGIS recipe. AEGIS schedules (6) on any host with posterior p_k, KL to data-free p_0, and per-mechanism predictor $\hat{r}_\theta$, via three switches: an optional leaf LCB policy

$$\pi^{\mathrm{RA}}_k(x) = \arg\max_{a\in\mathcal{A}^{(1)}} \widehat{\mu}_k(a,x) - \kappa_k \widehat{\sigma}_k(a,x), \tag{7}$$

$$\kappa_k = \sqrt{\tfrac{2(\mathrm{KL}(p_{k-1}\|p_0)+\log(1/\delta))}{k}}, \tag{8}$$

($x = \tilde{c}^{(1)}$; $\widehat{\mu}_k, \widehat{\sigma}_k$: mean, std of $\hat{r}_\theta(a,x)$ over n_{RA} draws $\theta \sim p_{k-1}$); a closed-form bound-minimising

$$\lambda^{\star}_k \;=\; \min\!\Big(\tfrac{1-\eta_{\mathrm{clip}}}{B^{\mathcal{I}}},\; \sqrt{\tfrac{2C_k}{k\,S^2_k}}\Big), \tag{9}$$

with $C_k = \mathrm{KL}(p_k\|p_0) + \log(1/\delta)$, episode-mean second-moment S^2_k, and clipping margin η_{clip} (App. E; evaluated on the prior split, Remark 6); and a sticky per-level handover gate Σ_k that adds ℓ to Σ_{k+1} once $G^{(\ell)}_k$, the right-hand side of (29) at level ℓ, reaches ε_ℓ. A flipped level keeps the π_{p_k} certified at its flip; no reported run plays (7) after a flip. The gate fires on the $(1-\delta)$ *lower bound* of the gain, so it can stay closed while $V(\pi^{(\ell)}_{p_k})$ already exceeds $V(\mu^{(\ell)}_{\mathrm{legacy}})$. This conservativeness follows from the log directly, as legacy data without randomisation over the candidate's support caps the certificate at $\widehat{V}(\mu)$ for every K (a deterministic legacy has $w \le 1$), so overlap in the log is a design-time precondition that a longer run cannot supply. Given it, the remaining requirement is computable before data, as the gain must exceed the approximate median firing point $G_{\min}(K) = c/\sqrt{K}$ (Appendix D, Effect 5). Sec. 7 evaluates the AEGIS gate on NCTS runs; rationale, hosts and notes: App. E.

Algorithm 2: AEGIS schedule (host-agnostic, L-level). δ is split over levels, epochs and checks (App. H.5); joint-scope diagnostic $G^{[L]}_k$ and design notes: App. E.

Input: Host: p_0, $\mathsf{HostUpdate}(p_{k-1},\mathcal{D}_k)\to p_k$, predictor $\hat{r}_\theta$. Schedule: $K, L, (I_\ell)_{\ell=1}^{L}$. AEGIS: legacy $(\mu^{(\ell)}_{\mathrm{legacy}})$, thresholds (ε_ℓ), $\delta, \eta_{\mathrm{clip}}$.
Output: Sticky handover (Σ_k); posteriors (p_k).
Σ_1, buffers$\leftarrow\emptyset$; $\lambda\leftarrow$(9) on prior split, declared KL;
Procedure `Traverse(`ℓ, $\tilde{c}^{(\ell+1)}$`)`**:**
  for $i_\ell = 1,\dots,I_\ell$ **do**
    Sample $C^{(\ell)} \sim P_{C^{(\ell)}}(\cdot\,|\,\tilde{c}^{(\ell+1)};\theta^*)$; extend $\tilde{c}^{(\ell)} \leftarrow (\tilde{c}^{(\ell+1)}, C^{(\ell)})$;
    if $\ell \in \Sigma_k$ **then**
      $a^{(\ell)} \sim \pi^{(\ell)}_{p_{k_\ell}}$;
    else
      $a^{(\ell)} \sim \mu^{(\ell)}_{\mathrm{legacy}}(\cdot\,|\,\tilde{c}^{(\ell)})$; append to $\mathcal{D}^{\mu,(\ell)}$, also $\mathcal{D}^{\mu,[L]}$ if $\Sigma_k = \emptyset$;
    $\tilde{c}^{(\ell)} \leftarrow (\tilde{c}^{(\ell)}, a^{(\ell)})$; **if** $\ell > 1$ **then**
      `Traverse(`$\ell-1$, $\tilde{c}^{(\ell)}$`)`;
    else observe Y, R; append $(\tilde{c}^{(\ell)}, Y, R)$ to $\mathcal{D}_k$;
for $k = 1,\dots,K$ **do**
  `Traverse(`L, $()$`)`;
  $p_k \leftarrow \mathsf{HostUpdate}(p_{k-1},\mathcal{D}_k)$; $\Sigma_{k+1}\leftarrow\Sigma_k$;
  for $\ell \in [L]\setminus\Sigma_k$ **with** $|\mathcal{D}^{\mu,(\ell)}| > 0$ **do**
    if $G^{(\ell)}_k(\pi_{p_k};\,\mathcal{D}^{\mu,(\ell)}) \ge \varepsilon_\ell$ **then**
      $\Sigma_{k+1} \leftarrow \Sigma_{k+1}\cup\{\ell\}$ *(sticky)*; $k_\ell \leftarrow k$; clear $\mathcal{D}^{\mu,(\ell')}$, $\ell' \ne \ell$;

6.2 Approximately Correct Graphs: Error Taxonomy and Corruption Study

Four error cases and the logging condition. We distinguish four ways the supplied graph, or the log it is applied to, can be wrong (Appendix D.1). *(i) Extra edges* cost compression, not validity (Remark 3 requires only an I-map). *(ii) Working-model misspecification* can widen the certificate without invalidating it (Remark 2). *(iii) A missing observation* no policy acted on cannot confound the action, so Lemma 2 still holds, even if the coarser graph is no I-map; it costs performance, not validity. *(iv) Missing logging*, a log that omits a variable the behaviour policy acted on or replaces it by a post-action measurement, breaks the identification step of Lemma 2 and alone of the four can falsely certify. The requirement thus reduces from "know the causal graph" to a checkable *logging condition*: every reweighted dataset must record, at their decision-time values, the variables its behaviour policy acted on.

Corruption study. On Γ-SCM (10 seeds, $\delta = 0.05$; Table 3, App. D.1), every valid arm stays at 0/10 false certifications, including a supplied graph that is not an I-map. Only mode 3a (confounded

logging) fails (up to 10/10) and opens the gain gate, on 4 seeds for a policy with negative true gain. Corrupted certificates are *not self-diagnosing*, so the logging condition is a design-time requirement, not a runtime check.

Logging requirement and post-handover buffers. The gate needs logged (x_t, a_t^μ, R_t) tuples and their propensities; for a deterministic legacy these are 1 and the IS ratio in (5) is $\pi_\theta^{(\ell)}(a_t^\mu \mid x_t)$ for each $\theta \sim p_k$. Algorithm 2 clears the other levels' buffers $\mathcal{D}^{\mu,(\ell)}$ at a flip, so the next gate certifies the current hybrid; uncleared buffers certify a time average over pre- and post-flip hybrids (Corollary 6), frozen ones the full-legacy hybrid (App. E).

7 Experiments

7.1 Setup

The $L{=}2$ benchmarks instantiate one parametric SCM (Γ-SCM; Table 4 lists every benchmark) with meta-action M outer, inner-action A inner (both bounded $[-2, 2]$), and latent noise U; unless stated otherwise, every bench runs at the Γ-Stress preset ($\eta{=}1$, $\sigma_Q{=}3$, $J{=}1$, $I{=}20$) with the `RFF-GP-Gibbs` posterior backend. The four body bandits share a common RFF-GP function class and differ only in reward model and commit shape: Flat-TS (joint regression $(T, M, C, A){\to}R$, no SCM), Factored-TS (SCM-factorised f_C, f_Y, f_R, flat commit), Joint-Commit (SCM-factorised with (M, A) committed jointly at meta-start), and NCTS (full NCTS, Algorithm 1); Table 5 lists all arms. Unless stated otherwise, numbers use 10 seeds, $K{=}2000$ episodes, mean $\pm\, 1\sigma$ across seeds; significance is Welch's t-test with reported d and p. Deviations are stated inline; details: App. G.

7.2 Constructive Ablations

Three claims motivate NCTS: a factorised SCM-mechanism posterior is the dominant gain under shift (Sec. 7.2.1); re-selecting A per realised inner context dominates committing (M, A) jointly (Sec. 7.2.2; at $L{=}3$ by $+12.8$ per episode, App. K); and each level can be certified in isolation, which drives the handover of Section 6.1 (Sec. 7.2.3).

7.2.1 Factorisation gains *significantly* under exogenous shift.

Flat-TS and Factored-TS differ only in the reward model; trained on the Γ-Stress source, frozen and replanned zero-shot onto shifted targets, the factorised bandit attains $\mathbf{113.6{\pm}4.9}$ versus $56.1{\pm}10.1$ under Shift-T and $\mathbf{194.7{\pm}4.9}$ versus $146.1{\pm}6.6$ under Shift-$\mathcal{N}_Q$ (both $p < 10^{-3}$; mild Shift-U: $+1.2$, App. H.4). The joint regression collapses once the shift exits the source support while the ICM-invariant mechanisms hold, the empirical complement to the necessity result of Richens and Everitt [25] (Fig. 2a in App. H).

7.2.2 Recursive nesting *significantly* dominates joint commit.

Joint-Commit reuses the SCM posterior but commits A with M at meta-start (conditioning on $\widehat{C} = \mathbb{E}[C \mid T, M]$, reused for all $I{=}20$ inner steps); NCTS re-selects A at every realised C. Causal factorisation contributes $+3.3$ reward and nesting a further $+3.9$ (Table 1, Fig. 2b); joint commitment forfeits 37.3 against Factored-TS, since A can no longer adapt to the realised C_t. The gate earns no reward itself; it decides *when* to deploy. On Depth-Bin at $L{=}3$ (App. K.5.2) it fires on $10/10$ seeds for a good candidate at -12.71 per episode against *always* (the pre-crossing delay) and on $0/10$ for the negated policy, saving $+23.56$ (Table 1): asymmetric protection, not reward.

7.2.3 The bound contracts, and the certified gain crosses zero.

Two certified quantities must be kept apart: the *absolute* bound V_{lower} of Theorem 1 on $V(\pi_Q; \theta^*)$, and the *gain* bound (29) on which the gate acts, which subtracts the incumbent's estimated value; an absolute bound far below zero is compatible with a positive certified gain.

Absolute side: contraction. The per-episode bound contracts $2.51\times$, from $-50{,}274{\pm}2{,}149$ to $-20{,}053{\pm}1{,}503$ (Table 7). We re-apply the theorem to each prefix of an NCTS run on Contract (App. H); clipping rewards to $[-15, 15]$ makes Assumption 4(i) hold by construction. At this budget

Table 1: Effect attribution: per-episode reward, mean ± sd, 10 seeds; Welch p, Cohen's d (pooled sd). Top: decomposition on Γ-Stress. Bottom: gate ablation on Depth-Bin (L=3), AEGIS spliced at the crossing, Δ paired. More rows: App. H.4.

Effect	**Contrast**	**A**	**B**	Δ	d	p
Factorisation	Factored-TS vs Flat-TS	114.7 ± 0.3	111.5 ± 0.5	$+3.3$	8.21	5.1×10^{-11}
Nesting (on top)	NCTS vs Factored-TS	118.6 ± 0.6	114.7 ± 0.3	$+3.9$	7.81	7.7×10^{-10}
Both (total)	NCTS vs Flat-TS	118.6 ± 0.6	111.5 ± 0.5	$+7.1$	12.34	1.7×10^{-15}
Nesting vs joint	NCTS vs Joint-Commit	118.6 ± 0.6	77.4 ± 3.4	$+41.2$	16.95	7.8×10^{-12}
Gate, good cand.	AEGIS vs *always*	10.86 ± 3.90	23.56 ± 0.53	-12.71 ± 3.73	fires $10/10$	
Gate, bad cand.	AEGIS vs *always*	0.00 ± 0.37	-23.56 ± 0.66	$+23.56\pm0.50$	fires $0/10$	

the bound is vacuous: it lies far below the floor of -300 per episode (N_L=20), because the regulariser $(\mathrm{KL}_\Phi + \log(1/\delta))/(\lambda K)$ dominates, with $\mathrm{KL}_\Phi \approx 2.3 \times 10^3$ nats at K=2000 for the D=128 RFF basis.

Gain side: crossings. To leading order the gain bound crosses zero after a number of episodes linear in $\mathrm{KL}_\Phi + \log(1/\delta)$ and quadratic in $B_w(R_{\mathrm{hi}} - R_{\mathrm{lo}})/g$, g the candidate's advantage (Eq. (31)). With a uniform incumbent on a 3-point leaf grid (B_w=3, attained exactly) and the posterior *learning*, the leaf certificate crosses zero on $10/10$ seeds at K=3,000 to 16,000 (δ=0.05, KL 2.55 to 25.73 nats), reaching $+25.69$ to $+42.86$ per episode by K=52,000; with the posterior frozen it crosses under non-additive noise, on Bounded and at every level for L=3, 4, 5 (App. H.5, App. K). We recommend deploying only once the gain stays positive at every later checkpoint; this rule still fires on all ten seeds (K=3,000 to 16,000).

7.3 External-Data Checks: Open Bandit and HeartSteps V1

Two public logs test the certificate alongside simulated scenarios (App. M). On the **Open Bandit Dataset** [28], the Bernoulli-TS policy certified from the platform's uniform-Random log ($B_w = 8.84$ exact, $\mathrm{KL} = 0$, $\delta = 0.05$) covers its directly measured value in all 81 cells; the bound contracts from -2.38 at $K = 100$ to -0.21 at $K = 10{,}000$ (as does a zero-click control), and the reverse direction ($B_w = 277.8$) is about $32\times$ looser at every K. With binary clicks the gap to the on-policy value is 52 to 570 times that value, so these cells rule out only gross miscalibration. On the micro-randomised **HeartSteps V1** trial [12] ($B_w = 2.5$ exact), the outer certificate is positive ($V_{\mathrm{lower}} = +5.63$ and $+5.39$ per user-day) and the inner one lies $+2.43$ and $+1.33$ above its trivial floor of -2.77. Neither log certifies improvement over its logging policy, and the inner level lies in the endogenous regime Assumption 2 excludes, so the coverage claim rests on Open Bandit.

8 Conclusion

We introduced the NCCB framework, an L-level hierarchical SCM for progressive certified handover across decision timescales, and proposed NCTS as a first concrete algorithm (recursive Thompson sampling, one belief per episode, known graph). Theorem 1 is a causal PAC-Bayes bound with a KL regulariser that decomposes along the causal mechanisms; it is off-policy valid for linear and nonlinear ANMs and certifies any intervention scope $\mathcal{I} \subseteq [L]$, exposing the safest deployment scope as the certificate-maximising one (Section 6.1). Under its conditions, Corollary 5 gives excess risk $\widetilde{O}(1/\sqrt{K})$ in episodes.

Three limitations stand out (extended discussion in Appendix J). *(i) Within-level exogeneity* (Assumption 2) is exact for batch domains with parallel or exchangeable inner units (wafer lots, field plots, quality samples), and state carried across timescales is covered through $C^{(\ell)}$. Its boundary is treatment titration: each dose step reacts to the previous response, so the within-level context is endogenous and the level-factored weight (3) no longer identifies the value. *(ii) Mechanism class*: additive-noise models; the non-additive case is a stress test only. *(iii) Evaluation scope*: exact validity verification remains synthetic, since it requires ground-truth policy values; a real log allows checking coverage (Section 7.3), not exact validity. NCTS is a baseline; Apps. J and L give open directions.

Acknowledgments and Disclosure of Funding

We thank the reviewers and the AC for their constructive feedback, which helped to improve this work. This work was funded by the European Union (Grant Agreement no. 101120763 - TANGO). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. This work was also supported by the Engineering and Physical Sciences Research Council [EP/S023917/1]. We thank our supervisors and colleagues Kristian Kersting, Thomas Bohné, Moritz Willig, Raban Emunds, Yi-Shan Wu, and Yevgeny Seldin for the fruitful and inspiring discussions and feedback.

References

[1] Y. Annadani, N. Pawlowski, J. Jennings, S. Bauer, C. Zhang, and W. Gong. BayesDAG: Gradient-based posterior inference for causal discovery. *Advances in Neural Information Processing Systems*, 36:1738–1763, 2023.

[2] I. Aouali, V.-E. Brunel, D. Rohde, and A. Korba. Unified PAC-Bayesian study of pessimism for offline policy learning with regularized importance sampling. In *Conference on Uncertainty in Artificial Intelligence*, pages 88–109. PMLR, 2024.

[3] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14:3207–3260, 2013.

[4] K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10 (3):273–304, 1995.

[5] S. Girard, A. Bibaut, and H. Zenati. Online policy learning via a self-normalized maximal inequality. *arXiv preprint arXiv:2510.15483*, 2025.

[6] J. Guan and H. Xiong. Improved Bayes regret bounds for multi-task hierarchical Bayesian bandit algorithms. *Advances in Neural Information Processing Systems*, 37:72964–72999, 2024.

[7] M. Haddouche and B. Guedj. Online PAC-Bayes learning. *Advances in Neural Information Processing Systems*, 35:25725–25738, 2022.

[8] M. Haddouche and O. Sakhi. Logarithmic smoothing for adaptive PAC-Bayesian off-policy learning. *arXiv preprint arXiv:2506.10664*, 2025.

[9] T. Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995.

[10] J. Hong, B. Kveton, S. Katariya, M. Zaheer, and M. Ghavamzadeh. Deep hierarchy in bandits. In *International Conference on Machine Learning*, pages 8833–8851. PMLR, 2022.

[11] J. Hong, B. Kveton, M. Zaheer, and M. Ghavamzadeh. Hierarchical Bayesian bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 7724–7741. PMLR, 2022.

[12] P. Klasnja, S. Smith, N. J. Seewald, A. Lee, K. Hall, B. Luers, E. B. Hekler, and S. A. Murphy. Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of HeartSteps. *Annals of Behavioral Medicine*, 53(6):573–582, 2019.

[13] R. Laroche, P. Trichelair, and R. Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pages 3652–3661. PMLR, 2019.

[14] S. Lee and E. Bareinboim. Structural causal bandits: Where to intervene? *Advances in Neural Information Processing Systems*, 31, 2018.

[15] S. Lee and E. Bareinboim. Characterizing optimal mixed policies: Where to intervene and what to observe. *Advances in Neural Information Processing Systems*, 33:8565–8576, 2020.

[16] S. Li, L. Salvat Niell, S. W. Choi, I. Nahum-Shani, G. Shani, and S. Murphy. Dyadic reinforcement learning. *arXiv preprint arXiv:2308.07843*, 2023.

[17] Y. Lu, A. Meisami, and A. Tewari. Efficient reinforcement learning with prior causal knowledge. In *Conference on Causal Learning and Reasoning*, pages 526–541. PMLR, 2022.

[18] M. Mutti, R. De Santi, M. Restelli, A. Marx, and G. Ramponi. Exploiting causal graph priors with posterior sampling for reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[19] T. NeCamp, S. Sen, E. Frank, M. A. Walton, E. L. Ionides, Y. Fang, A. Tewari, and Z. Wu. Assessing real-time moderation for developing adaptive mobile health interventions for medical interns: micro-randomized trial. *Journal of Medical Internet Research*, 22(3):e15033, 2020.

[20] I. Osband and B. Van Roy. Near-optimal reinforcement learning in factored MDPs. *Advances in Neural Information Processing Systems*, 27, 2014.

[21] I. Osband and B. Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International Conference on Machine Learning*, pages 2701–2710. PMLR, 2017.

[22] I. Osband, D. Russo, and B. Van Roy. (More) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

[23] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

[24] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

[25] J. Richens and T. Everitt. Robust agents learn causal world models. In *The Twelfth International Conference on Learning Representations*, 2024.

[26] D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. *Advances in Neural Information Processing Systems*, 27, 2014.

[27] D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[28] Y. Saito, S. Aihara, M. Matsutani, and Y. Narita. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

[29] O. Sakhi, I. Aouali, P. Alquier, and N. Chopin. Logarithmic smoothing for pessimistic off-policy evaluation, selection and learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[30] Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.

[31] A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[32] P. Thomas, G. Theocharous, and M. Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[33] J. Wang, Y. Fang, E. Frank, M. A. Walton, M. Burmeister, A. Tewari, W. Dempsey, T. NeCamp, S. Sen, and Z. Wu. Effectiveness of gamified team competition as mHealth intervention for medical interns: a cluster micro-randomized trial. *npj Digital Medicine*, 6(1):4, 2023.

[34] D. Williams. *Probability with martingales*. Cambridge University Press, 1991.

[35] Y. Yu, T. Gao, N. Yin, and Q. Ji. DAGs with no curl: An efficient DAG structure learning approach. In *International Conference on Machine Learning*, pages 12156–12166. PMLR, 2021.

A Notation

Table 2: Notation used throughout the paper. Symbols are defined at first use, in the body or in the appendix cited here; this table is for reference.

Causal model.	
$\mathcal{G}$, $\mathrm{Pa}(X)$	Causal structure (notational shorthand for the parent assignment); parent set of mechanism X
$\mathbf{do}(X{=}x)$	Pearl intervention setting X to x
$\theta^* = (\theta^*_C, \theta^*_Y, \theta^*_R)$	True SCM mechanism parameters
Φ, $d_X := \dim(X)$, $D := \sum_{X\in\Phi} d_X$	Set of learnable mechanisms (Def. 1); per-mechanism output dimension; total parameter dimension
Decision protocol.	
K, L, $(I_\ell)_{\ell=1}^{L}$, $N_L := \prod_\ell I_\ell$	Episodes; nesting depth ($\ell{=}1$ innermost, $\ell{=}L$ outermost); per-level loop counts; total inner samples per episode
$J = I_2$, $I = I_1$, $H = JI = N_2$	$L{=}2$ shorthand: meta-steps; inner steps; per-episode horizon
$\mathcal{A}$, $\mathcal{M}$, $\mathcal{X}$	Inner-action, meta-action, inner-context spaces
$T_{k,j}$, $m_{k,j}$, $C_{k,j,i}$, $a_{k,j,i}$, $R_{k,j,i}$	Task context, meta-action, inner context, inner action, inner reward at episode k, meta-step j, inner step i
$x_{k,j,i} = (m_{k,j}, C_{k,j,i}, T_{k,j}) \sim P_X$	Inner-step context tuple; its marginal distribution
$\mathcal{D}_k$, $\mathcal{D}_{\mathrm{off}}$	Episode-k online data; pre-online offline data
$\mathcal{D}_{\mathrm{prior}}$, $\mathcal{D}_{\mathrm{cert}}$	Partition of $\mathcal{D}_{\mathrm{off}}$ into prior-construction and certification portions (Appendix D, Effect 4)
Policies, value, off-policy estimators.	
μ_k, π_k, π_Q	Behaviour, target, PAC-Bayes-induced policy
$w_t = \pi(a_t \mid x_t)/\mu_k(a_t \mid x_t)$	Importance weight (Assumption 4)
$R_{\max}$, $B_w^{(\ell)}$, $B^{\mathcal{I}}$	Reward bound; per-level weight bound; loss bound $B^{\mathcal{I}} \le R_{\max}\prod_{\ell\in\mathcal{I}} B_w^{(\ell)}$
$V(\pi;\theta^*)$	Policy value under true parameters
$\widehat{V}_k^{\mathrm{IPS}}$, $\widehat{V}_{\mathrm{pes}}^{\mathrm{LS}}$	IPS / LS estimators (Eq. (5))
$\mathcal{L}_t(\pi)$, $c_t = B^{\mathcal{I}} - \mathcal{L}_t$	Per-step loss; per-step cost
PAC-Bayes machinery.	
$\mathrm{KL}(P\|Q)$	KL divergence
p_0, Q, p_k	Reference prior, data-free with respect to $\mathcal{D}_{\mathrm{cert}}$ (Bayesian update on $\mathcal{D}_{\mathrm{prior}}$); generic PAC-Bayes posterior; episode-k Bayesian posterior
$\lambda > 0$	PAC-Bayes temperature
$(\mathcal{F}_t)_{t\ge0}$, $(\mathcal{H}_k)_{k\ge0}$	Inner-step filtration; episode-level filtration ($\mathcal{H}_k = \mathcal{F}_{kJI} \vee \sigma(\theta^*, \mathcal{G}, \mathcal{D}_{\mathrm{prior}})$)
F	Fisher information (D-optimal design)

B Instantiation Example: $L = 2$

The body's experiments run at $L = 2$ unless stated otherwise, with the shorthand $T = C^{(2)}$, $M = A^{(2)}$, $C = C^{(1)}$, $A = A^{(1)}$, $J = I_2$, $I = I_1$, $N_2 = JI$ summarised in the notation table (Appendix A) and pictured in Figure 1a. The example below records the explicit structural equations, the latent-confounder absorption, and the mechanism class; it is not used in any proof and may be skipped on a first reading.

Example 1 (Linear / nonlinear ANM at $L = 2$)**.** The mechanism-generated C, Y, R obey

$$C_{k,j,i} = f_C(m_{k,j}, T_{k,j};\, \theta^*_C) + \varepsilon'_{C,k,j,i}, \tag{10}$$

$$Y_{k,j,i} = f_Y(a_{k,j,i}, C_{k,j,i};\, \theta^*_Y) + \varepsilon'_{Y,k,j,i}, \tag{11}$$

$$R_{k,j,i} = f_R(Y_{k,j,i}, C_{k,j,i}, m_{k,j}, T_{k,j}, a_{k,j,i};\, \theta^*_R) + \varepsilon'_{R,k,j,i}, \tag{12}$$

where $\varepsilon'_{X,k,j,i} = \delta_{UX}\, U_{k,j,i} + \varepsilon_{X,k,j,i}$ realises Assumption 1 via additive loadings δ_{UX} on a Bernoulli inner-level confounder $U_{k,j,i} \sim \mathrm{Ber}(\lambda_U)$ shared across $\{C, Y, R\}$. The mechanisms fall within the additive noise model (ANM) family [24]: *Case A1* is linear ANM with Gaussian noise, admitting NIG-conjugate posteriors; *Case A2* is nonlinear ANM with additive noise, treated by the online

Gibbs posterior of Haddouche and Guedj [7]. Identifiability of $V(\pi;\theta^*)$ from logged data uses the conditional expectations $\mathbb{E}[C \mid M,T]$, $\mathbb{E}[Y \mid A,C]$, $\mathbb{E}[R \mid Y,C,M,T,A]$, all available under Assumption 3.

Example 2 (NCTS recursion at $L=2$)**.** We unroll the recursion of Eqs. (1) and (2) at $L=2$ for the algorithm and experiments. At $L=2$, the augmented contexts collapse to $\tilde{c}^{(2)} = T$ and $\tilde{c}^{(1)} = (T,M,C) =: x$.

Innermost-level closed form (independent of L). At $\ell=1$, Eq. (2) reduces to $\hat{R}^{(1)}_{\mathcal{I}}(a,\tilde{c}^{(1)};\theta) = \mathbb{E}[R \mid A^{(1)}=a,\tilde{c}^{(1)};\theta]$, and the level-1 argmax of Eq. (1) is the best-action map

$$\pi_k(\tilde{c}^{(1)}) = \arg\max_{a\in\mathcal{A}^{(1)}} \mathbb{E}\big[R \mid A^{(1)}=a,\ \tilde{c}^{(1)};\ \theta^{(k)}\big]. \tag{13}$$

This applies at every L: the innermost level is always a contextual bandit whose context is the augmented context $\tilde{c}^{(1)}$. The expectation unfolds via SCM forward-propagation through f_Y then f_R:

$$\mathbb{E}\Big[R \mid \mathbf{do}(A^{(1)}{=}a),\ \tilde{c}^{(1)};\ \theta^{(k)}\Big] = \mathbb{E}_{\varepsilon'_Y,\varepsilon'_R}\left[f_R\big(\underbrace{f_Y(a,C^{(1)};\theta^{(k)}_Y)+\varepsilon'_Y}_{=:\hat{Y}},\ \tilde{c}^{(1)},\ a;\ \theta^{(k)}_R\big) + \varepsilon'_R \right]. \tag{14}$$

At $L=2$, $\tilde{c}^{(1)} = (T,M,C)$ and the parents of R are (Y,C,M,T,A) as in Eq. (12). For linear ANMs with Gaussian noise, (14) is closed-form and the $\arg\max$ in (13) reduces to evaluating $|\mathcal{A}^{(1)}|$ linear-quadratic expressions; for nonlinear ANMs the expectation is evaluated by Monte Carlo with n_{mc} draws. Equation (14) integrates the residuals marginally, which is exact when they are independent of $\tilde{c}^{(1)}$. If a latent $U^{(1)}$ is shared with $C^{(1)}$ (Remark 1), the expectation must instead be conditional on $\tilde{c}^{(1)}$: the policy then also depends on the mechanism of $C^{(1)}$, outside $\Phi_{\mathcal{I}}$ of Definition 2. Theorem 1 and Lemma 3 then hold with $\Phi_{\mathcal{I}} \cup \{C^{(1)}\}$ in place of $\Phi_{\mathcal{I}}$, and the KL sum charges that mechanism. The marginal form keeps $\Phi_{\mathcal{I}}$ but need not be optimal; reading $\pi^{\mathcal{I}}_{\theta^*}$ as the best policy in scope (Appendix C.5) needs the conditional form. Corollary 5 is relative to $\pi^{\mathcal{I}}_{\theta^*}$ in either case.

Outer-level argmax at $L=2$. At $\ell=2$, the agent observes $T_{k,j}$ ($\tilde{c}^{(2)} = T_{k,j}$), and Eq. (2) integrates over the freshly drawn $C \sim P_C(\cdot \mid T_{k,j}, m, \theta^{(k)}_C)$. With both levels in $\mathcal{I}$, the $\ell{-}1 \in \mathcal{I}$ case of (2) applies, giving the meta-action argmax

$$m_{k,j} = \arg\max_{m\in\mathcal{M}} \mathbb{E}_{C\sim P_C(\cdot\mid T_{k,j},m,\theta^{(k)}_C)}\big[\max_a \mathbb{E}[R \mid A^{(1)}{=}a, C, m, T_{k,j};\theta^{(k)}]\big]. \tag{15}$$

D-optimal alternative. A Bayesian experimental-design alternative [4] replaces the Thompson $\arg\max$ at any chosen level ℓ by an information-gain $\arg\max$,

$$a^{(\ell),\star}_k = \arg\max_{a\in\mathcal{A}^{(\ell)}} \mathbb{E}_{p_{k-1}}\Big[\log\det F\big(\theta;\ \mathbf{do}(A^{(\ell)}{=}a),\ \tilde{c}^{(\ell)}\big)\Big],$$

where F is the posterior Fisher information. This is tractable in closed form under the factorised posterior (Section 5.2) for linear Gaussian ANMs at any L. Specialised to the meta level at $L=2$ ($\ell=2$, $a=m$, $\tilde{c}^{(\ell)} = T_{k,j}$), this gives the $L=2$ D-optimal meta-rule. Proving that the D-optimal variant accelerates posterior contraction beyond the Thompson sampling baseline of (15) is an open problem.

C Full Proofs of Theorem 1 and Corollary 5

We prove Theorem 1 in four steps and Corollary 5 in Appendix C.5. Throughout this appendix, we work in the *off-policy startup phase* (episodes $1,\dots,K$ under a fixed legacy policy $\mu = (\mu_M,\mu_A)$). In this phase the meta-action is supplied by the legacy μ_M rather than chosen by the agent, and together with the i.i.d. meta-contexts $T_{k,j} \sim P_T$ (Assumption 2) this makes the per-episode IPS estimates target a common value $V(\pi;\theta^*)$.

Filtration conventions. Throughout the main text, $(\mathcal{F}_t)_{t\ge0}$ denotes the inner-step-level filtration accumulating all observations up to the t-th inner step (flattening $t=(k-1)JI+(j-1)I+i$). In this appendix we work at the *episode level*: $(\mathcal{H}_k)_{k\ge0}$ (Eq. (16)) is the σ-algebra generated by θ^*, $\mathcal{G}$, $\mathcal{D}_{\mathrm{prior}}$, and all episode data $\mathcal{D}_1,\dots,\mathcal{D}_k$. The two are related by $\mathcal{H}_k = \mathcal{F}_{kJI} \vee \sigma(\theta^*,\mathcal{G},\mathcal{D}_{\mathrm{prior}})$. Under Assumption 4(iii), $\mathcal{L}^{\mathcal{I}}_t(\pi) \in [0,B^{\mathcal{I}}]$ and $\widehat{V}^{\mathrm{IPS}}_k(\pi) \in [0,B^{\mathcal{I}}]$.

C.1 Step 1: Episode-Level Filtration and Unbiasedness

Define the between-episode filtration:

$$\mathcal{H}_0 := \sigma(\theta^*, \mathcal{G}, \mathcal{D}_{\text{prior}}), \quad \mathcal{H}_k := \sigma\big(\mathcal{H}_0, \mathcal{D}_1, \ldots, \mathcal{D}_k\big), \quad k \geq 1, \tag{16}$$

where $\mathcal{D}_{\text{prior}} \subseteq \mathcal{D}_{\text{off}}$ is the portion of offline data used to construct p_0 (Eq. (27); $\mathcal{D}_{\text{prior}} = \emptyset$ recovers the data-free hyper-prior special case discussed in Appendix D, Effect 4); and the LS-summed episodes $\mathcal{D}_1, \ldots, \mathcal{D}_K$ comprise both the certification portion $\mathcal{D}_{\text{cert}} = \mathcal{D}_{\text{off}} \setminus \mathcal{D}_{\text{prior}}$ (re-indexed as offline records) and any subsequent online episodes. By construction, the data-free prior p_0 is $\mathcal{H}_0$-measurable (since it is a deterministic function of $\mathcal{D}_{\text{prior}}$ via Eq. (27)); in the off-policy phase the $\mathcal{D}_k$ ($k \geq 1$) are independent of one another given $\mathcal{H}_0$ (Assumptions 1 and 2).

Lemma 2 (Per-sample and per-episode unbiasedness)**.** *Fix any intervention scope $\mathcal{I} \subseteq \{1, \ldots, L\}$. Under Assumptions 1–3, for every θ, every episode k, and every within-episode multi-index $\mathbf{i} \in [N_L]$:*

$$\mathbb{E}\big[\mathcal{L}^{\mathcal{I}}_{k,\mathbf{i}}(\pi^{\mathcal{I}}_\theta) \mid \mathcal{H}_{k-1}\big] \;=\; \mathbb{E}\big[\bar{\mathcal{L}}^{\mathcal{I}}_k(\pi^{\mathcal{I}}_\theta) \mid \mathcal{H}_{k-1}\big] \;=\; V(\pi^{\mathcal{I}}_\theta; \theta^*), \tag{17}$$

where $\bar{\mathcal{L}}^{\mathcal{I}}_k := \frac{1}{N_L} \sum_{\mathbf{i} \in [N_L]} \mathcal{L}^{\mathcal{I}}_{k,\mathbf{i}}$ is the episode-mean loss. Moreover, the episode-mean estimates $\{\bar{\mathcal{L}}^{\mathcal{I}}_k(\pi^{\mathcal{I}}_\theta)\}_{k=1}^K$ are conditionally independent given $\mathcal{H}_0$.

Remark 4 (Why the lemma does *not* claim within-episode independence)**.** *The samples $\{\mathcal{L}^{\mathcal{I}}_{k,\mathbf{i}}\}_{\mathbf{i} \in [N_L]}$ within episode k are* not *mutually conditionally independent given $\mathcal{H}_{k-1}$ alone, because they share within-episode parent-level draws (e.g. at $L = 2$, all I_1 inner-step samples within meta-step j share $(T_{k,j}, m_{k,j})$). Their conditional means coincide regardless: $\mathbb{E}[\mathcal{L}^{\mathcal{I}}_{k,\mathbf{i}} \mid \mathcal{H}_{k-1}]$ averages over the random within-episode parent draws under P_X and yields the same value $V(\pi^{\mathcal{I}}_\theta; \theta^*)$ for every $\mathbf{i}$. The episode-level supermartingale of Step 3 only uses this constancy of conditional means, not within-episode independence; cross-episode independence given $\mathcal{H}_0$ gives the constant conditional mean given $\mathcal{H}_{k-1}$, which is all the supermartingale uses.*

Proof. **Sequential factorisation of the trajectory.** Within one episode, the agent draws actions level by level from $\ell = L$ down to $\ell = 1$ (Section 5.1); each action $a^{(\ell)}$ is sampled given the augmented context $\tilde{c}^{(\ell)}$, which records all higher-level decisions and exogenous/SCM-generated contexts at and above level ℓ. The chain rule along this fixed level-ordering then factorises the joint trajectory under the legacy behaviour $\mu = (\mu^{(\ell)})_{\ell=1}^L$ as

$$\begin{aligned} P_{\text{legacy}}(c^{(1:L)}, a^{(1:L)}, R) = \prod_{\ell=1}^{L} & p\big(c^{(\ell)} \mid \tilde{c}^{(\ell+1)}, a^{(\ell+1)}; \theta^*\big)\, \mu^{(\ell)}\big(a^{(\ell)} \mid \tilde{c}^{(\ell)}\big) \\ & \cdot p\big(R \mid a^{(1:L)}, \text{contexts}; \theta^*\big). \end{aligned}$$

The hybrid distribution under $\pi^{\mathcal{I}}_\theta$ replaces the policy factor $\mu^{(\ell)}$ by $\pi^{(\ell)}_\theta$ at every $\ell \in \mathcal{I}$ while leaving every other factor, including the context kernels, unchanged (Assumption 1):

$$\begin{aligned} P_{\text{hybrid}}(c^{(1:L)}, a^{(1:L)}, R) = \prod_{\ell=1}^{L} & p\big(c^{(\ell)} \mid \tilde{c}^{(\ell+1)}, a^{(\ell+1)}; \theta^*\big) \prod_{\ell \in \mathcal{I}} \pi^{(\ell)}_\theta \prod_{\ell \notin \mathcal{I}} \mu^{(\ell)} \\ & \cdot p(R \mid a^{(1:L)}, \text{contexts}; \theta^*). \end{aligned}$$

The Radon–Nikodym derivative is therefore $dP_{\text{hybrid}}/dP_{\text{legacy}} = \prod_{\ell \in \mathcal{I}} \pi^{(\ell)}_\theta / \mu^{(\ell)} = w^{\mathcal{I}}_{k,\mathbf{i}}$, with the legacy/legacy factors at $\ell \notin \mathcal{I}$ cancelling and the reward factor cancelling because both joints share the same $p(R \mid a^{(1:L)}, \text{contexts}; \theta^*)$. The level-wise overlap clause of Assumption 3 on $\mathcal{I}$ ensures these ratios are well defined.

Causal identification of the reward expectation. The reward factor $p(R \mid a^{(1:L)}, \text{contexts}; \theta^*)$ is the same conditional under both joints, but the conditional expectation $\mathbb{E}[R \mid a^{(1:L)}, \text{contexts}; \theta^*]$ taken under the legacy joint is the *interventional* reward $\mathbb{E}_{\text{hybrid}}[R]$ only when each $a^{(\ell)}$ is identifiable as a do-intervention from observed quantities. The level-wise backdoor admissibility (Assumption 3) supplies this identification at every $\ell \in \mathcal{I}$. Combining,

$$\mathbb{E}\big[\mathcal{L}^{\mathcal{I}}_{k,\mathbf{i}} \mid \mathcal{H}_{k-1}\big] \;=\; \mathbb{E}\big[w^{\mathcal{I}}_{k,\mathbf{i}} \cdot R_{k,\mathbf{i}} \mid \mathcal{H}_{k-1}\big] \;=\; \mathbb{E}_{\text{hybrid}}[R_{k,\mathbf{i}}] \;=\; V(\pi^{\mathcal{I}}_\theta; \theta^*),$$

where the conditional expectation is taken under the marginal of the within-episode SCM-generated parent contexts given $\mathcal{H}_{k-1}$. Linearity of expectation gives the same identity for $\bar{\mathcal{L}}_k^{\mathcal{I}}$, proving (17). Cross-episode conditional independence given $\mathcal{H}_0$ holds because each $\mathcal{D}_k$ is generated from the same fixed (P_X, μ, θ^*) in the off-policy phase and is independent of $\mathcal{D}_{k'}$ for $k' \neq k$ given $\mathcal{H}_0$ (Assumptions 1 and 2). The episode-mean estimates $\bar{\mathcal{L}}_k^{\mathcal{I}}$ are deterministic functions of $\mathcal{D}_k$, hence inherit this independence. □

C.2 Step 2: Causal KL Decomposition

Lemma 3 (Causal KL decomposition)**.** *Let $Q = \prod_{X\in\Phi} Q_X$ and $p_0 = \prod_{X\in\Phi} p_{0,X}$ be product distributions over $\theta_\Phi = \bigtimes_{X\in\Phi} \theta_X$, induced by the per-mechanism factorisation. Then:*

$$\mathrm{KL}(Q\|p_0) = \sum_{X\in\Phi} \mathrm{KL}(Q_X\|p_{0,X}). \tag{18}$$

For any intervention scope $\mathcal{I} \subseteq \{1,\dots,L\}$, under the convention $Q_X = p_{0,X}$ for $X \notin \Phi_{\mathcal{I}}$, the chain rule collapses to a sum over the post-intervention learnable mechanisms:

$$\mathrm{KL}(Q\|p_0) = \sum_{X\in\Phi_{\mathcal{I}}} \mathrm{KL}(Q_X\|p_{0,X}). \tag{19}$$

Proof. KL divergence is additive over product measures (both sides are $+\infty$ unless $Q_X \ll p_{0,X}$ for every X): $\mathrm{KL}(Q\|p_0) = \int \log \prod_X (dQ_X/dp_{0,X}) \prod_\beta dQ_\beta = \sum_X \int \log(dQ_X/dp_{0,X})\, dQ_X = \sum_X \mathrm{KL}(Q_X\|p_{0,X})$. The posterior factorises as $p_k(\theta) = \prod_{X\in\Phi} p_k(\theta_X)$ because (i) the prior factorises by assumption (Section 5.2), a choice that mechanism autonomy justifies [23], and (ii) the likelihood of the mechanism-factorised working model factorises by construction (Section 5.2). For the restricted form (19), the convention $Q_X = p_{0,X}$ for $X \notin \Phi_{\mathcal{I}}$ gives $\mathrm{KL}(Q_X\|p_{0,X}) = 0$ for those terms, leaving only the sum over $\Phi_{\mathcal{I}}$. The policy $\pi_\theta^{\mathcal{I}}$ depends on θ only through $\theta_{\Phi_{\mathcal{I}}}$, since Eq. (2) integrates only over contexts below $\ell_{\mathcal{I}}^\star$; a posterior over Φ is therefore certified through its $\Phi_{\mathcal{I}}$-marginal. The exception is $\ell_{\mathcal{I}}^\star = 1$ with a latent $U^{(1)}$ shared with $C^{(1)}$: the conditional policy then also reads the mechanism of $C^{(1)}$ (Appendix B). □

C.3 Step 3: Logarithmic Smoothing Supermartingale

We adapt the proof of Haddouche and Sakhi [8, Thm. 3.2] from within-deployment rounds to between-episode samples. The key adaptation is that we apply the LS construction to the *episode-mean* loss $\bar{\mathcal{L}}_k^{\mathcal{I}}$ (Lemma 2), not the per-sample stream $\mathcal{L}_{k,\mathbf{i}}^{\mathcal{I}}$; this side-steps the within-episode dependence highlighted in Remark 4.

Setup. For $\lambda > 0$ and every θ, define the per-episode gap $\Delta\bar{\mathcal{L}}_k^{\mathcal{I}}(\pi_\theta) := B^{\mathcal{I}} - \bar{\mathcal{L}}_k^{\mathcal{I}}(\pi_\theta) \in [0, B^{\mathcal{I}}]$ (since $\bar{\mathcal{L}}_k^{\mathcal{I}} \in [0, B^{\mathcal{I}}]$ by Assumption 4(iii)). True gap $\overline{\Delta\bar{\mathcal{L}}}(\pi_\theta;\theta^*) := B^{\mathcal{I}} - V(\pi_\theta;\theta^*)$. For each episode k, define

$$Y_k(\theta) := -\log\bigl(1 - \lambda\, \overline{\Delta\bar{\mathcal{L}}}(\pi_\theta;\theta^*)\bigr) + \log\bigl(1 - \lambda\, \Delta\bar{\mathcal{L}}_k^{\mathcal{I}}(\pi_\theta)\bigr), \tag{20}$$

and $f_K(\theta) := \sum_{k=1}^K Y_k(\theta)$. Under Assumption 4 and the fixed $\lambda \in (0, 1/B^{\mathcal{I}})$ of Theorem 1, both logarithms are well-defined.

Lemma 4 (Episode-level LS supermartingale)**.** *Define $M_k := \mathbb{E}_{\theta\sim p_0}[\exp(f_k(\theta))]$ for $k \geq 1$ and $M_0 := 1$. Under Assumption 4 and Lemma 2, $(M_k)_{k\geq 0}$ is a non-negative supermartingale w.r.t. $(\mathcal{H}_k)_{k\geq 0}$.*

Proof. We show $\mathbb{E}[M_k \mid \mathcal{H}_{k-1}] \leq M_{k-1}$. Since p_0 is $\mathcal{H}_0$-measurable (Eq. (16); p_0 depends only on $\mathcal{D}_{\text{prior}}$, and $\mathcal{H}_0 \subseteq \mathcal{H}_{k-1}$), the law p_0 is fixed under conditioning on $\mathcal{H}_{k-1}$, so Fubini gives $\mathbb{E}[M_k \mid \mathcal{H}_{k-1}] = \mathbb{E}_{\theta\sim p_0}[e^{f_{k-1}(\theta)}\mathbb{E}[e^{Y_k(\theta)} \mid \mathcal{H}_{k-1}]]$ (noting that $f_{k-1}(\theta)$ is $\mathcal{H}_{k-1}$-measurable in ω and jointly Borel in (θ,ω), the argmax policies of Eq. (1) being jointly measurable in $(\theta, \tilde{c}^{(\ell)})$ under any fixed measurable tie-breaking rule (we adopt the smallest maximising index on the finite grid

$\mathcal{A}^{(\ell)}$), and that the integrand is non-negative and bounded by $(1-\lambda B^{\mathcal{I}})^{-k}$, so Fubini applies and M_k is integrable). The exponent $Y_k(\theta)$ is a single log-ratio, so

$$\mathbb{E}\big[e^{Y_k(\theta)} \mid \mathcal{H}_{k-1}\big] \;=\; \frac{\mathbb{E}\big[1-\lambda\,\Delta\bar{\mathcal{L}}_k^{\mathcal{I}}(\pi_\theta) \,\big|\, \mathcal{H}_{k-1}\big]}{1-\lambda\,\overline{\Delta\bar{\mathcal{L}}}(\pi_\theta;\theta^*)}. \tag{21}$$

By Lemma 2, $\mathbb{E}[\bar{\mathcal{L}}_k^{\mathcal{I}}(\pi_\theta) \mid \mathcal{H}_{k-1}] = V(\pi_\theta;\theta^*)$, hence $\mathbb{E}[\Delta\bar{\mathcal{L}}_k^{\mathcal{I}}(\pi_\theta) \mid \mathcal{H}_{k-1}] = B^{\mathcal{I}} - V(\pi_\theta;\theta^*) = \overline{\Delta\bar{\mathcal{L}}}(\pi_\theta;\theta^*)$. The numerator equals $1-\lambda\,\overline{\Delta\bar{\mathcal{L}}}(\pi_\theta;\theta^*)$, the ratio is 1, and $\mathbb{E}[M_k \mid \mathcal{H}_{k-1}] = M_{k-1}$, which is in particular $\le M_{k-1}$.[2] □

C.4 Step 4: Ville's Inequality and Change of Measure

By Ville's inequality, which is Markov's inequality applied to the stopped supermartingale [34, Secs. 6.4 and 10.9], since (M_k) is a non-negative supermartingale with $M_0 = 1$:

$$\Pr\big[\exists\, k \ge 1 : M_k \ge 1/\delta\big] \le \delta. \tag{22}$$

On $\{M_K < 1/\delta\}$, apply Donsker–Varadhan: for any posterior Q,

$$\mathbb{E}_{\theta\sim Q}[f_K(\theta)] \le \mathrm{KL}(Q\|p_0) + \log M_K \le \mathrm{KL}(Q\|p_0) + \log(1/\delta). \tag{23}$$

Expanding $\mathbb{E}_Q[f_K(\theta)]$ via the definition of $Y_k(\theta)$:

$$\mathbb{E}_Q[f_K(\theta)] = -K\,\mathbb{E}_Q\big[\log(1-\lambda\,\overline{\Delta\bar{\mathcal{L}}}(\pi_\theta;\theta^*))\big] + \sum_{k=1}^{K}\mathbb{E}_Q\big[\log(1-\lambda\,\Delta\bar{\mathcal{L}}_k^{\mathcal{I}}(\pi_\theta))\big]. \tag{24}$$

Jensen's inequality applied to the convex $u \mapsto -\log(1-\lambda u)$ under the pushforward of $\theta \mapsto B^{\mathcal{I}} - V(\pi_\theta;\theta^*)$, with the mixture convention $V(\pi_Q^{\mathcal{I}};\theta^*) = \mathbb{E}_Q[V(\pi_\theta^{\mathcal{I}};\theta^*)]$, gives $-\mathbb{E}_Q[\log(1-\lambda\,\overline{\Delta\bar{\mathcal{L}}}(\pi_\theta;\theta^*))] \ge -\log(1-\lambda\,(B^{\mathcal{I}} - V(\pi_Q^{\mathcal{I}};\theta^*)))$. Combining with (23), dividing by λK, and using $-\log(1-\lambda x)/\lambda \ge x$ for $\lambda x < 1$:

$$V(\pi_Q^{\mathcal{I}};\,\theta^*) \ge \widehat{V}_{\mathrm{pes}}^{\mathrm{LS}}(\pi_Q^{\mathcal{I}}) - \frac{\sum_{X\in\Phi_{\mathcal{I}}}\mathrm{KL}(Q_X\|p_{0,X}) + \log(1/\delta)}{\lambda K}, \tag{25}$$

with $\widehat{V}_{\mathrm{pes}}^{\mathrm{LS}}(\pi_Q^{\mathcal{I}}) = B^{\mathcal{I}} + \frac{1}{\lambda K}\sum_{k=1}^{K}\mathbb{E}_Q[\log(1-\lambda\,\Delta\bar{\mathcal{L}}_k^{\mathcal{I}}(\pi_\theta))]$ as in Eq. (5), and we used the causal KL decomposition (Lemma 3) to replace $\mathrm{KL}(Q\|p_0)$ by $\sum_{X\in\Phi_{\mathcal{I}}}\mathrm{KL}(Q_X\|p_{0,X})$. The bound holds simultaneously for all factorised Q and all $K \ge 1$ (time-uniform, inherited from Ville).

C.5 Excess-Risk Bound: Proof of Corollary 5

Fix the scope $\mathcal{I}$ and write $\theta = (\theta_X)_{X\in\Phi_{\mathcal{I}}}$. By backward induction on Eq. (2) (in the conditional form of Appendix B when $U^{(1)}$ is shared with $C^{(1)}$), $\pi_{\theta^*}^{\mathcal{I}}$ maximises $V(\cdot\,;\theta^*)$ among all policies that follow $\mu^{(\ell)}$ at every $\ell \notin \mathcal{I}$; for $\mathcal{I} = \{1,\dots,L\}$ it is the optimal policy $\pi^* := \arg\max_\pi V(\pi;\theta^*)$ on the true environment. For $K \ge 1$ let $\lambda_K := 1/(2B^{\mathcal{I}}\sqrt{K})$, with $B^{\mathcal{I}}$ the assumption-level constant of Assumption 4(iii), so that λ_K is data-free and $\lambda_K B^{\mathcal{I}} \le 1/2$. Let $\hat{Q}_K$ (written $\hat{Q}$ when K is clear) maximise $\widehat{V}_{\mathrm{pes}}^{\mathrm{LS}}(\pi_Q^{\mathcal{I}}) - \sum_{X\in\Phi_{\mathcal{I}}}\mathrm{KL}(Q_X\|p_{0,X})/(\lambda_K K)$ over factorised Q, with the estimator evaluated at $\lambda = \lambda_K$ on the first K episodes; the proof uses only that this objective at $\hat{Q}_K$ is at least its value at the reference posterior Q_K^* below.

Corollary 5 (Excess risk of the certificate-maximising posterior)**.** *Assume the conditions of Theorem 1. Let each $p_{0,X}$ be a non-degenerate Gaussian $\mathcal{N}(\theta_{0,X},\Sigma_{0,X})$ on θ_X with $\dim\theta_X = d_X(\sum_{\beta\in\mathrm{Pa}(X)} d_\beta + 1)$. Suppose there is $C < \infty$ with $V(\pi_{\theta^*}^{\mathcal{I}};\theta^*) - V(\pi_\theta^{\mathcal{I}};\theta^*) \le C\|\theta-\theta^*\|$ for all θ. Then for every $\delta \in (0,1)$ there is a constant c, depending on δ, $B^{\mathcal{I}}$, C, p_0 and θ^* but not on K, such that with probability at least $1-\delta$, simultaneously for all $K \ge 1$,*

$$V(\pi_{\theta^*}^{\mathcal{I}};\theta^*) - V(\pi_{\hat{Q}_K}^{\mathcal{I}};\theta^*) \;\le\; c\,\frac{p_\Phi\log(K+1)}{\sqrt{K}}, \qquad p_\Phi = \sum_{X\in\Phi} d_X\Big(\textstyle\sum_{\beta\in\mathrm{Pa}(X)} d_\beta + 1\Big).$$

[2] The conditional expectation is in fact equal to M_{k-1}, so (M_k) is a martingale; we state the supermartingale property because it is what Ville's inequality (Step 4) requires. The argument uses only the constancy of the conditional mean $\mathbb{E}[\bar{\mathcal{L}}_k^{\mathcal{I}} \mid \mathcal{H}_{k-1}]$ across k, never within-episode independence (Remark 4). If overlap fails, $R \ge 0$ gives $\mathbb{E}[\bar{\mathcal{L}}_k^{\mathcal{I}} \mid \mathcal{H}_{k-1}] \le V(\pi_\theta;\theta^*)$, the ratio in (21) is at most 1, and (M_k) stays a supermartingale. This covers the deterministic legacy of Appendix E, where the weights are evaluated only on the support of μ and the bound is valid but capped.

The certificate-maximising posterior approaches the best policy in scope at this rate. For $\mathcal{I} = \{1, \dots, L\}$ that policy is the optimal policy π^* on the *realised* environment; for a smaller scope it keeps the legacy at $\ell \notin \mathcal{I}$. The certificate behind the bound holds uniformly over posteriors, and p_Φ counts parameters per mechanism, not jointly. The Lipschitz condition holds whenever $|V(\pi;\theta) - V(\pi;\theta^*)| \le \frac{C}{2}\|\theta - \theta^*\|$ uniformly over that policy class, because $\pi^{\mathcal{I}}_\theta$ is optimal under model θ. It requires no unique argmax.

Proof. Theorem 1 at $\hat{Q}_K$ and the optimality of $\hat{Q}_K$ reduce the excess risk to three terms at a fixed reference posterior Q^*_K near θ^*. The Lipschitz condition, Hoeffding's inequality and the Gaussian KL formula bound these terms, and a union bound over K gives simultaneity.

(i) Decomposition. Let $\delta_K := 3\delta/(\pi^2 K^2)$, so that $\sum_{K\ge1} 2\delta_K = \delta$, and let $L_K := \log(1/\delta_K) = \log(\pi^2/(3\delta)) + 2\log K$. Fix K and write $\lambda = \lambda_K$. Let $Q^*_K := \mathcal{N}(\theta^*, \sigma_K^2 I)$ with $\sigma_K := K^{-1/2}$, a factorised posterior in dimension $p_{\mathcal{I}} := \sum_{X\in\Phi_{\mathcal{I}}} \dim\theta_X \le p_\Phi$. Apply (25) to $\hat{Q}_K$ at temperature λ and confidence δ_K, and use the optimality of $\hat{Q}_K$:

$$V(\pi^{\mathcal{I}}_{\theta^*};\theta^*) - V(\pi^{\mathcal{I}}_{\hat{Q}_K};\theta^*) \le \underbrace{V(\pi^{\mathcal{I}}_{\theta^*};\theta^*) - V(\pi^{\mathcal{I}}_{Q^*_K};\theta^*)}_{\text{(a) approximation error}} + \underbrace{V(\pi^{\mathcal{I}}_{Q^*_K};\theta^*) - \widehat{V}^{\mathrm{LS}}_{\mathrm{pes}}(\pi^{\mathcal{I}}_{Q^*_K})}_{\text{(b) estimation error at } Q^*_K} + \underbrace{\frac{\sum_{X\in\Phi_{\mathcal{I}}} \mathrm{KL}(Q^*_{K,X}\|p_{0,X}) + L_K}{\lambda K}}_{\text{(c) regulariser}}. \tag{26}$$

(ii) Term (a). The Lipschitz condition and Jensen's inequality give (a) $\le C\,\mathbb{E}_{\theta\sim Q^*_K}\|\theta - \theta^*\| \le C\sigma_K\sqrt{p_{\mathcal{I}}} = C\sqrt{p_{\mathcal{I}}/K}$.

(iii) Term (b). Let $g_k := \mathbb{E}_{\theta\sim Q^*_K}[-\log(1-\lambda\,\Delta\bar{\mathcal{L}}^{\mathcal{I}}_k(\pi_\theta))]/\lambda$, so that $\widehat{V}^{\mathrm{LS}}_{\mathrm{pes}}(\pi^{\mathcal{I}}_{Q^*_K}) = B^{\mathcal{I}} - \frac{1}{K}\sum_{k=1}^K g_k$. Since $\lambda\,\Delta\bar{\mathcal{L}}^{\mathcal{I}}_k \in [0, 1/2]$, the inequality $-\log(1-u) \le u + u^2$ on $[0,1/2]$ and Lemma 2 give $\mathbb{E}[g_k \mid \mathcal{H}_{k-1}] \le B^{\mathcal{I}} - V(\pi^{\mathcal{I}}_{Q^*_K};\theta^*) + \lambda(B^{\mathcal{I}})^2$. The last term is the pessimism of the LS estimator. Moreover $g_k \in [0, 2\log 2\cdot B^{\mathcal{I}}]$, because $u \mapsto -\log(1-u)/u$ is nondecreasing, and the g_k are i.i.d. across k given $\mathcal{H}_0$. Hoeffding's inequality gives, with probability at least $1-\delta_K$, (b) $\le \lambda(B^{\mathcal{I}})^2 + B^{\mathcal{I}}\sqrt{L_K/K} = B^{\mathcal{I}}(\frac{1}{2} + \sqrt{L_K})/\sqrt{K}$.

(iv) Term (c). We have $1/(\lambda K) = 2B^{\mathcal{I}}/\sqrt{K}$. The Gaussian KL formula with $\sigma_K^2 = 1/K \le 1$ gives $\sum_{X\in\Phi_{\mathcal{I}}} \mathrm{KL}(Q^*_{K,X}\|p_{0,X}) \le c_0 + \frac{1}{2}p_{\mathcal{I}}\log K$, where $c_0 := \frac{1}{2}\sum_{X\in\Phi_{\mathcal{I}}}[\mathrm{tr}\,\Sigma^{-1}_{0,X} + (\theta^*_X - \theta_{0,X})^\top\Sigma^{-1}_{0,X}(\theta^*_X - \theta_{0,X}) + \log\det\Sigma_{0,X}]$.

(v) Summing up. The three bounds give

$$V(\pi^{\mathcal{I}}_{\theta^*};\theta^*) - V(\pi^{\mathcal{I}}_{\hat{Q}_K};\theta^*) \;\le\; \frac{1}{\sqrt{K}}\Big(C\sqrt{p_{\mathcal{I}}} + B^{\mathcal{I}}\big[\tfrac{1}{2} + \sqrt{L_K} + 2c_0 + 2L_K + p_{\mathcal{I}}\log K\big]\Big),$$

which is at most $c\,p_\Phi\log(K+1)/\sqrt{K}$, since $p_{\mathcal{I}} \le p_\Phi$ and $L_K = O(\log(K+1))$. For this K the bound fails with probability at most $2\delta_K$ (the certificate and Hoeffding). A union bound over $K \ge 1$ gives total failure probability at most δ. □

C.6 Discussion

Role of the causal factorisation. The KL regulariser in (25) sums over $|\Phi|$ terms, each scaling with $d_X(\sum_\beta d_\beta + 1)$. Without the causal factorisation, it would scale with the full joint count $\sum_{X\in\mathbf{X}} d_X(\sum_\beta d_\beta + 1)$. For $\mathcal{I} = \{1\}$ ($\Phi_{\{1\}} = \{Y, R\}$), the saving relative to $\Phi = \{C, Y, R\}$ is $d_C(d_M + d_T + 1)$ parameters (coefficients of the parent set $\{M, T\}$ plus an intercept, per output dimension); substantial when d_C, d_M, or d_T is large.

Remark 5 (NCCB episodes as nested deployment rounds)**.** *Theorem 3.2 of Haddouche and Sakhi [8] is stated for per-sample costs $c \in [-1, 0]$ in conditionally independent batches, with the smoothed term $-\frac{p}{\lambda}\log(1-\lambda c/q)$ and any $\lambda > 0$. We keep its proof template (log-ratio process, Fubini, Ville, change of measure) and change the summand: one term per episode, built from the episode-mean*

loss $\bar{\mathcal{L}}_k^{\mathcal{I}}$ of Lemma 2, with the logarithm applied to the gap $B^{\mathcal{I}} - \bar{\mathcal{L}}_k^{\mathcal{I}} \geq 0$. Eq. (5) *is therefore a variant of their estimator, not an instance of it, and the sign of the gap is why it needs $\lambda < 1/B^{\mathcal{I}}$. This form is a design choice, not a requirement of the proof: applying the logarithm to the loss itself, $\log(1 + \lambda \bar{\mathcal{L}}_k^{\mathcal{I}})$, gives a valid bound for every $\lambda > 0$ by the same steps, without Assumption 4(ii), and measures the second moment from zero rather than from $B^{\mathcal{I}}$. We leave this variant to future work. The N_L samples within an episode are* not *i.i.d. given $\mathcal{H}_{k-1}$ (they share within-episode parent draws; Remark 4), and the supermartingale construction does* not *require them to be: it operates on the episode mean, whose conditional expectation is the same scalar $V(\pi_\theta^{\mathcal{I}}; \theta^*)$ for every k (Lemma 2). The supermartingale property carries over because (i) the episode-mean is conditionally unbiased; (ii) μ and P_X are fixed across episodes in the off-policy phase, so $\bar{\mathcal{L}}_k^{\mathcal{I}}$ are conditionally i.i.d. given $\mathcal{H}_0$; and (iii) p_0 is data-free w.r.t. $\mathcal{D}_1, \ldots, \mathcal{D}_K$. We use a single global temperature $\lambda \in (0, 1/B^{\mathcal{I}})$ at the episode level; the accelerated "Adjusted LS" variant of their Sec. 4 bounds a difference of risks and is outside our scope. The within-episode samples are still useful: they reduce the variance of the empirical episode mean $\bar{\mathcal{L}}_k^{\mathcal{I}}$ around its conditional expectation, which tightens the empirical LS term in Eq.* (5) *(Jensen: aggregating samples before the log is tighter than averaging after). The adaptive online setting with $\mathcal{H}_{k-1}$-measurable μ_k is treated in Appendix C.7 below.*

C.7 Adaptive online extension and mixed-policy data

Theorem 1 treats μ as fixed across episodes, so the level-wise factor cancellation $\mu^{(\ell)}/\mu^{(\ell)} = 1$ at $\ell \notin \mathcal{I}$ in (3) is automatic. Two operational regimes break this assumption: *(i)* the online phase, where the agent acts under its own evolving policy from episode to episode; and *(ii)* progressive handover (Section 6.1), where successive certification rounds reuse data collected under partially-handover-deployed behaviour. A single adaptive corollary covers both under overlap.

Corollary 6 (Adaptive online extension)**.** *Let $(\mu_k)_{k \geq 1}$ be a sequence of behaviour policies (each $\mu_k = (\mu_k^{(\ell)})_{\ell=1}^L$ specifying level-wise action choices), with μ_k measurable w.r.t. $\mathcal{H}_{k-1}$ extended by the randomisation drawn before episode k (e.g., NCTS draws $\theta^{(k)} \sim p_{k-1}$ and acts via $\pi_{\theta^{(k)}}^{(\ell)}$ at every $\ell \in \mathcal{I}$, so $\theta^{(k)}$ joins the filtration before episode k). Suppose Assumption 3 holds with $\mu_k^{(\ell)}$ in place of $\mu^{(\ell)}$ for every k and every $\ell \in \mathcal{I}$, and $\sup_k \sup_{\theta,\, a, \tilde{c}^{(\ell)}} \pi_\theta^{(\ell)}(a \mid \tilde{c}^{(\ell)})/\mu_k^{(\ell)}(a \mid \tilde{c}^{(\ell)}) \leq B_w^{(\ell)}$ for every $\ell \in \mathcal{I}$ ($\mathcal{H}_{k-1}$-uniform per-level overlap). Then Theorem 1 holds with the per-sample weight $w_{k,\mathbf{i}}^{\mathcal{I},(\mu_k)}(\theta) = \prod_{\ell \in \mathcal{I}} \pi_\theta^{(\ell)}/\mu_k^{(\ell)}$ in place of $w_{k,\mathbf{i}}^{\mathcal{I}}$, and the corresponding episode-mean $\bar{\mathcal{L}}_k^{\mathcal{I},(\mu_k)} := \frac{1}{N_L} \sum_{\mathbf{i}} w_{k,\mathbf{i}}^{\mathcal{I},(\mu_k)}(\theta) R_{k,\mathbf{i}}$ in place of $\bar{\mathcal{L}}_k^{\mathcal{I}}$, certifying the per-episode-averaged value $\frac{1}{K} \sum_k V(\pi_Q^{\mathcal{I}}; \theta^*, \mu_k^{\mathcal{I}^c})$, where $\mu_k^{\mathcal{I}^c}$ is the behaviour at levels $\ell \notin \mathcal{I}$, and covering simultaneously:* (i) *the offline startup ($\mu_k = \mu$ fixed legacy) and* (ii) *any mixture across episodes of behaviour policies that satisfy the overlap condition. A deterministic online behaviour does not satisfy it for $\pi_\theta^{(\ell)} \neq \mu_k^{(\ell)}$: the ratio at the played action is $\pi_\theta^{(\ell)}(a \mid \tilde{c}^{(\ell)}) \in \{0, 1\}$, the identity of Lemma 2 becomes $\leq$ because $R \geq 0$, and the bound stays valid but credits only the episodes in which π_θ agrees with the played action. The bound certifies any Q, in particular $Q = p_K$ (final posterior) or $Q = p_k$ for any $k \leq K$ (anytime over K).*

Proof. Throughout, $\mathcal{H}_{k-1}$ includes the randomisation drawn before episode k. The key observation is that $V(\pi_\theta^{\mathcal{I}}; \theta^*, \mu_k^{\mathcal{I}^c})$ is $\mathcal{H}_{k-1}$-measurable. Lemma 2 extends to adaptive μ_k: the identity $\mathbb{E}[\mathcal{L}_{k,\mathbf{i}}^{\mathcal{I},(\mu_k)}(\pi_\theta^{\mathcal{I}}) \mid \mathcal{H}_{k-1}] = V(\pi_\theta^{\mathcal{I}}; \theta^*, \mu_k^{\mathcal{I}^c})$ holds because (a) μ_k is $\mathcal{H}_{k-1}$-measurable, so conditioning on $\mathcal{H}_{k-1}$ fixes μ_k; (b) the level-wise $\mu_k^{(\ell)}/\mu_k^{(\ell)}$ cancellation in the IPS step still produces $\sum_{a^{(\ell)}} \pi_\theta^{(\ell)}(a^{(\ell)} \mid \tilde{c}^{(\ell)})\, \mathbb{E}[R \mid a^{(\ell)}, \tilde{c}^{(\ell)}; \theta^*]$ at every $\ell \in \mathcal{I}$; (c) level-wise backdoor identification (Assumption 3 on $\mathcal{I}$) holds independently of μ_k. Linearity gives the same identity for $\bar{\mathcal{L}}_k^{\mathcal{I},(\mu_k)}$. The supermartingale construction (Lemma 4) and Ville's inequality go through unchanged with $f_K(\theta) := \sum_k Y_k(\theta)$ constructed from $\Delta\bar{\mathcal{L}}_k^{\mathcal{I},(\mu_k)} = B^{\mathcal{I}} - \bar{\mathcal{L}}_k^{\mathcal{I},(\mu_k)}$, with true gap $B^{\mathcal{I}} - V(\pi_\theta^{\mathcal{I}}; \theta^*, \mu_k^{\mathcal{I}^c})$ now depending on k. Step 4 then applies Jensen's inequality jointly over k and $\theta \sim Q$, which gives the averaged value. The $\mathcal{H}_{k-1}$-uniform per-level overlap together with Assumption 4(i) ensures $\bar{\mathcal{L}}_k^{\mathcal{I},(\mu_k)} \in [0, B^{\mathcal{I}}]$ and $1 - \lambda\Delta\bar{\mathcal{L}}_k^{\mathcal{I},(\mu_k)} > 0$ uniformly for $\lambda < 1/B^{\mathcal{I}}$. This is the discretised, mechanism-factorised, scope-aware analogue of the adaptive PAC-Bayes machinery of Haddouche and Guedj [7] with the logarithmic smoother of Haddouche and Sakhi [8, Thm. 3.2]. □

Mixed-policy data: data-generating scope $\neq$ deployment scope. Corollary 6 certifies the time-averaged value $\frac{1}{K}\sum_k V(\pi_Q^{\mathcal{I}}; \theta^*, \mu_k^{\mathcal{I}^c})$, with $\mu_k^{\mathcal{I}^c}$ the per-episode behaviour at uncontrolled levels. In progressive handover this matters: the cancellation $\mu^{(\ell)}/\mu^{(\ell)} = 1$ at $\ell \notin \mathcal{I}$ in (3) requires $\mathcal{D}_{\mathrm{cert}}$ to have been collected under μ_k at every uncontrolled level, but as earlier handover candidates are accepted, the $\mu_k^{\mathcal{I}^c}$ evolves and the corresponding factors no longer cancel; the full L-product weight is then required. The candidate $\pi_\theta^{(\ell)}$ at $\ell > 1$ depends on the deployment scope $\mathcal{I}$ via the recursive value definition (2): the level-ℓ greedy is computed under the assumption that descendants follow $\pi_\theta^{(\ell')}$ for $\ell' \in \mathcal{I}, \ell' < \ell$, and $\mu^{(\ell')}$ otherwise. Hence the data-generating scope (call it $\mathcal{I}_{\mathrm{data}}$) and the deployment scope $\mathcal{I}$ are not interchangeable, even when both are subsets of $[L]$. Reusing data across handover rounds therefore requires recomputing both $\pi_\theta^{(\ell)}$ (under the new $\mathcal{I}$) and the full level-wise IS weight (3).

Scope gap benchmark. On Γ-SCM the gap is large. The meta action is a direct structural parent of the inner context: $f_Q = 2\tanh(2M) + N_Q + 0.3U$ at the Γ-Stress preset ($\gamma = 1$, $\eta = 1$, $\sigma_Q = 3$). So $\mathbb{E}[Q \mid do(M)]$ moves from -1.86 to $+2.15$ across the seven-point meta grid, a shift of 1.34 standard deviations of Q, while $\mathrm{sd}[Q]$ stays invariant (20,000 draws per grid point). A level-1 certificate on full-legacy logs therefore bounds $V(\pi_Q^{\{1\}}; \theta^*, \mu^{(2)})$, the value of handing over the inner level alone, not the value of the composition $(\pi_Q^{(2)}, \pi_Q^{(1)})$. On the same environment, at $20{,}000$ Monte Carlo episodes per arm, the inner candidate is worth -2.337 per episode under the legacy meta action and $+1.145$ once the meta level is handed over as well, a gap of $+3.48$. The meta action alone changes the legacy value by -0.103, so the gap is almost pure interaction, which no single-level weight can express. For a grid-oracle inner policy the gap is $+23.2$ per episode ($+96.0$ against $+119.2$). The sign is not guaranteed. Here the composition exceeds the certified object, which costs power; with a different meta candidate the same interaction can make the certified object exceed the deployment, which costs validity. Per-level certificates therefore certify per-level handovers, and each further round is re-certified on data collected under the levels already handed over.

C.8 Practical computation of safe-deployment selection

The proof of Theorem 1 (Steps 1–4 of Appendix C) handles arbitrary intervention scope $\mathcal{I} \subseteq [L]$ uniformly: Lemma 2 establishes hybrid-IS unbiasedness via the level-factored Radon–Nikodym derivative under level-wise overlap (Assumption 3 on $\mathcal{I}$), Lemma 3 delivers the restricted KL chain rule via the convention $Q_X = p_{0,X}$ for $X \notin \Phi_{\mathcal{I}}$, and the LS supermartingale construction (Lemma 4) applies verbatim to the episode-mean gap $\Delta\bar{\mathcal{L}}_k^{\mathcal{I}} = B^{\mathcal{I}} - \bar{\mathcal{L}}_k^{\mathcal{I}}$, requiring $\lambda < 1/B^{\mathcal{I}}$.

Computation of (6)**.** For each candidate $\mathcal{I} \in \mathfrak{I}$, the bound is evaluated on the same $\mathcal{D}_{\mathrm{cert}}$ as Theorem 1, with three quantities recomputed: the per-sample weights $\{w_{k,\mathbf{i}}^{\mathcal{I}}\}$ (one product per record, aggregated to episode-means $\tilde{\mathcal{L}}_k^{\mathcal{I}}$), the loss bound $B^{\mathcal{I}}$ of Assumption 4(iii) (sets the upper limit on λ), and the relevant mechanism set $\Phi_{\mathcal{I}}$ (sets the KL summation). The optimal temperature in the leading-order expansion of (4) is closed-form,

$$\lambda_{\mathcal{I}}^\star \;=\; \sqrt{\frac{2\big(\sum_{X\in\Phi_{\mathcal{I}}} \mathrm{KL}(Q_X \| p_{0,X}) + \log(1/\delta)\big)}{K \cdot \widehat{S}_{\mathcal{I}}^2}}, \quad \widehat{S}_{\mathcal{I}}^2 := \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_Q\big[(B^{\mathcal{I}} - \bar{\mathcal{L}}_k^{\mathcal{I}})^2\big],$$

clipped to $\min(\lambda_{\mathcal{I}}^\star, c/B^{\mathcal{I}})$ for some $c < 1$ to stay strictly inside the domain of (4). On the prior split the sum runs over its episodes, K is the planned certification budget and the KL a declared budget. Enumerating $\mathfrak{I}$ needs no extra data, since $\mathcal{D}_{\mathrm{cert}}$ is shared; the cost of evaluating the bound grows linearly in $|\mathfrak{I}|$.

Remark 6 (Grid-uniform validity of the empirical temperature)**.** *The supermartingale of Lemma 4 needs one fixed λ. The temperatures $\lambda_{\mathcal{I}}^\star$ above and $\lambda_k^\star$ of Eq.* (9) *meet this when computed on the prior split, with a declared KL budget in place of $\mathrm{KL}(p_k \| p_0)$ and the planned certification budget in place of k, since they are then $\mathcal{H}_0$-measurable. Computed on the certification data, they do not: a union bound over a data-free grid closes this gap at $O(\log\log K)$ cost and keeps the anytime property. Throughout this remark, $B^{\mathcal{I}}$ is the assumption-level constant of Assumption 4(iii) (e.g. $R_{\max}\prod_{\ell\in\mathcal{I}} B_w^{(\ell)}$), and $\eta_{\mathrm{clip}} \in (0,1)$ is the clipping margin.*

Fix the data-free countable grid $\Lambda := \{\lambda_j = (1-\eta_{\mathrm{clip}})\,2^{-j}/B^{\mathcal{I}} : j \geq 0\}$ and the confidence weights $\delta_j := \delta \cdot 6/(\pi^2(j+1)^2)$, so that $\sum_{j\geq 0}\delta_j = \delta$. Every λ_j lies strictly inside $(0, 1/B^{\mathcal{I}})$, so Theorem 1

applies at each λ_j with confidence δ_j. A union bound gives Eq. (4) *simultaneously for all $j \geq 0$, all posteriors and all $K \geq 1$.*

Deploy the rounded temperature $\widehat{\lambda} := \max\{\lambda_j \in \Lambda : \lambda_j \leq \lambda^\star\}$. Geometric spacing gives $\widehat{\lambda} > \lambda^\star/2$, so the regulariser term at $\widehat{\lambda}$ exceeds its $\lambda^\star$-value by at most a factor 2. The pessimism term does not increase, because $u \mapsto -\log(1-u)/u$ is nondecreasing. The term $\log(1/\delta)$ grows to $\log(1/\delta_j) = \log(1/\delta) + 2\log(j+1) + \log(\pi^2/6)$.

The index j is logarithmically bounded. Since $S_k \leq B^{\mathcal{I}}$ and $C_k \geq \log(1/\delta)$, Eq. (9) *gives $\lambda_k^\star \geq \sqrt{2\log(1/\delta)}/(B^{\mathcal{I}}\sqrt{k})$, unless the clip is active, in which case $j = 0$. So $2^{-j} \geq \lambda^\star/(2(1-\eta_{\mathrm{clip}})/B^{\mathcal{I}})$ yields $j \leq \log_2\big((1-\eta_{\mathrm{clip}})\sqrt{k}/\sqrt{2\log(1/\delta)}\big) + 1 = O(\log k)$ for fixed $\delta \in (0,1)$, and hence $2\log(j+1) = O(\log\log k)$.*

The same construction applies to $\lambda_{\mathcal{I}}^\star$ at fixed K once the clip constant of the preceding paragraph is $c := 1 - \eta_{\mathrm{clip}}$, so that $\lambda_{\mathcal{I}}^\star \leq \lambda_0$. Combined with the enumeration of $\mathfrak{I}$, the confidence δ is replaced by $\delta/|\mathfrak{I}|$.

D How offline data and the legacy policy shape the bound

The bound interacts with two quantities a practitioner controls before the online phase: the data $\mathcal{D}_{\mathrm{off}}$ from which the reference prior p_0 is built, and the legacy controller μ under which $\mathcal{D}_{\mathrm{off}}$ was collected (and which acts during the off-policy startup phase). The reference prior p_0 in Theorem 1 must be data-free with respect to the data used in the LS estimator. It need not be uninformed. Concretely, the workflow constructs p_0 from a portion $\mathcal{D}_{\mathrm{prior}} \subseteq \mathcal{D}_{\mathrm{off}}$ of the offline data via a single Bayesian update,

$$p_0(\theta) \;\propto\; p_{\mathrm{hyper}}(\theta) \cdot \prod_{X \in \Phi} p\big(\mathcal{D}_{\mathrm{prior}}^{(X)} \mid \theta_X, \mathrm{Pa}(X)\big), \tag{27}$$

where p_{hyper} is any data-free factorised hyper-prior (e.g. the broad NIG of Appendix F). The remainder $\mathcal{D}_{\mathrm{cert}} := \mathcal{D}_{\mathrm{off}} \setminus \mathcal{D}_{\mathrm{prior}}$ enters the LS estimator alongside any online episodes $\mathcal{D}_1, \ldots, \mathcal{D}_K$ collected later under behaviour policies μ_k. Since p_0 is a deterministic function of $\mathcal{D}_{\mathrm{prior}}$ alone, it qualifies as the data-free reference prior of Theorem 1; the two roles (Bayesian prior for the NCTS update loop and PAC-Bayes reference for the certificate) are served by the same object. The data split between $\mathcal{D}_{\mathrm{prior}}$ and $\mathcal{D}_{\mathrm{cert}}$ is itself a design choice, taken up explicitly in Effect 4 below. We now examine how the size, alignment, and coverage of $\mathcal{D}_{\mathrm{off}}$ enter the certificate.

Effect 1: Precision via prior concentration. The regulariser term in Theorem 1 scales as $\mathrm{KL}(Q^*\|p_0)/(\lambda K)$. For Gaussian per-mechanism marginals $p_{0,X} = \mathcal{N}(\theta_{0,X}, \Sigma_{0,X})$ and concentrated $Q_X^* = \mathcal{N}(\theta_X^*, \sigma^2 I)$ with $\sigma \to 0$, the per-mechanism KL is

$$\mathrm{KL}(Q_X^*\|p_{0,X}) \;=\; \tfrac{1}{2}(\theta_X^* - \theta_{0,X})^\top \Sigma_{0,X}^{-1}(\theta_X^* - \theta_{0,X}) \;+\; O(\dim\theta_X \log(1/\sigma))\,. \tag{28}$$

The first term is a Mahalanobis distance from the prior centre $\theta_{0,X}$ to the truth θ_X^*, weighted by the inverse prior covariance $\Sigma_{0,X}^{-1}$. Two regimes: *(a) tight and well-aligned prior*, $\Sigma_{0,X}^{-1}$ large and $\theta_{0,X} \approx \theta_X^*$, gives a small KL and a tight certificate; *(b) tight but misaligned prior*, $\Sigma_{0,X}^{-1}$ large and $\theta_{0,X}$ far from θ_X^*, inflates the Mahalanobis distance and *loosens* the certificate. The asymmetry of the KL is the protection mechanism: a prior that confidently concentrates mass in the wrong place pays for that confidence in the bound.

Effect 2: Accuracy via legacy suboptimality. The bound certifies $V(\pi_{\hat{Q}}; \theta^*)$ in absolute terms, not relative to $V(\mu; \theta^*)$. The certified *gain* over the legacy controller is therefore

$$V(\pi_{\hat{Q}}; \theta^*) - V(\mu; \theta^*) \;\geq\; \widehat{V}_{\mathrm{pes}}^{\mathrm{LS}}(\pi_{\hat{Q}}) - \widehat{V}_K^{\mathrm{IPS}}(\mu) - \frac{\mathrm{KL}(\hat{Q}\|p_0) + \log(2/\delta)}{\lambda K} - R_{\max}\sqrt{\frac{\log(2/\delta)}{2K_\mu}}, \tag{29}$$

with probability at least $1-\delta$, where $\widehat{V}_K^{\mathrm{IPS}}(\mu)$ is the unbiased on-policy estimate of $V(\mu)$ under μ itself (computed as the empirical mean of the K_μ episode-mean rewards logged under μ), and the last term is its Hoeffding deviation; each of the two events is taken at confidence $\delta/2$. Only the first event is time-uniform: the Hoeffding term holds at a fixed K_μ, so a gate that re-checks the bound must spread that confidence over its checks, levels and epochs, as App. H.5 does. All terms are in reward

units shifted by R_{lo}, that is Assumption 4(i) before its division by the range: rewards lie in $[0, R_{\max}]$, so $R_{\max}$ is the width of the reward range and bounds every episode-mean reward. The per-episode totals in Section 7 multiply every term on the right-hand side by N_L. When the incumbent enters as a measured constant rather than an estimate (the value form used on the external benches), the Hoeffding term and the factor 2 inside the logarithm drop. The right-hand side is large precisely when μ is suboptimal and $\pi_{\hat{Q}}$ has been certified high. Theorem 1 thus delivers an *accuracy boost*: the worse the legacy, the larger the certifiable improvement, with no dependence on the gap entering the regulariser. The expected reward $V(\mu; \theta^*)$ does not appear in Theorem 1, only on the gain side of (29).

Effect 3: Coverage, not quality, of the legacy. Suboptimality of μ (Effect 2) is not the same as misbehaviour. The regulariser and the admissible range of λ depend on μ through (i) the overlap clause of Assumption 3 and (ii) the loss bound $B^{\mathcal{I}}$ of Assumption 4, never through $V(\mu)$; the estimate $\widehat{V}^{\mathrm{LS}}_{\mathrm{pes}}$ and its Jensen gap read the logged rewards, so S in Effect 5 does depend on μ. A uniformly random high-coverage legacy yields a tighter certificate than a near-optimal but concentrated one: the latter inflates $B^{\mathcal{I}}$, which shrinks the admissible range $\lambda < 1/B^{\mathcal{I}}$ and widens the bound, and it may violate overlap entirely. Certifying improvement therefore requires μ to be *informative*, not good; a feature of PAC-Bayesian off-policy guarantees, not a bug.

Effect 4: Pre-TS certification via data splitting. Theorem 1 requires $K \geq 1$ episode-mean records in the LS sum and a reference prior p_0 that is data-free with respect to those records. A practitioner with only offline data $\mathcal{D}_{\mathrm{off}}$ (no online interaction yet) reaches certification by data-splitting: partition $\mathcal{D}_{\mathrm{off}}$ (taken to comprise N offline episodes, each producing a within-episode batch of N_L samples) into a prior portion $\mathcal{D}_{\mathrm{prior}}$ (αN episodes) and a certification portion $\mathcal{D}_{\mathrm{cert}}$ ($(1-\alpha)N$ episodes, $\alpha \in [0, 1)$), construct p_0 from $\mathcal{D}_{\mathrm{prior}}$ via Eq. (27), and evaluate $\widehat{V}^{\mathrm{LS}}_{\mathrm{pes}}(\pi_Q)$ on $\mathcal{D}_{\mathrm{cert}}$ ($K = (1-\alpha)N$ episodes contributing one episode-mean each). Setting $Q = p_0$ collapses the regulariser to $\log(1/\delta)/(\lambda K)$ since $\mathrm{KL}(p_0 \| p_0) = 0$, which gives a pre-TS certificate with no KL cost; a $Q \neq p_0$ certifies more when its LS estimate gains more than its KL costs. The choice of α trades the quality of the prior against the certification sample $K = (1-\alpha)N$; we report no optimal α. The boundary case $\alpha = 0$ is data-free by construction: p_0 reduces to the broad NIG hyper-prior of Appendix F and the LS estimator uses all N episodes. This is the special case in which no offline information enters the prior, so the regulariser pays $\mathrm{KL}(Q \| p_{\mathrm{hyper}}) = O(p_\Phi \log N)$ for any informative Q learned from the data; loose relative to splitting, but applicable when offline records are scarce or the practitioner declines to pre-commit to a single split.

Effect 5: When the gate can fire (the certifiability frontier). Setting the expected right-hand side of (29) to zero gives, to leading order and for a fixed candidate and KL budget, the true gain at which the certificate fires with probability about one half at budget K, before any data. Write $x_k := B^{\mathcal{I}} - \bar{\mathcal{L}}^{\mathcal{I}}_k \in [0, B^{\mathcal{I}}]$. Since $\log(1-\lambda x) = -\lambda x - [-\log(1-\lambda x) - \lambda x]$, the LS estimator (5) is the plug-in IS estimate minus a Jensen gap,

$$\widehat{V}^{\mathrm{LS}}_{\mathrm{pes}} = \widehat{V}^{\mathrm{IS}} - D_K(\lambda), \qquad D_K(\lambda) := \frac{1}{\lambda K} \sum_{k=1}^{K} \big[-\log(1-\lambda x_k) - \lambda x_k \big] = \frac{\lambda}{2} S^2 + O(\lambda^2),$$

with $S^2 := \frac{1}{K} \sum_k x_k^2$ the second moment of the loss measured from the *top* of the declared range (the $\widehat{S}^2_{\mathcal{I}}$ of Section 6.1). The certified gain is therefore the *coverage deficit* $\widehat{\Delta} := \widehat{V}^{\mathrm{IS}}(\pi_{\hat{Q}}) - \widehat{V}^{\mathrm{IPS}}_K(\mu)$ minus $D_K(\lambda)$, the regulariser $C/(\lambda K)$ with $C := \mathrm{KL}(\hat{Q} \| p_0) + \log(2/\delta)$, and the Hoeffding term. At the interior minimiser $\lambda^\star = \sqrt{2C/(KS^2)}$ of the leading-order expression the two λ-dependent terms are equal and sum to $S\sqrt{2C/K}$, so a candidate whose true gain is G has a non-negative expected certified gain at K only if

$$G \;\geq\; G_{\min}(K) \;=\; S\sqrt{\frac{2C}{K}}\,\big(1 + O(K^{-1/2})\big) \;+\; R_{\max}\sqrt{\frac{\log(2/\delta)}{2K_\mu}} \;=\; \frac{c}{\sqrt{K}} \tag{30}$$

(for $K_\mu \propto K$; multiply by N_L for per-episode totals). The realised $\widehat{\Delta}$ fluctuates around G at the same order $1/\sqrt{K}$, so under a normal approximation $G_{\min}(K)$ is the median firing point, not a hard threshold; selecting $\hat{Q}$ on the same data can shift it. The constant c is fixed by the reward span and the overlap constant (through $S \leq B^{\mathcal{I}}$), by δ, and by a KL budget that is declarable when the

posterior is pinned. Three conditions qualify (30), and each corresponds to a failure we observed. *(C0)* $\lambda^\star$ must be interior, $K \geq 2C(1-\eta_{\text{clip}})^{-2}(B^{\mathcal{I}}/S)^2$; below this the cap of (9) binds and the requirement is larger. *(C1)* λ must fall with K by a licensed route: data-free constants, $\lambda^\star$ fitted on an independent split, or an a-priori geometric grid closed by a union bound (rounding multiplies the two λ-dependent terms by at most $5/4$ and adds $O(\log\log K)$ to C). At a fixed λ_0 the Jensen gap $D(\lambda_0)$ does not shrink with K, so $G_{\min}(K)$ converges to a positive floor, which is the plateau of fixed-λ runs. *(C2)* $\mathbb{E}[\widehat{\Delta}]$ equals the true gain only when no importance weight is truncated, $B_w \geq \sup \pi/\mu$ exactly. Otherwise the certified functional is the truncated one, whose deficit Δ_∞ is K-independent and, for rewards mapped to $R \geq 0$ before weighting, below G; if $\Delta_\infty \leq 0$, no budget certifies any gain. A deterministic legacy is the extreme case: $\mu(a^\mu \mid x) = 1$ at every logged action, so $w \leq 1$ and the deficit is non-positive at every K, which caps the certificate at $\widehat{V}(\mu)$. Positivity of μ over the candidate's support is therefore necessary, and (30) gives the budget it then takes. On the Open Bandit log and the two HeartSteps levels of Section 7, with the loss moments taken from the certification sample (a consistency check, not a prediction), the derived c of the value form agrees with the measured one to within 1% (3.5% on the HeartSteps outer level) once the positive third-order term of D_K is included.

D.1 Graph-error taxonomy and corruption study: full statement

Four error cases and the logging condition. We distinguish four ways the supplied graph, or the log it is applied to, can be wrong. Only the fourth threatens validity. *(i) Extra edges* cost compression, not validity. Remark 3 requires only an I-map, so every d-separation read off a superset DAG holds in the true distribution, and backdoor admissibility (Assumption 3) survives. The price is a larger KL sum. *(ii) Working-model misspecification* can widen the certificate without invalidating it, because the LS estimator touches the data only through the importance weights and realised rewards (Remark 2). Appendix D discusses the observable symptoms. *(iii) Missing observation*, a context component the practitioner cannot see, costs performance, not validity. In an NCCB, a variable no policy acted on cannot confound the action: both policies read only the observed history, so the laws of a trajectory under π and μ differ only in the policy factors, also if the variable acts across levels. Lemma 2 then holds as long as the variable is independent across episodes given $\mathcal{H}_0$. If the legacy did act on it, the certificate stands provided the log records it; only the planner's candidate quality suffers. *(iv) Missing logging* is the dangerous case: the log omits a variable the behaviour policy acted on, or replaces it by a post-action measurement. The recorded propensities then misrepresent how the actions were chosen, the identification step of Lemma 2 fails, and the bias can point upward. Within this taxonomy only case (iv) can produce false certification, and no off-policy method escapes it without additional signals (instruments, proxies). Cases (i) and (iii) therefore reduce the requirement from "know the causal graph" to a checkable *logging condition*: every dataset a certification round reweights must record, at their decision-time values, the variables its behaviour policy acted on. Incompleteness on the practitioner's side is safe; incompleteness on the log's side is not.

Corruption study: design. We test the taxonomy on Γ-SCM at $\sigma_Q = 1$ (10 seeds, $K = 8000$, $\delta = 0.05$, NIG posterior, rewards mapped to $[0,1]$ by the declared range, then IS weights truncated at $w_{\max} = 4$). The study compares a reference certificate (correct graph, true propensities) with six corruption modes. Mode 1 adds a spurious edge (case i). Mode 2 deletes the true edge $Q \to Y$, so the supplied graph is not an I-map. Modes 3a and 3b record propensities marginalised over the trigger the legacy acted on, using its design prior (3a) or the empirical action frequencies (3b). Mode 4 evaluates them at a post-action measurement. Mode 5 records propensities at a permuted context column. Modes 3a, 3b, 4 and 5 are all case iv. The metric is the *false-certification rate*: the fraction of seeds whose certified estimand falls below $\widehat{V}^{\text{LB}}$ at the final K. The weight covers only the inner action, so the bound certifies a hybrid policy: at the meta level it plays whatever the logging agent chose in that episode, and at the inner level it plays π_Q, drawing one $\theta \sim Q$ per episode. Because the meta choices change from episode to episode, the certified value is this hybrid's value averaged over the certification episodes (Corollary 6). We estimate it by Monte Carlo (4000 episodes per checkpoint, standard error at most 1.3). The handover gate is the gain certificate (29) against the on-policy legacy mean, with its Hoeffding term and δ split over the three checkpoints.

Table 3 confirms the taxonomy. Every valid arm, including the graph that is not an I-map, stays at 0/10 at both caps, and at $w_{\max} = 2.5$ no valid certificate is vacuous. Only mode 3a fails (2/10 at $w_{\max} = 2.5$, 10/10 at $w_{\max} = 4$), and at $w_{\max} = 4$ it hands over to a harmful policy: four of its

Table 3: Corruption study ($\sigma_Q = 1$, 10 seeds, $K = 8000$, $\delta = 0.05$), at $\lambda^\star$ rounded to the grid of Remark 6; $\widehat{V}^{\mathrm{LB}}$ and the IS point estimate $\widehat{V}^{\mathrm{IS}}$ in per-episode reward units, against a certified value of 32.3 (37.9 and 77.0 for modes 1 and 2). False cert.: seeds whose certified estimand lies below $\widehat{V}^{\mathrm{LB}}$ at final K, in parentheses the number of vacuous seeds ($\widehat{V}^{\mathrm{LB}} \le N_L R_{\mathrm{lo}} = -300$); gate: gain certificate (29) > 0 at any checkpoint $k \in \{2000, 4000, 8000\}$, in parentheses firings with negative true gain.

	$w_{\max} = 4$				$w_{\max} = 2.5$
mode	$\widehat{V}^{\mathrm{LB}}$ (mean $\pm$ sd)	$\widehat{V}^{\mathrm{IS}}$	false cert.	gate	false cert.
reference (correct graph, true propensities)	-308.5 ± 28.0	-90.1	0/10 (5)	0/10	0/10
1: spurious extra edge (case i)	-278.8 ± 55.5	-81.8	0/10 (2)	0/10	0/10
2: missing true edge $Q \to Y$ (not an I-map)	-201.6 ± 23.1	-32.3	0/10	0/10	0/10
3a: confounded, design-marginal μ (iv)	$+180.0 \pm 176.0$	358.9	**10/10**	**6/10** (4)	**2/10**
3b: confounded, empirical-marginal μ (iv)	-107.8 ± 37.0	100.7	0/10	0/10	0/10
4: descendant leakage (iv)	-312.2 ± 27.2	-93.6	0/10 (8)	0/10	0/10
5: permuted-context propensities (iv)	-75.3 ± 25.0	124.6	0/10	0/10	0/10

six gate firings are on seeds where the true gain is negative. Modes 3b and 5 do not falsely certify, but at $w_{\max} = 4$ their IS point estimates (100.7, 124.6) exceed the certified value (32.3); only the pessimism of the bound keeps them below it. The missing edge costs planning, not certification: the deployed policy collapses to a constant action and loses 11.7 reward units of true value against the reference's good-policy seeds, yet the bound stays valid. Remark 3 does not cover it, but Lemma 2 and Theorem 1 hold for any posterior, so the table is consistent with the theory. Corrupted certificates do not diagnose themselves: mode 2 and modes 3a, 3b and 5 produce certificates *narrower* than the reference, and mode 2 a lower split-KL. A range check would catch only 2 of mode 3a's 10 false bounds at $w_{\max} = 4$, those above $N_L R_{\mathrm{hi}} = 300$. Every quantity a practitioner watches looks better while the policy degrades. The logging condition is therefore a design-time requirement, not a runtime check.

Corruption study: protocol and per-seed reading. The study of Section 6.2 uses Γ-SCM at $\sigma_Q = 1$ ($\gamma = 1$, $\eta = 1$, $N_L = 20$, reward clipped to $[-15, 15]$), an ε-mixture threshold legacy ($\varepsilon = 0.2$ on the 7-point inner grid) acting on the observed trigger Q, an even prior/certification split, and true values from 2×10^4 Monte Carlo episodes per seed on the uncorrupted environment. Mode 1 trains under a spurious edge $R \leftarrow \{Y, T\}$. Mode 2 trains under a graph with $Q \to Y$ deleted, the effect modifier in $f_Y = A \cdot Q$, so the supplied graph omits a true parent. Modes 3a, 3b, 4 and 5 reuse the reference runs and corrupt only the recorded propensities (the trigger stays in the log for model fitting), so they isolate the logging failure. Table 3 reports the certificate at the closed-form $\lambda^\star$ rounded down to the grid of Remark 6, which conforms to the theorem. The exact overlap constant of this legacy is $n_a/\varepsilon = 35$, above both caps. Rewards are mapped to $[0, 1]$ by the declared range $[-15, 15]$ before the weight is applied (Assumption 4(i)), so truncation can only lower each loss and the certificate remains a lower bound on the untruncated value. At the fixed $\lambda = 1/(2B')$ every sound arm stays vacuous (reference at $w_{\max} = 4$: -847.8 at $k = 8000$), because the pessimism of the LS estimator at a fixed temperature barely shrinks with K; the grid temperature falls with K and lifts the reference to -308.5.

Per seed, one of the ten reference seeds converges to a wrong-signed policy (true $V \approx -71$); the other nine are good-policy seeds (true $V \approx 88$, mean 88.7). The certified object is a different policy: it keeps the logged meta actions and randomises the inner action over $\theta \sim Q$; its true gain over the legacy is positive only on seeds 5 and 6. Under mode 3a the gate fires on seeds 5 and 6 at $k = 2000$ and on seeds 0, 1, 7 and 8 at $k = 8000$, where the true gain is negative. Mode 2 changes what the agent deploys: a constant action worth 76.9 ± 0.1 replaces a Q-conditional sign policy worth 88.7 on good seeds. Its bound stays valid on every seed. Its certificate is narrower than the reference's (mean width $\widehat{V}^{\mathrm{IS}} - \widehat{V}^{\mathrm{LB}}$ of 169.4 against 218.4) and its split-KL lower (3.13 against 9.90 nats, paired $t = -3.59$ over 10 seeds), because a mechanism with fewer parents moves less from its prior. The propensity-corrupted modes are also narrower (178.9, 208.5 and 200.0 for 3a, 3b and 5).

Self-diagnosis without ground truth. A practitioner cannot separate Effect 1's two regimes (well-aligned vs misaligned prior) by inspecting p_0 alone, because the truth θ^*_X in (28) is unknown. The bound itself provides the diagnostic, but its two candidate symptoms differ in reliability. The

KL symptom, a large $\mathrm{KL}(Q_K \| p_0)$ as the online posterior pulls away from p_0 towards parameters consistent with the LS estimator, depends on the posterior family. Under the online Gibbs posterior, misspecification inflates the KL: in the wrong-graph contrast at $\sigma_Q = 1$ (Appendix H, item iv), the mean mechanism KL rises from 3363.2 under the correct graph to 19884.6 under the wrong one. Under the conjugate NIG working model, the KL can instead *shrink*: from 468.3 to 58.5 in the same contrast, while the bound still looks valid. A small KL is therefore no evidence of alignment. The reliable symptom is the certificate itself: misalignment that matters shows as a low $\widehat{V}^{\mathrm{LS}}_{\mathrm{pes}}(\pi_{Q_K})$, a certificate that widens, or a certificate that fails to clear the incumbent, whichever backend produced the posterior. Widening is the analytical mechanism of Effect 1; in the NIG contrast the shrunken-KL certificate stayed valid, so no alarm was owed. The bound is therefore self-conservative with respect to prior and working-model misalignment: *it does not falsely tighten in the wrong-direction case, it loosens or refuses to certify*. This self-diagnosis does not extend to corrupted inputs. When the logged propensities are wrong (case iv of Section 6.2), the certificate can come out narrower, not wider, while being invalid (mode 3a in Table 3). An omitted true edge likewise narrows the certificate and lowers its KL while the deployed policy degrades (mode 2, which stayed valid). Auditing the logged propensities is a precondition for trusting the certificate, not something the certificate can check for itself.

E AEGIS Design Choices and Algorithm Coverage

Section 6.1 presents AEGIS's three deployment-time modifications. This appendix collects: the host instantiations covered by the recipe, the per-switch design justifications, an implementation note tying the switches to a single configurable agent class, and the gate's correctness clauses on legacy-action logging and post-handover data buffers.

Host instantiations

Theorem 1 applies to any algorithm that produces a Bayesian posterior p_k with a defined $\mathrm{KL}(p_k \| p_0)$ to a data-free prior, together with a per-mechanism reward predictor $\hat{r}_\theta(a, x)$. A posterior that does not factorise over mechanisms pays the joint KL in place of the sum. We name four host families:

- *AEGIS-NCTS*: AEGIS over Algorithm 1; the host whose runs Sec. 7 certifies.
- *AEGIS-UCB*: per-mechanism upper-confidence-bound exploration during data collection, the certified π_{p_k} at deployment.
- *AEGIS-IDS*: information-directed sampling [26] as the exploration rule, identical handover gate.
- *AEGIS-PB*: direct PAC-Bayes minimisation $Q_k = \arg\min_Q \big(-\widehat{V}^{\mathrm{LS}}_{\mathrm{pes}}(\pi_Q) + \mathrm{KL}(Q\|p_0)/(\lambda K)\big)$ replacing passive Bayesian updating.

The three switches touch only p_k, $\hat{r}_\theta$, and the legacy log (x_t, a_t^μ, R_t), so they apply verbatim across the four hosts; none of them depend on *how* p_k was produced.

Why the LCB form of (7)

The deviation factor $\kappa_k = \sqrt{2(\mathrm{KL} + \log(1/\delta))/k}$ has the rate of the PAC-Bayesian deviation bounds of Seldin et al. [30], evaluated at episode k on the k accumulated episode means. It is large when the posterior is broad (early episodes) and shrinks at $\widetilde{O}(1/\sqrt{k})$. At $\kappa_k = 0$, (7) reduces to greedy posterior-mean acting; at κ_k from (8), it is a lower confidence bound on the action value, in the spirit of the pessimistic learning principles of Aouali et al. [2].

Two alternatives we considered and rejected: *(a) Pessimistic Thompson sampling* (sample θ from a tilted lower-quantile of p_{k-1}). Equivalent in expectation but reintroduces the per-step sampling variance the LCB removes; loses the deterministic-action advantage at deployment time. *(b) McAllester-style* $\sqrt{\mathrm{KL}/k}$. The Logarithmic-Smoothing supermartingale of Haddouche and Sakhi [8, Thm. 3.2] that underlies Theorem 1 gives this form after tuning λ; using a McAllester rate would either require a separate proof or pay a redundant constant factor.

Why the closed-form of (9)

The derivation expands $-\log(1-\lambda c) = \lambda c + \lambda^2 c^2/2 + O(\lambda^3 c^3)$ in (5), then minimises the leading-order λ-dependent terms in $\widehat{V}^{\mathrm{IPS}} - \frac{\lambda}{2} S^2 - C/(\lambda k)$. The first-order condition gives $\lambda^\star = \sqrt{2C/(kS^2)}$; the higher-order correction is $O(\lambda^2 B^3)$ and immaterial for $\lambda B \ll 1$. The clip $(1-\eta_{\mathrm{clip}})/B$ enforces the domain constraint with safety margin η_{clip} (subscripted to distinguish it from the SCM nonlinearity axis η of Appendix G).

Comparison: The bounds of Haddouche and Guedj [7] hold for a fixed λ, so tuning it requires a grid over $\mathbb{R}_+$. (9) replaces the grid with a single algebraic step. Validity requires a temperature fixed independently of the data it certifies. So S_k^2 and C_k must come from data outside the LS sum (the prior split), or λ must come from an a-priori grid closed by a union bound. At $S_k^2 = 0$ only the cap in (9) applies. Theorem 1 does not cover a $\lambda^\star$ read off the certification sample itself (Appendix D, Effect 5, condition C1).

Why sticky per-level Σ_k, not adaptive

The sticky semantics (ℓ in $\Sigma_k \implies \ell$ in $\Sigma_{k'}$ for all $k' > k$) is a deliberate restriction. An adaptive variant that allows ℓ to leave Σ_k would make the realised behaviour of episode $k+1$ a function of all of $\mathcal{F}_k$, but Theorem 1 certifies a fixed deployment scope, not a re-optimised one. In a regulated handover setting (the practical use case), reverting from agent control back to legacy control is also operationally disruptive and rarely warranted.

Implementation note

The three switches above are flags on a single `ICTSAgent` class in the codebase: `risk_adjust_inner`, `optimal_lambda`, `progressive_handover`. The YAML preset `aegis_cts_handover_tsign_nig.yaml` flips all three on. With all three off, `ICTSAgent` reduces to vanilla NCTS (Algorithm 1); the preset flag is data, not type. This is also why the body keeps two algorithm boxes rather than collapsing into one with conditional branches: the proof in Appendix C reasons about Algorithm 1's policy directly and the three switches are independent ablation knobs over it. Theorem 1 certifies the Thompson mixture $\pi_Q^{\mathcal{I}}$. The LCB action of (7) is a deterministic function of the posterior and lies outside this class, so Algorithm 2 plays $\pi_{p_{k_\ell}}$ after the flip at episode k_ℓ; playing (7) there forgoes the certificate. A flipped level keeps this policy, so the played and the certified policy coincide; the level stops learning, and a later posterior would need its own certificate. In Algorithm 2, $\pi_{p_{k_\ell}}$ denotes one draw $\theta \sim p_{k_\ell}$ per episode, shared by the steps of that level. The loop over ℓ runs in increasing order and the first flip clears the other buffers, so at most one level flips per episode. A buffer holds the logged steps of the episodes in which ℓ acted under the legacy; Eq. (29) reads it as one mean per episode, and K and K_μ count these episodes. The temperature of the first line is computed per level, with $B^{\{\ell\}} = R_{\max} B_w^{(\ell)}$ declared before the data. The joint-scope diagnostic $G_k^{[L]}$ is the right-hand side of (29) at scope $[L]$ on $\mathcal{D}^{\mu,[L]}$; it is logged, not gated on.

Edge-case implementation notes

Why a deterministic legacy needs no density. A deterministic legacy $\mu^{(\ell)}$ has $\mu^{(\ell)}(a_t^\mu \mid x_t) \equiv 1$ at every observed action by construction, so the weighted loss in (5) collapses to $\pi_\theta^{(\ell)}(a_t^\mu \mid x_t) \cdot R_t$ per draw $\theta \sim p_k$, with the expectation outside the logarithm, which the agent can compute from its own posterior p_k. For the joint scope $[L]$ the weight becomes the per-draw product $\prod_{\ell \in [L]} \pi_\theta^{(\ell)}(a_t^\mu \mid x_t)$, still without any μ density. The empirical leg $\widehat{V}_K^{\mathrm{IPS}}(\mu)$ is simply the average reward observed under μ during the corresponding legacy-phase steps; data is on-policy under μ, so no IS reweighting is needed. A deterministic legacy therefore requires logging (x_t, a_t^μ, R_t) tuples only; a randomising legacy also needs its propensities.

Single-level scope while another level has flipped. Algorithm 2 clears the per-level buffers $\mathcal{D}^{\mu,(\ell)}$ of the other levels when a gate flips, because the data distribution shifts: e.g. once $1 \in \Sigma_k$, the meta-level data in $\mathcal{D}^{\mu,(2)}$ is generated under $(\mu_{\mathrm{legacy}}^{(2)}, \pi_{p_{k_1}}^{(1)})$ rather than full-legacy. An uncleared

buffer would certify, by Corollary 6, the average over episodes of the value of $(\pi_Q^{(2)}, \mu_k^{(1)})$, which mixes the full-legacy hybrid before the flip with the post-handover hybrid after it, and neither alone. A flipped level keeps its policy, so the behaviour is fixed between flips, and a cleared buffer certifies the gain of $\pi_Q^{(\ell)}$ over $\mu_{\text{legacy}}^{(\ell)}$ under the current hybrid, with Eq. (29) as stated. Practitioners targeting $(\pi_Q^{(2)}, \mu_{\text{legacy}}^{(1)})$ should instead freeze $\mathcal{D}^{\mu,(2)}$ at the moment Σ_k first becomes non-empty.

F Environment Parameters

As a worked example, we instantiate the NCCB (Definition 1) for the linear-ANM setting (Case A1). The experiments of Section 7 do not use this instance; they use Γ-SCM (App. G). All variables are scalar ($d_X = 1$).

Domains. Task type $T \in \{1, \dots, 3\}$ (observed, exogenous); machine setting $M \in \{1, \dots, 5\}$ (between-episode); action $A \in \mathcal{A} = \{1, \dots, 10\}$ (within-episode); context $C \in \mathbb{R}$, yield $Y \in \mathbb{R}$, reward $R \in [0, 1]$ (clipped). Discretisation: $n = 10$ values per scalar component.

Legacy logging policy $\mu = (\mu_M, \mu_A)$. In the offline startup phase, M and A are produced by a legacy logging policy μ. For the simulation, μ_M draws M per episode as $M = 0.6\,T + 1.0 + \varepsilon_M$, $\varepsilon_M \sim \mathcal{N}(0, 0.5^2)$, and $A_t \sim \mu_A(\cdot \mid M, C_t, T)$ is drawn uniformly over $\mathcal{A}$.

True mechanism parameters.

$$
\begin{aligned}
C &= 0.8\,M + \varepsilon'_C, && \varepsilon'_C \sim \mathcal{N}(0, \sigma^2_{C'}),\\
Y &= 0.5\,A + 0.4\,C + \varepsilon'_Y, && \varepsilon'_Y \sim \mathcal{N}(0, \sigma^2_{Y'}),\\
R &= \mathrm{clip}_{[0,1]}\big(0.3\,Y + 0.08\,C + 0.1\,M + 0.05\,T + 0.02\,A + \varepsilon'_R\big), && \varepsilon'_R \sim \mathcal{N}(0, \sigma^2_{R'}).
\end{aligned}
$$

Confounder. $U_t \sim \mathrm{Ber}(0.3)$ i.i.d. per round. Compound noise: $\delta_{UC} = 0.2$, $\delta_{UY} = 0.15$, $\delta_{UR} = 0.1$. Marginalised variances: $\sigma^2_{C'} = \delta^2_{UC}\lambda_U(1-\lambda_U) + \sigma^2_C \approx 0.10$, similarly for $\sigma^2_{Y'}, \sigma^2_{R'}$ (with $\sigma_C = 0.3, \sigma_Y = 0.2, \sigma_R = 0.15$).

Prior hyperparameters (NIG). For each mechanism $X \in \{C, Y, R\}$, $p_0(\theta_X, \sigma^2_X) = \mathrm{NIG}(\mu_0^{(X)}, \Lambda_0^{(X)}, a_0^{(X)}, b_0^{(X)})$ with $\mu_0^{(X)} = \mathbf{0}$, $\Lambda_0^{(X)} = 0.1\,\mathbf{I}$, $a_0^{(X)} = 2$, $b_0^{(X)} = 1$.

G Experimental Setup Details

This appendix expands Sec. 7.1: full Γ-SCM structural equations, the Γ-Stress preset, posterior-backend hyperparameters, and the per-bench deviations table.

Structural equations. Γ-SCM is a single parametric SCM with five axes $(\gamma, \eta, \nu, \sigma_Q, q)$ selected per bench via Hydra overrides:

$$
\begin{aligned}
&T \sim \mathcal{N}(0, 0.3^2), \quad U \sim \mathrm{Ber}(0.5),\\
&M = 0.8T + N_M, \quad N_M \sim \mathcal{N}(0, 0.2^2),\\
&Q = \gamma\big((1-\eta)\,1.5M + \eta\,2\tanh(2M)\big) + 0.3U\\
&\qquad + \big((1-\nu) + \nu(1 + 0.4\gamma|M|)\big)N_Q,\ N_Q \sim \mathcal{N}(0, \sigma_Q^2),\\
&A \sim \mathcal{N}(0, 0.1^2),\\
&Y = (1-q)\,AQ + q\big(-(A-0.5Q)^2 + 4\big) + 0.3U + N_Y,\ N_Y \sim \mathcal{N}(0, 0.08^2),\\
&R = Y + N_R,\ N_R \sim \mathcal{N}(0, 0.5^2).
\end{aligned}
$$

Here Q is the inner context $C = C^{(1)}$ of Sec. 5, not the posterior Q of Theorem 1. Axis semantics: γ (lever strength on $M \to Q$); η (mechanism nonlinearity, linear $1.5M$ vs. saturating $2\tanh(2M)$); ν (homoscedastic additive vs. heteroscedastic noise, $\propto |M|$); σ_Q (inner-context noise std); q (bilinear AQ reward vs. quadratic $-(A-0.5Q)^2 + 4$ with interior optimum $A^\star = 0.5Q$). Hierarchy: meta-action M outer, inner-action A inner, both bounded $[-2, 2]$. The simulator redraws the latent U at every context reveal and at the outcome, so in the reported runs it acts as noise per mechanism and links no two variables; identifiability holds via the Causal Markov property over (T, M, Q, A, Y, R).

The Γ-Stress preset. Unless stated otherwise, every benchmark runs at Γ-Stress: $\eta=1$ (nonlinear f_Q), $\nu=0$ (additive noise), $\sigma_Q=3$ (high inner-noise: $P(\mathrm{sign}(Q)\neq\mathrm{sign}(M))\approx 25\%$ at $|M|=2$), $q=0$ (bilinear), $J=1$ meta-step per episode, $I=20$ inner steps. In this regime M is costly to change between episodes and A is cheap; below this noise floor the main signals (factorisation+commit shape in Sec. 7.2.1/Sec. 7.2.2, the parameter-count contrast in Sec. H.3) are not empirically separable.

Benchmark registry. Table 4 names every benchmark once; the text refers to them by these names.

Table 4: Benchmark registry. L: number of levels. The Γ-benches are presets of Γ-SCM; the methods are defined in Table 5.

Name	Question it answers	L	Key settings	Used in
Γ-SCM	parametric SCM family	2	axes $\gamma, \eta, \nu, \sigma_Q, q$ (equations above)	throughout
Γ-Stress	default; do factorisation and nesting pay?	2	η=1, ν=0, σ_Q=3, q=0, J=1, I=20	Sec. 7.2
Shift-T, -$\mathcal{N}_Q$, -U	does the factorised posterior transfer?	2	zero-shot Γ-Stress targets with shifted T, N_Q or U	Sec. 7.2.1
Contract	does the absolute bound contract?	2	NCTS on Γ-Stress, reward clipped to $[-15, 15]$, prefixes $k \in \{200, \ldots, 2000\}$, D=128 RFF	Sec. 7.2.3
Crossing	does the certified gain cross zero?	2	uniform incumbent, 3-point leaf grid, B_w=3, reward clipped to $[-15, 15]$, N_L=20	Fig. 1, App. H.5
Bounded	crossing under an almost-sure reward bound	2	interaction basis $[Q, A, QA, 1]$	App. H.5
Γ-Add	does a deterministic legacy let the gate fire?	2	$f_Y=2A+0.5Q$, reward clipped to $[-20, 20]$, γ=1, J=1, I=20	App. I
Γ-SCM, σ_Q=1	does a wrong graph falsely certify?	2	six corruption modes	Sec. 6.2
Bin-3L	does the gate fire at depth?	3	binary actions, uniform legacy, B_w=2 per level	App. K.5.1
Depth-Bin	how does firing scale with depth? does the gate protect?	3–5	binary actions, b=2, $R \in [-3, 3]$	App. K.5.2, Table 1
Depth-Grid	how does overlap compound with depth?	3	Γ-SCM plus an outer level, 7-point grids, $B_w^{(\ell)}$=7	App. K

Posterior backends. Two backends, both implementing the Theorem 1 bound. `NIG-Gibbs`: the linear-conjugate Normal-Inverse-Gamma posterior with closed-form update (used as a misspecification contrast against the high-noise $f_Y=AQ$ at $\sigma_Q=3$). `RFF-GP-Gibbs`: D=128 random Fourier features with RBF kernel $\exp(-\|x-x'\|^2)$ and Gibbs posterior at λ chosen per bench; RFF-GP is the default backend for every bandit in Sec. 7 except where NIG is named explicitly. Both use the prior $\mu_0=\mathbf{0}$, $\Lambda_0=\mathbf{I}$, $a_0=2$, $b_0=1$.

Agent taxonomy. Table 5 fixes the bandit vocabulary. All bandits share the same RFF-GP function class unless noted; bandits differ only in (i) the reward model, (ii) the meta-to-inner commit shape, and (iii) the meta-action rule, isolating each NCCB component as a single axis of comparison.

Table 5: Agent bandits used across Sec. 7 and App. H. Same RFF-GP basis (D=128) unless noted.

Bandit	Reward model	Commit shape	Meta-rule	First in
Flat-TS	joint $(T, M, C, A)\to R$	none (flat)	TS over A, μ_M exo.	§7.2.1
Factored-TS	SCM f_C, f_Y, f_R	none (flat)	TS over A, μ_M exo.	§7.2.1
Joint-Commit	SCM factorised	(M, A) joint at meta-start	TS at $\widehat{C}$	§7.2.2
NCTS	SCM factorised	meta→inner recursive	Algorithm 1	§7.2.2
AEGIS-NCTS	= NCTS	+ closed-form $\lambda^\star$, optional LCB inner action	+ per-level handover gate, legacy log μ_{leg}	App. I
Tabular-TS	tabular (M, A) cells	joint	Thompson per cell (= PSRL)	App. H.3

Per-bench deviations. Table 6 lists the main deviations from the Sec. 7.1 defaults (K=2000, 10 seeds, RFF-GP, Γ-Stress preset).

Table 6: Per-bench deviations from Sec. 7.1 defaults.

Sec.	Deviation	Reason
7.2.1	K_{src}=2000, K_{tgt}=25, freeze + replan	zero-shot transfer protocol
7.2.3	K=2000, IS-weighted random-split α=0.5	expose bound's $\sqrt{1/K}$ contraction
H.1	K=200, two posterior backends	posterior-class contrast
H.2	K=100, NIG posterior, four meta-action rules	meta-rule contrast
H.3	K=100, q=1 (quadratic f_Y), $n_a \in \{21, 41, 81\}$	action-grid scaling

H Contrastive Comparisons

Three contrastive comparisons probe the bound's behaviour along orthogonal axes: the posterior class (Sec. H.1), the meta-action rule (Sec. H.2), and the action-grid resolution (Sec. H.3). They live in the appendix because they support the bound's predicted invariances rather than the body's constructive claim; the Sec. 7.1 agent taxonomy and Γ-Stress preset apply unchanged.



(a) Zero-shot OOD reward. (b) Four-arm decomposition. (c) NCTS lower bound contracts.

Figure 2: Constructive ablations of Section 7.2: detailed setup, numerics, and interpretation in Sec. 7.2.1–Sec. 7.2.3.

Sensitivity-analysis map. The four NCTS-robustness dimensions (action-grid resolution, posterior-class misspecification, overlap, graph error) map onto the existing benches as follows. *(i) Action-grid resolution.* Sec. H.3 sweeps $n_a \in \{21, 41, 81\}$ and verifies the invariance in n_a the bound predicts (factorised arms vary by $\Delta < 1$ reward across the grid; Tabular-TS collapses to negative reward). *(ii) Posterior-class misspecification.* Sec. H.1 contrasts well-specified RFF-GP-Gibbs against shape-misspecified NIG-Gibbs: the wrong class costs about 34 reward, not validity, since the bound holds for every posterior. *(iii) Overlap.* The handover gate's per-level overlap $B_w^{(\ell)}$ enters the bound through $\lambda < 1/B^{\mathcal{I}}$ and is automatically respected by the closed-form $\lambda^\star$ (9). The AEGIS demonstration (App. I) stresses overlap through a *deterministic* inner-level legacy $a^{(1)}_{\text{legacy}} = 2\,\text{sign}(T)$ with ε-greedy positivity at the meta level ($\varepsilon = 0.05$) and truncated IPS (a capped variant of the clipped weights of Bottou et al. [3]) at $w_{\max} = 1.0$; capping lowers the estimate for non-negative rewards ($\mathbb{E}[\min(w, w_{\max})\, R] \le V$), but the rewards here are signed, so this bench has no such guarantee. *(iv) Graph error.* A wrong-graph contrast (parents permuted under the same SCM) is inconclusive at the Γ-Stress preset under the RFF-GP backbone (graph-misspecification gap below the noise floor); the principled prediction (Pearl-truncated factorisation collapses under wrong parents) holds at the level of the bound's KL decomposition but is empirically dominated by inner-context noise at $\sigma_Q = 3$. At $\sigma_Q = 1$ the contrast resolves: the wrong graph inflates the mean mechanism KL from 3363.2 to 19884.6 under the Gibbs backend but shrinks it from 468.3 to 58.5 under the conjugate NIG backend (with the bound remaining valid), the backend-dependence taken up in Appendix D; graph and logging error are probed against the certificate directly by the corruption study of Section 6.2.

NCTS bound trajectory under the data-split protocol. Per-k values of the random-split lower bound from Sec. 7.2.3. These are *absolute* value bounds on $V(\pi_{Q_k}; \theta^\star)$, not gain bounds. They are stated per episode under the episode-level aggregation of Lemma 2: one episode-mean record per certified episode, so the regulariser is divided by the episode count k, not by the sample count $N_L k$. Protocol: NCTS runs for K=2000 episodes, and each episode is assigned independently to

a prior or a certificate set with probability α=0.5. For each k the theorem is applied separately to the first k episodes: p_0 is fitted once on the prior portion (size αk), and the episode-mean LS estimator (Eq. (5)) and $\mathrm{KL}(Q\|p_0)$ are evaluated on the certificate portion (size $(1-\alpha)k$). The loss range of the LS supermartingale is data-free: $B' = B_w(R_{\mathrm{hi}} - R_{\mathrm{lo}})$ for the declared reward range $[R_{\mathrm{lo}}, R_{\mathrm{hi}}] = [-15, 15]$. The legacy logs A uniformly over the 7 grid points that the certified policy plays, so the propensity $\mu(a) = 1/7$ is exact and the weight $\pi(a)/\mu(a) = 7\,\pi(a)$ attains its ceiling $B_w = 7$ exactly; hence $B' = 210$ and the temperature $\lambda = 1/(2B')$ is data-free as well. Assumption 4(i) holds by construction, because the environment clips every reward to this range. The empirical range of the observed losses would make B' depend on the certification data. The implementation shifts the weighted loss, $wR - B_w R_{\mathrm{lo}}$, not the reward; since μ puts mass $1/7$ on every grid point, this has mean $V - B_w R_{\mathrm{lo}}$ and lies in $[0, B']$, so the theorem applies to it directly. The paired comparison $\widehat{V}_K \geq V_{\mathrm{lower}}$ has mean difference +20,056 (t=42.2, $p<10^{-11}$). The contraction itself is significant: paired over seeds, $V_{\mathrm{lower}}(2000) - V_{\mathrm{lower}}(200)$ has mean +30,221 (t=31.3, df=9, $p<10^{-8}$). The naive unsplit LS estimator, with the same B' and λ, gives $V_{\mathrm{lower}}^{\mathrm{naive}} = -153{,}066 \pm 1{,}941$ against $V_{\mathrm{lower}}^{\mathrm{split}} = -20{,}053 \pm 1{,}503$ (paired t=156.6, $p<10^{-16}$); it is not a valid bound on $V(\pi_{Q_K})$, since its weights $\pi_{\theta_k}(a)/\mu(a)$ use the single Thompson draw $\theta_k \sim Q_{k-1}$ of episode k, not $\theta \sim Q_K$, while its KL term is $\mathrm{KL}(Q_K\|Q_0)$.

Table 7: Absolute bound $V_{\mathrm{lower}}(k)$ under nested random split, per episode, episode-level aggregation (regulariser over k episodes); mean $\pm$ 1σ across 10 seeds, K=2000, NCTS, no AEGIS switches.

k	200	500	1000	1500	2000
V_{lower}	−50,274±2,149	−36,609±3,185	−27,309±1,339	−23,137±1,297	−20,053±1,503

Anytime trajectory under one pre-fixed temperature. Table 7 and Sec. 7.2.3 report a sample-size sweep: each plot point re-applies Theorem 1 to a prefix. Fig. 3 instead evaluates *two* Ville supermartingales per seed, one per level of the hierarchy, along a single continuous run. The prior p_0 is fixed once from an offline prior portion. Rewards are clipped to $[-15, 15]$ by the environment and mapped affinely to $[0, 1]$ before weighting, so with $B_w = 2$ the loss lies in $[0, B_w]$, $B' = B_w + 10^{-3} = 2.001$, and Assumption 4(i) holds by construction. One global temperature $\lambda = 1/(2B') = 0.2499$ is fixed from these declared constants; the per-episode $\lambda_k^\star$ of Eq. (9) is disabled. Both levels aggregate one term per certified episode (Lemma 2). The two levels share one budget $\delta = 0.05$ ($\delta/2$ per level). The LS leg is time-uniform at the fixed λ, and the Hoeffding leg on $V(\mu)$ is made anytime by spending $\delta_2\, 6/(\pi^2 j^2)$ at the j-th of the 20 chunk boundaries evaluated. The posterior keeps learning during the run, so every evaluation rescores the whole legacy log under the current posterior and charges that posterior's KL: each point is a certificate for one Q, and all 40 points of a seed hold simultaneously with probability at least $1 - \delta$. We read $V_{\mathrm{lower}}(k)$ at the pre-fixed checkpoints $k \in \{200, 500, 1000, 1500, 2000\}$.

No point lies above its seed's Monte Carlo true value (+83.83, +75.84, +86.21); the smallest margin over all 120 evaluated points is 609.0 per episode. From $k = 200$ to $k = 2000$ the meta bound rises from between −910.6 and −897.5 to between −805.0 and −671.7, and the inner bound from between −625.9 and −615.9 to between −545.0 and −522.8, without retuning λ. The regulariser and the Hoeffding term shrink with k (together 41.2 to 63.9 per episode at $k = 2000$). Two terms do not: the Jensen gap $D_K(\lambda)$ of the LS estimator (Appendix D, Effect 5), 340 to 464 per episode at this λ, and the coverage deficit of the deterministic legacies, between −303.4 and −144.5 at $k = 2000$ and non-positive at every evaluated point (largest value −141.9). What the figure does *not* demonstrate is a gate with a $\lambda_k^\star$ fitted on the certification data, which lies outside the scope of a fixed-temperature supermartingale.

H.1 Posterior-class misspecification costs reward, not validity.

On Γ-Stress f_Q (10 seeds, K=200, λ=1), the matched RFF-GP backend earns 115.8 ± 0.5 per episode and the shape-misspecified NIG-Gibbs backend 81.9 ± 1.9 (mean $\pm$ SE): the conjugate form cannot calibrate σ^2 at $\sigma_Q = 3$, and the wrong class costs about 34 reward. It cannot cost validity: Theorem 1 holds for every posterior, so the certificate stays valid for the misspecified class.

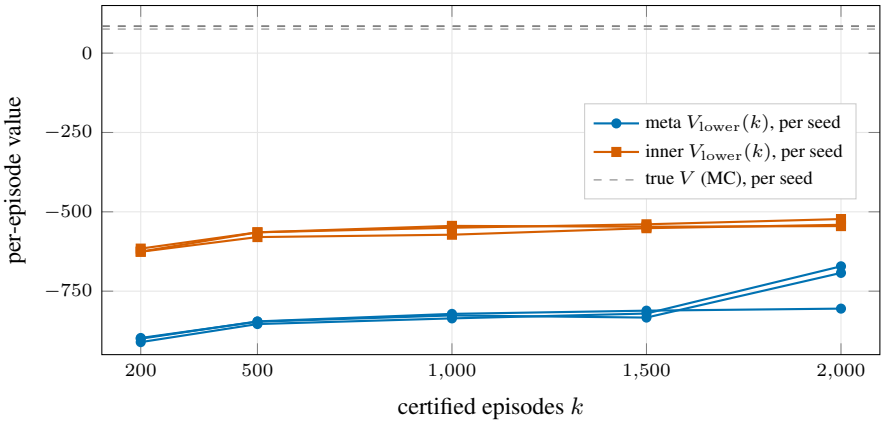


Figure 3: **Anytime trajectory under a single pre-fixed temperature.** Two Ville supermartingales per seed, one per level of the 2-level hierarchy (3 seeds, $K{=}2000$ certified episodes), sharing $\delta = 0.05$ ($\delta/2$ per level, Hoeffding leg anytime over the 20 evaluations): one a-priori $\lambda = 1/(2B') = 0.2499$ on the $[0,1]$-mapped rewards, fixed from the declared constants and never re-optimised, with $V_{\text{lower}}(k)$ read at five pre-fixed checkpoints out of the 20 evaluated, for the meta level (blue circles) and the inner level (vermillion squares). Both levels aggregate one term per certified episode, and each evaluation rescores the legacy log under the current posterior. Every certificate stays below its seed's Monte Carlo true value (dashed grey) at all 120 evaluated points and rises from $k = 200$ to $k = 2000$ at every seed and level; the remaining gap is the LS Jensen gap at this λ plus the non-positive coverage deficit of the deterministic legacy policies, neither of which shrinks with k. A gate with a $\lambda_k^\star$ fitted on the certification data is not covered by this demonstration.

H.2 Under the NIG posterior, D-optimal collapses under stress; Thompson and greedy dominate.

Question. Across three meta-action rules (Thompson as in Eq. (15), D-optimal $\arg\max_m \mathbb{E}_{p_{k-1}}[\log\det F(\theta;\, \mathbf{do}(M{=}m))]$, greedy posterior-mean), plus a top-k Thompson-tiebreak hybrid, which survives at high inner-context noise? All four rules use the NIG posterior ($K{=}100$, 10 seeds). *Result* (Fig. 4a, Table 8). *Interpretation.* Pure D-optimal collapses to single-digit reward: under high inner-context noise the info-gain objective concentrates on m values informative about θ_C but noisy on the reward channel; the top-k Thompson-tiebreak hybrid only partially recovers. Nor does D-optimal buy a better fit of f_Q here: its mechanism MSE on f_Q is 9.1, against 9.8 for Thompson. The collapse is specific to the NIG posterior: with the RFF-GP backend, pure D-optimal earns 112.4 per episode against 115.8 for Thompson ($K{=}200$).

Table 8: (Sec. H.2) Mean reward by meta-action rule on Γ-Stress f_Q (NIG posterior, $K{=}100$, 10 seeds); bold marks the winner.

Rule	Reward
Thompson	**80.2**
Greedy	77.5
D-opt-Hybrid	10.0
D-optimal	7.0

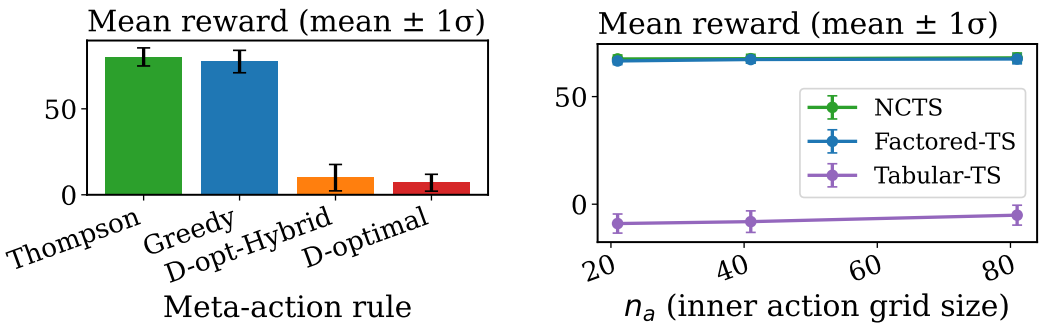
H.3 Invariance in n_a for the factorised arms; Tabular-TS goes negative.

Question. The bound predicts that factorised-arm reward depends on the parameter count of the factorised model, $D{=}128$ RFF features per mechanism and independent of action-grid resolution n_a, while a tabular flat-joint basis has one parameter per cell of $|\mathcal{A}_S| = |\mathcal{M}|\cdot n_a$. Does this invariance survive at the stress regime? *Setup.* Γ-Stress with quadratic reward ($q{=}1$) ($Y = -(A{-}0.5\,Q)^2 + 4$); $A^\star = 0.5\,Q$ is continuous in Q, and the grid on $[-2,2]$ has spacing $h = 4/(n_a{-}1) \le 0.2$, so discretisation costs at most $(h/2)^2 \le 0.01$ reward per step against the best action in $[-2,2]$; the n_a axis tests exploration cost up to this term. Arms: NCTS (RFF-GP, factorised + targeted), Factored-TS (RFF-GP, factorised, no targeting), Tabular-TS (tabular). *Result* (Fig. 4b, Table 9). Factorised arms vary by $\Delta < 1$ reward across the grid; Tabular-TS stays in the negative-reward floor at every n_a. *Interpretation.* Factorised arms are flat in n_a, the invariance the

Table 9: (Sec. H.3) Mean reward across action-grid resolutions n_a on Γ-Stress with quadratic reward ($q{=}1$).

Arm	$n_a{=}21$	41	81
NCTS	67.49	67.60	68.00
Factored-TS	66.56	67.23	67.47
Tabular-TS	−8.99	−8.10	−5.09

bound predicts. Tabular goes negative because with $K \cdot J \cdot I/(7 \cdot n_a)$ samples per cell, even $n_a = 21$ leaves only ~ 14 samples per arm against $\mathrm{Var}(R) \sim 9$, so the Thompson argmax is dominated by noise. The $|\mathcal{A}_S|$-penalty the bound predicts within tabular is swamped at the negative-reward floor; what is visible is the $|\tilde{\mathcal{Z}}|$-vs-$|\mathcal{A}_S|$ *level* gap of ~ 75 reward between the structured factorised arms and unstructured tabular. The bound's prediction holds at the level; the within-tabular n_a-scaling is swamped by the noise.



(a) Reward by meta-action rule, NIG posterior (Sec. H.2).

(b) Reward vs. n_a, three arms (Sec. H.3).

Figure 4: Contrastive comparisons: detailed setup, numerics, and interpretation in Sec. H.2 and Sec. H.3.

H.4 Effect attribution: shift rows

Nesting dominates in distribution, factorisation under shift. Table 1 in the body carries the in-distribution decomposition and the gate ablation. Table 10 adds the Welch t statistics and the exogenous-shift rows of Sec. 7.2.1. On the mild Shift-U ($\lambda_U : 0.5 \to 0.9$) the factorised bandit attains 116.1 ± 5.0 against 115.0 ± 4.3 ($\Delta = +1.2$): at $K = 2000$ the joint regression has accumulated enough data to compensate on the mild shift. The gate itself is isolated on a bench where it fires (Table 1, bottom).

Table 10: Attribution with Welch t, including the exogenous-shift bench. Per-episode reward, mean ± sd over 10 seeds.

Effect	**Contrast**	**A**	**B**	Δ	d	t	p
Factorisation	Factored-TS vs Flat-TS	114.7±0.3	111.5±0.5	+3.3	8.21	18.36	5.1×10^{-11}
Nesting (on top)	NCTS vs Factored-TS	118.6±0.6	114.7±0.3	+3.9	7.81	17.46	7.7×10^{-10}
Both (total)	NCTS vs Flat-TS	118.6±0.6	111.5±0.5	+7.1	12.34	27.60	1.7×10^{-15}
Nesting vs joint	NCTS vs Joint-Commit	118.6±0.6	77.4±3.4	+41.2	16.95	37.90	7.8×10^{-12}
Factorisation (Shift-T)	Factored-TS vs Flat-TS	113.6±4.9	56.1±10.1	+57.5	7.2	16.17	5.8×10^{-10}
Factorisation (Shift-$\mathcal{N}_Q$)	Factored-TS vs Flat-TS	194.7±4.9	146.1±6.6	+48.6	8.4	18.69	1.3×10^{-12}

Gate ablation protocol. The deployment rules *never*, *always* and AEGIS run on the meta level of Depth-Bin at $L = 3$ (App. K.5.2; binary action sets, uniform legacy, interaction strength $b = 2$, per-level weight ceiling $B_w = 2$), over a 6,000-episode horizon and 10 seeds, with the same certificate and δ as the depth sweep. The good candidate is the learned meta policy (its posterior mean; the certified Thompson mixture differs by at most 0.03 per episode); the bad candidate is the same policy negated. AEGIS is not rolled out: its reward is the legacy reward up to the crossing and the candidate reward after it. Against the good candidate the certified crossing occurs at $K = 1{,}268$ to 4,872; per-episode rewards are 0.00 ± 0.37 (*never*), 23.56 ± 0.53 (*always*) and 10.86 ± 3.90 (AEGIS). Against the bad candidate the gate never fires, so AEGIS coincides with *never* on every seed (0.00 ± 0.37), while *always* collects -23.56 ± 0.66.

H.5 Certified-gain crossings

Crossing scale. The implemented per-level certificate is the gain bound (29) with its two confidence legs made explicit: the certified gain per episode is $\widehat{V}^{\mathrm{LS}}_{\mathrm{pes}} - \widehat{V}^{\mathrm{IPS}}(\mu)$ minus a regulariser $N_L(R_{\mathrm{hi}} -$

$R_{\mathrm{lo}})\,(\mathrm{KL}_\Phi + \log(1/\delta_1))/(\lambda K)$ minus a Hoeffding width at confidence δ_2, with $\delta_1 + \delta_2 \le \delta$. At the j-th checkpoint the Hoeffding width is evaluated at confidence $\delta_2 \cdot 6/(\pi^2 j^2)$; since $\sum_{j \ge 1} 6/(\pi^2 j^2) = 1$, a union bound makes it hold at all checkpoints simultaneously, as the LS term does by Ville's inequality. On the first four rows of Table 11 each of the L levels is certified at confidence δ/L, as on the depth benches of App. K. Algorithm 2 restarts a buffer at a flip, which starts a fresh supermartingale at a stopping time, and a level is gated in at most L such epochs. A union bound over the at most L^2 pairs of level and epoch then certifies a run that checks a gate after a flip at δ/L^2 per pair, or at any fixed weights that sum to at most δ over these pairs; no run here does so. With the posterior learning, the LS sum at each checkpoint is evaluated under the current posterior, the posterior whose KL enters the bound (Theorem 1). Let g be the candidate's per-step advantage over the incumbent. Optimising λ, the bound crosses zero, to leading order, at

$$K_{\mathrm{cross}} \;\approx\; \Big(\frac{(R_{\mathrm{hi}} - R_{\mathrm{lo}})\,A}{g}\Big)^2, \qquad A = B_w\sqrt{2\,(\mathrm{KL}_\Phi + \log(1/\delta_1))} + \sqrt{\log(1/\delta_2)/2}, \tag{31}$$

and N_L cancels. This takes $K_\mu = K$ and is an upper estimate: Effect 5 has $S \le B_w$ in place of B_w. At a fixed λ the Jensen floor of condition C1 applies instead. So K_{cross} is linear in the KL budget and quadratic in B_w, the reward span and $1/g$. Contract (Sec. 7.2.3), whose KL_Φ grows with the feature dimension, therefore contracts without crossing at $K{=}2000$. Benches with a small exact B_w and a low-dimensional mechanism class cross within a few thousand episodes.

The benches. Table 11 lists the benches on which the certified gain crosses zero at $\delta{=}0.05$ (Bin-3L: App. K.5.1). On the first four rows the declared weight ceiling is the exact realised supremum (attained at every checkpoint and never exceeded), so no weight is truncated and Assumption 4(ii) holds with this ceiling. There λ is fitted once on the offline prior split and held constant, an $\mathcal{H}_0$-measurable choice licensed by Theorem 1. The two-level bench Crossing uses a uniform incumbent on a 3-point leaf grid, so $B_w{=}3$, with a degree-2 feature map, reward clipped to $[-15, 15]$ and $N_L{=}20$. Only the leaf level certifies; the meta level does not cross on any seed. The first four rows run with (7) off, as Algorithm 2 specifies.

Table 11: Certified-gain crossings at overall $\delta{=}0.05$ (δ/L per level). Gains are certified per-episode lower bounds on $V(\pi) - V(\mu)$ at the crossing checkpoint; the last column is the certificate later in the run, with the gate suppressed so that it keeps being evaluated after the crossing. Frozen: posterior fixed on the offline split, so $\mathrm{KL} \equiv 0$.

Bench	**Posterior**	**Cross**	K_{cross}	KL at cross	Gain at cross	Later
Crossing	live	10/10	3,000 to 16,000	2.55 to 25.73	+0.50 to +18.89	+25.69 to +42.86 at 52,000
Crossing	frozen	10/10	6,000 to 7,000	0	+1.50 to +12.87	
non-additive noise	frozen	3/3	5,000	0	+16.74, +18.40, +17.75	
bounded reward	frozen	1/1	24,000	0	+4.915	+115.927 at 1,380,000
nested, $L{=}3, 4, 5$	frozen	every level at every depth, 10 seeds each (App. K)				

Live posterior. With the posterior left learning, the KL at the crossing is 2.55 to 25.73 nats and grows to 14.11 to 62.44 nats by $K{=}52{,}000$. At the crossing it accounts for 40.4 to 87.2% of the regulariser. Two qualifications apply. First, the live and frozen arms differ in more than the freeze: the prior-split λ budgets the KL it anticipates, which is zero under the freeze, so switching the freeze off also raises λ from about 0.0044 to about 0.030 (a factor of 6.9); both values are $\mathcal{H}_0$-measurable. Second, the regulariser charges the aggregate KL over all of Φ, not the restriction to $\Phi_{\mathcal{I}}$ that Lemma 3 licenses, so the quoted KL over-charges the certificate. The frozen arm crosses at $K{=}6{,}000$ or $7{,}000$ on all ten seeds. The freeze buys a certificate that is smoother in K, not one that fires where the live certificate cannot.

Deployment trigger. The gate as run deploys at the first crossing. On 6 of the 10 live seeds the certified gain stays positive at every checkpoint after the first crossing; on the other 4 it dips below zero at least once before staying positive. A trigger that requires the gain to stay positive at every later checkpoint, evaluated on the gate-suppressed counterfactual trajectories, still fires on all ten seeds, at $K{=}3{,}000$ to $16{,}000$, costing at most $11{,}000$ extra episodes on the worst seed. We recommend this sustained trigger for deployment.

Non-additive noise. The non-additive variant keeps the hierarchy of Crossing and makes the Q-noise scale with the parent action, which violates the additive-noise form of the working model by

construction. It is a noise variant of the same hierarchy, not a different environment family. With the gate suppressed, the leaf certificate crosses at $K{=}5{,}000$ on all three seeds.

Bounded reward. Bounded has an almost-sure reward envelope, so clause (i) of Assumption 4 holds by derivation, not as a tail bound. The outcome mechanism uses a four-coefficient interaction basis $[Q, A, QA, 1]$. The crossing lies between two checkpoints of a 2,000-episode grid (-5.101 at $K{=}22{,}000$, $+4.915$ at $K{=}24{,}000$), and all 679 checkpoints from there on are positive. At $K{=}1{,}380{,}000$ the certificate of $+115.927$ per episode is 96.4% of the paired Monte Carlo advantage of 120.198 (95% interval $[119.892, 120.504]$). The certified policy is context-dependent, and its advantage survives setting the mean of U to zero (120.202 against 120.198). This is a single run.

I Systematic Handover Evaluation

The deployment story of NCCB / AEGIS is *progressive certified handover*: control passes from a legacy controller to the agent at each timescale only once the per-level certificate meets a practitioner-specified risk threshold. This appendix covers the benches with a deterministic legacy. It reports (i) the reference trajectory on Γ-Add, which shows no certified crossing, and (ii) the consolidated handover statistics (Table 12). On these benches the gate does not fire. Section 7 reports the benches on which the gate fires, among them the Crossing bench of Fig. 1c.

Reference trajectory: Γ-Add

Setup. Environment: Γ-Add ($\gamma{=}1$, $J{=}1$, $I{=}20$, reward clipped to $[-20, 20]$), a linear-additive variant of Γ-SCM in which $f_Y{=}2A + 0.5Q$ is correctly specified by the agent's linear-NIG posterior. Legacy controller μ_{legacy}: deterministic $a^{(1)}_{\text{legacy}} = 2\,\text{sign}(T)$, which matches the oracle argmax $a^\star{=}{+}2$ on 50% of meta-contexts. Agent: AEGIS-NCTS over the linear-NIG posterior, with LCB inner action (7) and per-level handover gate, whose own gate test is replaced by the certificate below. Hyperparameters: ε-greedy positivity on M at $\varepsilon{=}0.05$; truncated IPS with $w_{\max}{=}1.0$ (a capped variant of the weights of Bottou et al. [3]); gain-bound check stride 50; certificate confidence $\delta = 0.5$ (a median-confidence setting; at $\delta = 0.05$ every bound below is strictly lower). The posterior is not frozen (KL reaches about 735 nats by $K = 5{,}000$). Budget: $K{=}5{,}000$ episodes $\times$ 3 seeds. The certificate is evaluated at every gain check in the form Theorem 1 licenses: one term per certified episode at both levels (Lemma 2); the expectation over Q outside the logarithm, over 4 posterior draws; rewards mapped to $[0, 1]$ before weighting, so that $B' = B_w + 10^{-3} = 1.001$ with $B_w = w_{\max} = 1$; the data-free $\lambda = 1/(2B') = 0.4995$, since no prior split exists on this bench; δ shared by the two levels; the Hoeffding leg anytime over the 101 checks; and the whole legacy log rescored under the current posterior at every check, with that posterior's KL. The legacy holds control throughout; up to the first crossing of this certificate this is the trajectory of a gate driven by it.

No certified crossing. Under the licensed accounting the certificate contracts from about $-4.8{\times}10^4$ at $K = 1$ to -596.7, -599.5 and -598.6 at $K = 5{,}000$ at the inner level (-978.0, -969.2 and -956.2 at the meta level), and never crosses 0 up to $K = 5{,}000$. Every weight is at most 1 (deterministic inner legacy; truncation at $w_{\max} = 1$ at the meta level), and with rewards mapped to $[0, 1]$ the coverage deficit is then non-positive (Effect 5). The measured inner deficit is 0 at $K = 1$ on one seed, where every logged weight equals 1, and lies in $[-168.1, -142.8]$ from $K = 100$ on; the meta deficit is at most -307.5 at every check. At $K = 5{,}000$ the remaining terms are the regulariser (about 236 per episode, from the KL of the learning posterior), the Jensen gap at the data-free λ (172 to 173 at the inner level, 302 to 309 at the meta level) and the Hoeffding term (27.5). We claim no handover time for this bench and report no pre/post-handover reward comparison, because it would be conditioned on a gate event the bound does not produce. The trajectory illustrates the deployment contract (a per-level guarantee, a practitioner threshold, a sticky handover) and the contraction of the bound, not a certified firing.

Consolidated handover statistics

Table 12 collects the handover evidence. The long-K runs (one seed per σ_Q, $2{\times}10^5$ certified episodes after 2×10^5 offline legacy episodes, $\delta = 0.5$) certify in the form of Theorem 1: the posterior is fitted on the offline split and frozen (KL $= 0$); λ is Eq. (9) on the offline split, held constant (2.91×10^{-3}

at every σ_Q); rewards are mapped to $[0, 1]$ before weighting ($B' = 2.001$); both levels aggregate one term per certified episode (Lemma 2) and share δ ($\delta/2$ per level); and the Hoeffding leg is anytime over the 200 checkpoints.

Table 12: Firing on the deterministic-legacy benches; long-K bounds are certificates (frozen posterior, prior-split λ).

Inner level, long-K runs	no crossing at any $\sigma_Q \in \{1, 3, 6\}$ or checkpoint; from $K =$ 1,000 to 200,000 the certificate rises from -602.0 to -157.0 per episode at $\sigma_Q = 1$ (-601.9 to -157.0 and -602.4 to -157.4 at $\sigma_Q = 3, 6$), below a coverage deficit of at most -146.8
Meta level	never fires in any seed; true meta advantage over the legacy: mean -0.10 per episode, 95% CI $[-0.98, +0.77]$ (20,000 paired Monte Carlo episodes per arm)
Per-level firing order	not observable: neither level fires on these benches

Reading. The meta-level refusal is correct: the true meta advantage is indistinguishable from zero. With a deterministic legacy and rewards mapped to $[0, 1]$ the coverage deficit is non-positive (Effect 5), and no level fires at any checkpoint. The measured inner deficit is at most -146.8, -146.8 and -147.2 per episode at $\sigma_Q = 1, 3, 6$. Firing order and firing across σ_Q are therefore not observable on these benches. The three arms share the prior-split temperature to within 0.1%, and their final inner certificates differ by at most 0.4 per episode: the deficit, which σ_Q barely moves, sets the level. AEGIS gates on the $(1-\delta)$ lower bound of the per-level gain rather than on its empirical mean; the resulting closed-gate stretch and its certified-safety price are discussed where the gate is introduced (Section 6.1).

J Extended Limitations, Graph Provenance and Future Work

Limitations. Three limitations stand out. The first is within-level exogeneity (Assumption 2). The assumption is exact for batch domains whose inner steps are parallel or exchangeable units under one decision (wafer lots, field plots under one input allocation, per-batch quality samples in process control), and state carried *across* timescales is covered as well, since it is modelled by the parent-level context $C^{(\ell)}$, which the parent action shifts. Its boundary is treatment titration: in a stepwise dose adjustment, each inner step reacts to the response of the previous one, so the within-level context process is *endogenous* in the within-level actions, which is precisely what the assumption excludes; the level-factored weight (3) then no longer identifies the deployed policy's value from logged data. With an engineered self-destabilising candidate the level-factored target overstates the candidate's value; at the data-free temperature the bound's slack absorbs this overstatement at AR(1) coefficient $\rho = 0.9$, but not in a near-unit-root variant ($\rho = 0.95$, shock half-life 13.5 instead of 6.6 inner steps; state penalty $c = 0.7$ instead of 0.3), and the certificate then falsely certifies and would trigger a false handover in every trial (measurements below). Titration and process control therefore need the carried state modelled in $C^{(\ell)}$, as our instantiation does. The natural route to the endogenous case replaces the level-factored weight by trajectory-product weights within each episode, whose overlap constant compounds over the I_ℓ inner steps; that extension is left to future work. The second limitation is the mechanism class: additive-noise models (linear with NIG, nonlinear with online Gibbs), with the non-additive case exercised only as a misspecification stress test rather than modelled directly. The third is evaluation scope: exact validity verification remains synthetic, since checking that a certificate holds requires ground-truth policy values, which external logs do not provide; on a real log, coverage can be checked (Section 7.3), exact validity cannot.

Measured boundary of Assumption 2. A one-dimensional linear state $s_{i+1} = \rho\, s_i + \beta\, a_i + \varepsilon_i$ (an AR(1) process within the level; for $\beta = 0$ the autoregressive coefficient ρ is the lag-1 autocorrelation of the state and sets its persistence: a shock decays to half after $\ln 2 / \ln(1/\rho)$ inner steps, 6.6 at $\rho = 0.9$ and 13.5 at $\rho = 0.95$, against $I_\ell = 6$ steps per level) placed inside one level separates the two within-level demands of the assumption ($\beta = 0$: exogenous serial dependence; $\beta > 0$: state driven by the within-level actions). The reward $a_i s_i - c\, s_i^2$ is clipped to $[-4, 2]$ and used signed (no rescaling); with the uniform logger over $\{-1, 0, 1\}$ the level-factored weight (3) is at most $B_w = 3$, so the declared loss range gives $B' = 3 \cdot 6 + 10^{-3} = 18.001$, and the temperature is the data-free $\lambda = 1/(2B') = 0.0278$, fixed before any data as Theorem 1 requires (KL $= 0$, $K = 4000$, $I_\ell = 6$,

$\delta = 0.05$; values per inner step). The column "target" is what the level-factored certificate estimates, $\mathbb{E}_\mu[\frac{1}{T_\ell}\sum_i w_i r_i]$ (2×10^6 Monte Carlo episodes); the certificate is sound only where it does not exceed $V(\pi)$ by more than the bound's slack.

ρ	β	c	trials	$V(\pi)$	target	mean V_{lower}	false cert.
0.9	0.0	0.3	60	0.51	0.51	0.01	0%
0.9	0.8	0.3	60	0.17	0.52	0.03	0%
0.95	0.8	0.7	40	−1.95	−0.31	−0.99	100%

Serial dependence alone never falsely certifies: the target equals $V(\pi)$ (0 of 60 trials at each $\rho \in \{0, 0.5, 0.9\}$; violation rate below 4.87% at one-sided 95% confidence). Action-driven state breaks identification. At $\rho = 0.9$, $\beta = 0.8$ the target overstates $V(\pi)$ by 0.35 per step, less than the logarithmic-smoothing pessimism at $\lambda = 1/(2B')$ (about 0.5 per step, not shrinking in K), so no trial falsely certifies: coverage through slack, not detection. A smaller temperature exposes the failure: the closed-form $\lambda^\star$ fitted on the certification sample (outside the theorem) falsely certifies in $60/60$ trials, and $\lambda^\star$ rounded to the grid of Remark 6 ($\lambda = 0.0066$) in $58/60$. In the near-unit-root variant (third row: $\rho = 0.95$, $c = 0.7$) the overstatement is 1.64 and exceeds the slack: every trial falsely certifies, and since $V_{\text{lower}} = -0.99 > V(\mu) = -1.17 > V(\pi) = -1.95$, every trial would trigger a false handover. The candidate is engineered and self-destabilising ($\pi = \text{sign}(s)$, myopically attractive); at $\rho \in \{0, 0.5\}$, $\beta = 0.8$ the target lies below $V(\pi)$ and the certificate stays valid.

Two assumptions are stress-tested by the paper's experiments: posterior-class misspecification in Appendix H.1 (a shape-misspecified NIG backend against the matched RFF-GP class costs about 34 reward, not validity, since the bound holds for every posterior) and the additive-noise violation (frozen posterior, 3 seeds; App. H.5). Shared within-level confounding enters only Bin-3L and Depth-Bin, with a frozen posterior (App. K); the working model of Remark 2 is not tested under it.

NCTS is a baseline; the NCCB class admits a wider design space along two axes. First, sharper bounds: PRISM's multiplicative overlap penalty $B^{\mathcal{I}}$ is sidestepped level by level by the progressive workflow of Section 6.1 but remains loose under one-shot full deployment, and self-normalised importance sampling [9, 31], PAC-Bayes-Bernstein refinements on the per-mechanism variance structure, and self-normalised maximal inequalities [5] are promising tighteners. Second, alternative NC-X hosts, including NC-UCB [17], NC-IDS [26], NC-PB (direct PAC-Bayes minimisation), NC-DOpt [4], and pessimistic warm-starts [2], all wrapped by AEGIS (Algorithm 2) via its host-agnostic interface. Further open problems (AEGIS-conservativeness reduction, graph learning, continuous domains, adaptive episode length, confounder estimation, real-data deployment) are collected in App. L.

Where the graph comes from. When is such an I-map available? In the motivating domains the superset structure is typically documented rather than discovered: process engineering fixes mechanism structure in plant flowsheets and P&IDs, clinical pharmacology in trial-established dose-response pathways, and agronomy in managed input experiments. Each such document is signed off by a domain expert and may safely contain edges the true mechanism does not use, since only an I-map is required; none must get the mechanism form right, that being a working model whose misspecification costs tightness and not validity (Remark 2). The strongest case is *randomised logging*, as in micro-randomised trials and A/B logs: there the logged action probabilities depend on nothing outside the log by experimental design, so the logging condition of Section 6.2 is met by construction. The fully unknown-graph case is the graph-posterior extension sketched in App. L.

K Depth: the $L = 3$ bench and the depth sweep

Most body experiments run at $L = 2$. This appendix asks two separate questions at greater depth. First, does NCTS still behave as Algorithm 1 intends when a third level is added (Section K.2)? Second, how does the certificate of Theorem 1 behave as L grows, given that the importance weight of Eq. (3) is a product over handed-over levels (Sections K.3, K.4 and K.5)? Throughout, $\delta = 0.05$ unless stated otherwise, effect sizes are Cohen's d with pooled standard deviation, and tests are Welch tests over seed means. Table 13 records, for every exhibit, whether the posterior is learning or frozen, since a frozen posterior sets the KL term to zero by construction.

Table 13: Depth exhibits and their posterior regime. "Frozen": the posterior is fitted on a disjoint offline split and pinned as its own prior, so $\mathrm{KL} = 0$ exactly. "Learned": a non-zero mechanism KL is charged.

Exhibit	Depth	Posterior	Certificate crosses zero?
Reward triple (§K.2)	3	online NCTS (learning)	not a certificate
Depth-Grid, deterministic legacies (§K.3)	3	frozen	no, up to $K = 2 \times 10^5$
Per-scope compounding, Depth-Grid (§K.4)	3	learned	no, at $K = 2{,}000$
Bin-3L (§K.5.1)	3	frozen	yes, $9/9$ and $30/30$
Depth-Bin (§K.5.2)	$3, 4, 5$	frozen	yes, $120/120$

The live-KL crossing, in which the certified gain crosses zero with the posterior still learning, is an $L = 2$ result and is reported with the other crossing exhibits; nothing in this appendix depends on it.

K.1 The three-level Depth-Grid bench

The three-level bench Depth-Grid extends Γ-SCM by an outer strategic level whose action shifts the distribution of the meta-context. In the paper's inside-out indexing, $C^{(3)} = T$, $A^{(3)} = M$ (outer, strategic), $C^{(2)} = Q$, $A^{(2)}$ (mid), and $C^{(1)} = S$, $A^{(1)} = B$ (leaf), with hierarchy $(T, M) \to (Q, A^{(2)}) \to (S, B) \to Y \to R$. Under an intervention on all three actions the structural equations are

$$T = \varepsilon_T, \tag{32}$$

$$Q = \gamma\big[(1-\eta)\,1.5\,M + \eta\,2\tanh(2M)\big] + \varepsilon_Q + 0.3\,U, \tag{33}$$

$$S = 1.5\,\gamma\,A^{(2)} + 0.5\,Q + \varepsilon_S + 0.3\,U, \tag{34}$$

$$Y = B \cdot S + 0.3\,U + \varepsilon_Y, \qquad R = Y + \varepsilon_R, \tag{35}$$

with $U \sim \mathrm{Ber}(0.5)$ a latent noise term, independent Gaussian noises $\varepsilon_T \sim \mathcal{N}(0, 0.3^2)$, $\varepsilon_Q \sim \mathcal{N}(0, \sigma_Q^2)$, $\varepsilon_S \sim \mathcal{N}(0, \sigma_S^2)$, $\varepsilon_Y \sim \mathcal{N}(0, 0.08^2)$, $\varepsilon_R \sim \mathcal{N}(0, 0.5^2)$, and actions $M, A^{(2)}, B \in [-2, 2]$. The Γ-Stress preset uses $\gamma = 1$, nonlinearity $\eta = 1$, $\sigma_Q = 3$ and $\sigma_S = 0.04$; the episode has one outer step, four mid steps per outer step and five leaf steps per mid step, so $N_L = 20$ leaf rewards per episode. The leaf mechanism is bilinear, $\mathbb{E}[Y \mid B, S] = BS$ up to the offset from U, so the optimal leaf action depends on the realised leaf context, and the realised mid context Q enters S with coefficient 0.5.

The environment implements the context semantics of Algorithm 1: an ancestor context realised at its decision time is held fixed for all descendant decisions within that decision, so each leaf context is drawn from f_S at the *observed* Q and the committed $A^{(2)}$. As a check, the ordinary least-squares slope of the leaf S on the observed Q is 0.5013 against the structural coefficient 0.5. The declared reward envelope is $R \in [-25, 25]$ per leaf step; it was fixed before the runs below but calibrated on a pilot scan of corner interventions in the same domain, which observed $R \in [-20.2, 20.5]$, so it is a pilot-calibrated envelope rather than one derived from the model. Every certificate on this bench uses R clipped to this envelope inside the SCM, so Assumption 4(i) holds by construction.

K.2 Reward at $L = 3$

We compare three arms on the Γ-Stress preset ($K = 2{,}000$ episodes per seed, 10 seeds, all arms in one batch) with a matched function class (random Fourier features, 64 features, NIG posterior) and identical one-draw Thompson exploration in every arm: the nested factorised agent (pure NCTS), a joint-commit ablation that commits all three actions at episode start on a single value model over $(T, M, A^{(2)}, B)$ and never observes Q or S, and a structure-free flat agent with the factorisation switched off.

The factorisation axis carries most of the gain at depth 3: the 10 nested and 10 flat seed means separate completely. One caveat applies to this contrast. The flat arm has no mechanism model of $Q \mid M$ or $S \mid (Q, A^{(2)})$ and imputes unrevealed deeper contexts at 0 at decision time, so an unquantified part of the $+133.91$ is a handicap of the baseline rather than pure factorisation.

On this bench, re-deciding on realised contexts beats committing all levels at episode start by $+12.77$ per episode. Separation is near-complete rather than complete: 9 of 10 nested seed means exceed the

Table 14: $L = 3$ reward triple: mean episode reward over all $K = 2{,}000$ episodes and over the last 25%, mean $\pm$ sd over 10 seeds.

Arm	Mean reward	Last 25%
Nested (NCTS)	$+180.52 \pm 6.90$	$+181.91 \pm 7.00$
Joint commit	$+167.75 \pm 1.37$	$+171.14 \pm 0.82$
Flat (no factorisation)	$+46.61 \pm 6.42$	$+51.16 \pm 10.88$
Nested vs. flat: $\Delta = +133.91$, $d = 20.08$, $t(17.9) = 44.91$, $p < .001$		
Nested vs. joint: $\Delta = +12.77$, $d = 2.56$, $t(9.7) = 5.74$, $p < .001$ (permutation $p = 1.2 \times 10^{-4}$)		
Joint vs. flat: $\Delta = +121.14$, $d = 26.09$, $t(9.8) = 58.34$, $p < .001$		

joint maximum, and one nested seed ($+161.15$) lies below the joint minimum. We report this as a measurement on this bench and do not claim it as a law in L; the $L = 2$ ablation of Section 7.2.2 remains the primary evidence for recursive re-decision. The contrast depends on the context semantics above: a run in which every descendant decision redraws the whole SCM, which makes the mid context causally inert, reverses its sign, so the joint-commit comparison is informative only under the semantics Algorithm 1 specifies. Reproduction is statistical rather than bitwise, because one posterior draw consumes an unseeded random stream; means and verdicts reproduce, exact per-seed values do not.

K.3 Certificate on Depth-Grid: why no scope crosses

On the bench of Section K.1 the legacies are deterministic at all three levels. A deterministic legacy has $\mu^{(\ell)}(a^{\mu} \mid \tilde{c}^{(\ell)}) = 1$ at every logged action, so every importance weight is at most 1. Since rewards are mapped to $[0, 1]$ before weighting, the coverage deficit of the gain bound is then non-positive at every λ and every K (Appendix D, Effect 5), and no episode budget can produce a crossing. A certified long-K run measures the trajectory (one seed, 2×10^5 certified episodes after 2×10^5 offline legacy episodes, reward clipped to $[-25, 25]$). The posterior is fitted on the offline split and frozen, so $\mathrm{KL} = 0$; the temperature $\lambda = 2.85 \times 10^{-3}$ is Eq. (9) evaluated on the offline split and held constant; $B' = B_w + 10^{-3} = 2.001$ with $B_w = 2$; every level aggregates one term per certified episode (Lemma 2); the three levels share $\delta = 0.5$ ($\delta/3$ each), and the Hoeffding leg is anytime over the 400 checkpoints. The largest coverage deficit over all checkpoints is -485.6 (outer), -495.0 (mid) and -245.3 (leaf) per episode, and no level crosses. From the first checkpoint ($K = 500$) to the last, the certified bounds rise from $-2{,}305.7$, $-2{,}303.9$ and $-2{,}048.3$ to -513.6, -511.8 and -261.8 per episode (outer, mid, leaf), a 4.5-, 4.5- and 7.8-fold contraction towards the respective deficits; the leaf, whose deficit is smallest in magnitude, is the tightest at every checkpoint. The non-crossing is a property of the legacy and not of the depth; Section K.5 holds the depth at three and replaces the legacy with one that randomises.

K.4 Compounding: the KL adds, the overlap multiplies

The two penalties of Theorem 1 scale differently in L. At levels the legacy still runs, the deployed and logging policies coincide and their ratio cancels in Eq. (3), so a single-level scope pays one overlap factor at any depth, while one-shot full handover pays the product $\prod_{\ell} B_w^{(\ell)}$. The regulariser, by contrast, depends on the scope only through $\max(\mathcal{I})$ (Definition 2): the deepest single-level scope $\{L\}$ and the full scope $\{1, \ldots, L\}$ charge an identical KL, so at a temperature $\lambda \propto 1/B^{\mathcal{I}}$, as used below, the regulariser of the full scope exceeds that of $\{L\}$ by exactly the overlap factor $\prod_{\ell<L} B_w^{(\ell)}$, while the width of both also carries a legacy-side concentration term that does not depend on the scope.

We measure this on Depth-Grid at the Γ-Stress preset of Section K.1, with its deterministic legacies replaced at every level by a legacy that samples uniformly at random from the $n_a = 7$ points of the level's action grid, the same grid the certified policy plays on, so that every scope has overlap (10 seeds, $K = 2{,}000$, $\delta = 0.05$, 7-point action grids on $[-2, 2]$, reward clipped to $[-25, 25]$). The certified policy is the nested agent with linear features, trained on-policy for 400 episodes separately from the certification data; every scope is then certified from the same logged legacy stream. Each seed retrains the policy and draws a fresh legacy log. The temperature is the data-free $\lambda = 1/(2B')$ with $B' = B_w^{\mathcal{I}}(R_{\mathrm{hi}} - R_{\mathrm{lo}})$, fixed per scope before the data, as on Contract (App. H). The legacy

propensity is exactly $1/n_a$ at every logged action, so the per-level weight is $n_a\,\pi^{(\ell)}(a)$ and the per-level ceiling is $B_w^{(\ell)} = n_a = 7$. The bound uses the declared ceiling $B_w^{\mathcal{I}} = \prod_{\ell\in\mathcal{I}} B_w^{(\ell)} = 7^{|\mathcal{I}|}$ in B' and λ, not the largest weight observed in the log. At $\{1\}$, $\{2\}$ and $\{1,2\}$ the largest observed weight equals this ceiling in every seed: the certified policy is deterministic at levels 1 and 2, so every logged action that matches its choice has weight 7 per level. At the scopes containing level 3 it stays below the ceiling, because the policy is stochastic there (largest action probability below 1).

Table 15: Per-scope certificates at $L = 3$: declared and realised overlap constant $B_w^{\mathcal{I}}$, mechanisms in $\Phi_{\mathcal{I}}$, regulariser and width ($K = 2{,}000$, $\delta = 0.05$, data-free $\lambda = 1/(2B')$; mean $\pm$ standard deviation over 10 seeds, realised $B_w^{\mathcal{I}}$ as the range over seeds). Levels are indexed inside out, so $\ell = 1$ is the leaf.

Scope $\mathcal{I}$	$B_w^{\mathcal{I}}$ declared	$B_w^{\mathcal{I}}$ realised	$\Phi_{\mathcal{I}}$	KL (nats)	Width
$\{1\}$	7	7.000	Y, R	386 ± 225	$2{,}755 \pm 1{,}573$
$\{2\}$	7	7.000	S, Y, R	481 ± 224	$3{,}425 \pm 1{,}571$
$\{3\}$	7	3.938 to 4.813	Q, S, Y, R	496 ± 224	$3{,}530 \pm 1{,}571$
$\{1,2\}$	49	49.000	S, Y, R	481 ± 224	$23{,}795 \pm 11{,}000$
$\{1,2,3\}$	343	192.9 to 235.8	Q, S, Y, R	496 ± 224	$171{,}507 \pm 76{,}977$

The comparison is clean because the scopes pair up at identical regularisers (Table 15). In every seed, $\{2\}$ and $\{1,2\}$ charge the same KL, and their widths stand in the ratio 6.93 ± 0.05 (range 6.83 to 6.97) against a predicted 7; $\{3\}$ and $\{1,2,3\}$ charge the same KL, and their widths stand in the ratio 48.47 ± 0.33 (range 47.77 to 48.77) against a predicted $7^{L-1} = 49$. Within each pair nothing but the overlap product separates the two scopes. At the data-free temperature the regulariser $(\mathrm{KL} + \log(2/\delta))/(\lambda K)$ is linear in B', so the regularisers stand in the ratios 7 and 49 in every seed, up to rounding; the small shortfall of the widths comes from the legacy-side concentration term (30.37 per episode), which does not depend on the scope. The ratios are therefore stable although the KL itself varies by about 225 nats across seeds with the trained policy. With the policy held fixed, 10 fresh legacy logs leave every width unchanged, since no term of the width reads the data, and move the lower bounds by at most 62.6 (standard deviation at most 17.8, both at $\{1,2,3\}$), so the spread across seeds comes from training, not from the log. Each single-level scope pays one factor of 7, at mean widths from 2,755 to 3,530; their spread comes from the regulariser alone, as $\Phi_{\mathcal{I}}$ takes in one more context mechanism per level, and we claim no fixed multiple for it. The theory predicts the same 7^{L-1} separation at every depth; we quote the measured point at $L = 3$ rather than the extrapolation. No scope crosses zero at $K = 2{,}000$ on this bench, in any seed (the highest lower bound is -916.9, at $\{1\}$). AEGIS never pays the full product: it certifies each level at single-level scope, so the full-scope factor is not charged; two levels deployed in turn are certified one at a time, the second on a buffer collected after the first handover (App. E).

K.5 Firing at depth: Bin-3L and Depth-Bin

Both benches below use binary action sets and uniform legacies, so that $B_w^{(\ell)} = 2$ per level exactly and the composed ceiling is 2^L. In both, the posterior is fitted on a disjoint offline split and frozen, so $\mathrm{KL} = 0$ by construction; these are therefore crossings of the overlap, Hoeffding and logarithmic-smoothing terms, not of the regulariser. Each per-level certificate is evaluated at δ/L, so that the L per-level statements tested by the gate hold jointly at $\delta = 0.05$; the composed certificate of Section K.5.1 is a single statement, not tested by the gate, and is evaluated at δ. Bounded rewards (Assumption 4(i)) hold exactly because contexts are squashed by $\tanh$ and the environment clips rewards to the declared range $[-3, 3]$; Assumption 4(ii) holds by measurement, the realised weight attaining the declared ceiling at every level and seed without exceeding it.

K.5.1 Bin-3L: a three-level NCCB with a randomising legacy

The nesting follows the two-level nested-timescale testbed of Li et al. [16], extended by one level and re-implemented from the published description: a cohort action reshapes the weekly context, a weekly action reshapes the daily context, and the daily action produces a bounded reward, with no separate outcome Y,

$$\begin{aligned} &X_0 \text{ exogenous}, \quad X_1 = f_1(X_0, A_0, \varepsilon_1), \quad X_2 = f_2(X_1, A_1, U, \varepsilon_2), \\ &R = \mathrm{clip}_{[-3,3]}\Big(\textstyle\sum_\ell b_\ell\, A_\ell\, (\vartheta_\ell^\top X_\ell) + c\,U + \varepsilon_R\Big), \end{aligned} \tag{36}$$

with a latent $U = \tanh(Z)$, $Z \sim \mathcal{N}(0,1)$, drawn per leaf step and shared by the leaf context X_2 and the reward with loading $c = 0.3$ (confounding within the leaf level only, Assumption 1), $A_\ell \in \{-1, +1\}$ (outer level indexed 0 here), four weeks of seven days ($N_L = 28$ leaf steps per episode), and the interaction feature map $[1, x, a, ax]$. Uniform legacy at every level, $K_{\mathrm{cert}} = 40{,}000$ certified episodes, λ chosen from an a-priori 16-point geometric grid closed by a union bound costing $\log 16$ nats.

Per-level certificates. All nine per-level certified gains (three seeds by three levels) cross zero, at K between $4{,}597$ and $8{,}517$; over ten seeds the crossing holds 30 times out of 30, at K between $4{,}411$ and $9{,}280$. A no-interaction control (the same bench with ax removed from the feature map) crosses at no level and no seed, so the gate is seen both to fire and to decline.

Composed policy. A per-level certificate flips one level and leaves the rest logging, so its overlap constant is 2 regardless of L and nothing compounds. The question of how the certificate behaves as the weight compounds is answered by the composed three-level policy, whose per-leaf-step weight is the product of three per-level weights with ceiling $2^3 = 8$, realised exactly. It crosses zero at a K in $(9{,}477, 15{,}315]$ on all three seeds, between 1.6 and 2.5 times the median per-level budget rather than a blow-up, with the composed certificate at δ and the per-level ones at δ/L. At $K = 40{,}000$ the certified gain is $+7.07$ to $+9.83$ per level and $+17.63$ to $+22.92$ composed, against measured advantages of 14.11 to 16.41 and 42.18 to 47.03: the bound recovers 49% to 60% of the measured advantage per level and 42% to 49% composed.

What is designed in. The bench is synthetic and its favourable properties were designed and then verified. The context-dependence of the optimal action is built in by the interaction term; that the fitted policy recovers it is measured (agreement with the oracle 0.992 to 0.999 over ten seeds). The uniform legacy is also the randomising legacy that minimises B_w on two actions. Raising only the declared B_w from 2 to 4 multiplies the crossing K by 4.86 to 6.71 (mean 5.37; $K_{\mathrm{cert}} = 80{,}000$ so that every level crosses at $B_w = 4$), a rise of the order of the B_w^2 factor the penalty predicts. Against a competent floor-softmax incumbent aligned with the true reward direction (floor 0.25, declared $B_w = 4$) no level crosses at $K = 40{,}000$, but all three cross at $K = 78{,}049$ to $84{,}598$ (one seed), about 13 times the uniform-incumbent budget, of which a factor of about 5.4 is the overlap constant and the rest the smaller advantage left to certify. This bench is a synthetic demonstration of the mechanism, not evidence of field-deployability: priced at published effect sizes of two mobile-health settings (the Intern Health Study [19, 33] and the nested-timescale testbed of Li et al. [16]), the certificate would need between about 300 and 3.5×10^5 times the data those settings provide.

K.5.2 Depth sweep at $L = 3$, 4 and 5

A second binary-action bench, Depth-Bin, varies the depth directly with the mechanisms of Eq. (36), including the per-leaf-step U (10 seeds per depth, interaction strength $b = 2$, declared reward range $[-3, 3]$, offline, legacy and certification splits of $1{,}500$, $3{,}000$ and $6{,}000$ episodes).

Table 16: Depth sweep with frozen posterior (KL $= 0$), $\delta = 0.05$ (δ/L per level), 10 seeds per depth. "Median K" is the range over levels of the per-level median crossing episode. Clip: fraction of rewards clipped to the declared range.

L	Levels firing	Median K at crossing	Mean certified gain	Clip	Composed ceiling 2^L
3	30/30	1,930 to 3,509	$+7.6$	0.196	8
4	40/40	136 to 182	$+76.7$	0.259	16
5	50/50	134 to 179	$+162.3$	0.320	32

Every one of the 120 per-level certificates crosses zero (Table 16). The crossing budget falls with L on this bench, and we claim no trend: the deeper benches carry more total signal, since each added level contributes its own action-by-context interaction, so depth buys both a larger overlap product and a larger available gain. Clipping binds on a fraction of rewards that grows with depth. What the sweep shows is that the per-level certificate does not degrade with depth on this bench: the same gate at the same joint δ fires at every level of a five-level hierarchy.

Across these two benches, per-level overlap rather than depth tracks feasibility. Taken together, the two benches separate depth from overlap. Nothing crosses on Depth-Grid (Section K.4), where

$B_w^{(\ell)} = n_a = 7$ makes the full-scope product 7^L (343 at $L = 3$); everything crosses to $L = 5$ on Depth-Bin, where $B_w^{(\ell)} = 2$ makes it 2^L. L enters only as the exponent, so a practitioner controls feasibility through the per-level action resolution, which is a design-time choice, and not through the depth, which the problem fixes. The benches differ in more than this constant (reward scale, interaction strength, horizon), so we report this as the pattern across both benches and not as the isolation of a single variable.

L Further Open Problems

Appendix J discusses the two open directions closest to the present results (tighter bounds and alternative NC-X hosts). The following are further open problems that the NCCB framework opens but which are out of the present paper's scope.

- **Reducing AEGIS's certified-safety conservativeness.** AEGIS gates on a $(1-\delta)$ lower bound, so the handover can stay closed while $V(\pi_{Q_K}^{(\ell)})$ already exceeds $V(\mu_{\text{legacy}}^{(\ell)})$ in expectation. Tightening the gap (PAC-Bayes-Bernstein refinements that absorb low per-step variance, dependence-aware overlap bounds, or a calibrated $(\varepsilon_\ell, \delta)$-schedule that anneals the guarantee level as K grows) without losing the anytime guarantee is open.
- **Joint graph-structure and parameter learning.** Replace the known-graph assumption with a posterior over graphs (BayesDAG [1] or DAG-NoCurl [35]); the resulting bound decomposes as $\mathrm{KL}(\hat{Q}(\mathcal{G}) \| Q_0(\mathcal{G})) + \mathbb{E}_{\hat{Q}(\mathcal{G})}[\sum_X \mathrm{KL}(Q_X \| p_{0,X})]$ marginalised over the graph posterior; online graph-discovery is open.
- **Continuous variable domains.** Extend the analysis from discrete-grid action and meta-action spaces to continuous domains via covering-number / metric-entropy arguments, with parametric policy networks π_θ and $\mathrm{KL}(\pi_\theta \| \pi_0)$ as the PAC-Bayes complexity term; in the spirit of PAC-Bayes deep RL but with the mechanism-factorised prior preserved.
- **Adaptive episode length.** Turn the between-episode problem into an optimal-stopping / semi-MDP with switching costs, deciding when to commit a fresh meta-action versus continuing the current one.
- **Latent confounder estimation.** Combine the PAC-Bayes mechanism posteriors with instrumental-variable or proximal causal inference to estimate U-loadings rather than absorbing them into the noise terms; tighten the bound on environments with identifiable confounders.
- **Real-data deployment.** The external-data checks of App. M exercise the certificate on two public logs, but neither can settle gate firing: both are value certificates well below their logging policies' measured values, and neither log randomises over a candidate with a certifiable advantage at its sample size. The most direct remaining test of the safe-deployment workflow is a logged industrial or clinical deployment (process control, semiconductor yield, demand-response) with randomised legacy logging and a candidate worth certifying.

M External-Data Checks

This appendix gives the constructions, caveats and full tables behind Section 7.3. Both logs are public and used under their licences (CC BY 4.0 in both cases); neither contains data we generated, and certificates need only the logged (x_t, a_t^μ, R_t) tuples and the logging propensities.

M.1 Open Bandit Dataset

Construction. The Open Bandit Dataset [28] (campaign `all`, 10,000-round sample per policy) logged the same platform in the same week under a uniform Random policy and a Bernoulli Thompson Sampling (BTS) policy on A/B-split traffic. The *forward* direction certifies the BTS policy from the Random log; its target is the empirical action frequency $\bar{\pi}_{\text{bts}}$ of the BTS log (a fixed policy, so $\mathrm{KL} = 0$), and its weight ceiling $B_w = 80 \times 0.1105 = 8.84$ is exact because uniform logging fixes $\mu = 1/80$ at every round. The *reverse* direction evaluates the Random policy from the BTS log; its $B_w = 277.8$ is the empirical propensity floor of the BTS log rather than an Assumption-level ceiling, so it is a poor-overlap contrast and not a certificate. Each direction has

81 cells ($K \in \{100, 300, 1000, 3000\}$ with 20 subsample seeds each, plus the full log), $\delta = 0.05$, and λ pre-fixed per K by the data-free rule $\lambda = \min\{(1 - \eta_{\text{clip}})/B_w,\ \sqrt{2\log(1/\delta)/(KB_w^2)}\}$, $\eta_{\text{clip}} = 0.05$. On-policy references are $\widehat{V}(\pi_{\text{bts}}) = 0.0042$ $[0.0030, 0.0054]$ and $\widehat{V}(\pi_{\text{rand}}) = 0.0038$ $[0.0027, 0.0050]$ (1000-resample bootstrap).

Table 17: Open Bandit Dataset: certified V_{lower} (mean over 20 subsample seeds; 1 cell at $K = 10{,}000$), $\delta = 0.05$. Violations: cells with V_{lower} above the on-policy estimate. The zero-click control sets every click to 0.

K	forward V_{lower}	zero-click control	reverse V_{lower}	reverse/forward	violations (fwd / rev)
100	-2.3767	-2.3810	-74.8036	31.5	0/20 / 0/20
300	-1.3108	-1.3153	-41.3244	31.5	0/20 / 0/20
1000	-0.6972	-0.7031	-22.0873	31.7	0/20 / 0/20
3000	-0.3970	-0.4012	-12.6026	31.7	0/20 / 0/20
10000	-0.2134	-0.2182	-6.8534	32.1	0/1 / 0/1

Reading. Zero violations in all 162 cells rule out gross miscalibration only: with clicks in $\{0, 1\}$ at a click-through rate of about 0.4%, a forward violation would need the bound to be biased upward by 52 to 570 times the on-policy value. The cells are also nested (subsamples of one 10,000-round log), so $0/81$ is not 81 independent successes. The forward log-log slope is -0.52 (OLS over the five K values), but the zero-click control, which contains no information beyond K and B_w, has slope -0.518 and reproduces the curve to within $+0.004$ to $+0.006$ at every K: the slope is consistent with the $O(1/\sqrt{K})$ rate, not a measurement of it. The reverse direction is 31.5 to 32.1 times looser at matched K, the overlap cost measured on external data. No positive certificate is reachable at this n, by design of the dataset, and none is claimed. The log is single-level, so it exercises the certificate and not the nested workflow.

M.2 HeartSteps V1

Construction. HeartSteps V1 [12] is a public micro-randomised mHealth trial (44 adults enrolled, 37 in the public release, nominally 42 days) with activity suggestions randomised at five decision points per day, $P(\text{send} \mid \text{available}) = 0.6$ by design. We instantiate two timescales inside the trial's randomisation: the day's slot-1 suggestion as the meta-action, evaluated on the daily step return, and the slot-2 to slot-5 suggestions as the inner level, evaluated on 30-minute proximal step returns (rewards are $\log(\text{steps} + 0.5)$). Slot-1 as meta-action is a designation of a decision point rather than a distinct intervention channel, and we disclose it as such: the trial's daily planning randomisation is not part of the public release. The estimand is availability-conditional (the standard MRT causal-excursion estimand), with user-day episodes as the independence unit: 1,319 outer and 1,641 inner episodes, each split $50/50$ into a fit half and a held-out evaluation half. The importance weights are bounded by $1/0.4 = 2.5$ exactly by design, and the LS temperature $\lambda^\star(K)$ is derived from the disjoint fit half, so the evaluation sample selects neither the policy nor its own λ. The one quantity that still reads the whole log is the reward range when R_{hi} is taken as the empirical maximum, whose maximising inner episode falls in the evaluation half; the headline numbers below therefore use a-priori physiological caps, and the empirical-max figures are reported as a sensitivity. Three structural caveats are stated with the result. All episodes come from the same 37 users (16–46 outer and 36–63 inner episodes each, with both halves drawing on all users), so cross-episode independence is an approximation at the user level and a cluster-robust analysis over the 37 users is the natural check. Inner episodes hold 1 to 4 available slots (mean 3.14) against the nominal $N_L = 4$, a convention shared by the certificate, by $\widehat{V}(\mu)$ and by the trivial floor, so their comparison is internally consistent while the fixed N_L is out-of-model for this trial. Finally, if suggestions carry over within a day, the inner certificate bounds the causal-excursion value, that is, deviating to π_Q at one randomised decision point with the remainder of the day left under μ, rather than the value of deploying π_Q at every slot; the outer level, with one decision point per episode, is immune to this distinction.

Results (Table 18, δ=0.05). Each level's trivial floor is $\log(0.5)$ per decision point (-0.69 outer, -2.77 inner), the a-priori minimum of any policy's value. The outer certificate is positive on this real log: $V_{\text{lower}} = +5.63$ and $+5.39$ per episode at full sample (660 evaluation episodes) under the two a-priori caps, $+5.68$ under the in-sample cap; under that in-sample cap, the one the K sweep was computed with, it clears the trivial floor from $K \approx 100$ and is positive in the mean from $K = 200$.

The inner certificate sits $+2.43$ and $+1.33$ above its trivial floor under the a-priori caps ($+3.36$ under the in-sample cap); its absolute sign is cap-dependent, positive only under the in-sample cap, and is disclosed rather than optimised over. Both levels contract at approximately $1/\sqrt{K}$. The randomised parent action shifts the inner-context distribution by $+0.18$ log-steps (SE 0.11, $z = 1.59$ at the user-day randomisation unit) and the inner day-mean response by $+0.11$ (SE 0.11, $z = 1.00$): point estimates consistent with the nested interface, not individually significant at the randomisation unit.

What is and is not claimed. These are value certificates plus contraction, not certified improvement: the trial's treatment effects are within about 1.5 standard errors of zero, so no valid 95% bound can place $V(\pi)$ above the logging policy's measured value ($\widehat{V}(\mu) = +8.78$ outer, $+12.54$ inner) at this sample size, at either level, and none is claimed.

Table 18: HeartSteps V1: per-episode V_{lower} on held-out evaluation subsamples of size K, mean $\pm\, 1\sigma$ across 20 subsample seeds (1 seed at full sample: K=660 outer, K=821 inner), δ=0.05. The K columns use the in-sample empirical-max reward cap, under which the sweep was computed; the last column gives the primary a-priori-capped full-sample certificates (50,000 / 100,000 steps per day outer, 10,000 / 20,000 steps per 30 min inner), which sit $+6.33$ / $+6.08$ (outer) and $+2.43$ / $+1.33$ (inner) above the trivial floors -0.69 and -2.77.

K	25	50	100	200	400	full	a-priori
outer	-18.24 ± 5.20	-7.58 ± 1.55	-0.11 ± 0.89	$+2.90\pm0.46$	$+4.77\pm0.25$	$+5.68$	$+5.63/+5.39$
inner	-103.90 ± 10.68	-53.48 ± 3.24	-27.08 ± 1.58	-13.02 ± 1.11	-4.99 ± 0.65	$+0.59$	$-0.35/-1.44$

The certificate treats user-days as independent episodes, and since suggestions within a day may influence later availability, the inner level sits in the endogenous within-level regime that Assumption 2 excludes; the coverage claim therefore rests on the Open Bandit check.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state three contributions, namely the NCCB problem class and NCTS algorithm (Sections 4–5), the causal PAC-Bayesian value bound certifying any intervention scope $\mathcal{I} \subseteq [L]$ off-policy and anytime (Theorem 1, Section 6.1), and empirical evidence on the hierarchical SCM with reported p-values (Section 7). Each contribution is delivered in the corresponding numbered section and appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 8 flags three limitations: within-level exogeneity (Assumption 2), whose boundary (endogenous within-level state, as in titration) is worked out there; the restriction to additive-noise (ANM) mechanisms, with the non-additive case a stress test only; and the synthetic scope of exact validity verification. The multiplicative overlap penalty $B^{\mathcal{I}} \leq R_{\max} \prod_{\ell \in \mathcal{I}} B_w^{(\ell)}$ in Theorem 1 is discussed in Appendix J alongside concrete tightening directions (self-normalised importance sampling, PAC-Bayes-Bernstein refinements, self-normalised maximal inequalities).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Theorem 1 explicitly references Assumptions 1–4 (Causal Markov plus exogenous noise, per-level i.i.d. contexts, overlap and backdoor admissibility, and per-level bounded importance weights). The full proof is given in Appendix C in four steps (filtration setup, causal KL decomposition, logarithmic-smoothing supermartingale, Ville's inequality), followed by the proof of the excess-risk corollary, with a proof sketch in the main text and supporting Lemmas 2–4 stated and proved. All theorems, formulas, and assumptions are numbered and cross-referenced.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 7.1 fixes the agent taxonomy, statistical protocol (10 seeds, $K{=}2000$ episodes, Welch's t-test), and Γ-Stress preset; Appendix G provides the full Γ-SCM structural equations, axis semantics, posterior-backend hyperparameters (RFF-GP-Gibbs with $D{=}128$, RBF kernel $\exp(-\|x - x'\|^2)$; NIG-Gibbs as alternative), and per-bench deviations from defaults; Appendix F gives a worked linear instance with its legacy logging policy and prior. Algorithms 1 (NCTS) and 2 (AEGIS) are stated in pseudocode.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: A code repository is provided at `https://github.com/PaulDavidZuercher/AEGIS`, with configuration files and reproduction scripts for the reported numbers. Synthetic data are simulated from the parametric Γ-SCM specified in Appendix G; the two external-data checks (Appendix M) use the public HeartSteps V1 release [12] and the Open Bandit Dataset [28], both openly available under CC BY 4.0, with preprocessing scripts in the repository.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 7.1 states the experiment defaults; Appendices F and G list all SCM parameters, prior hyperparameters (NIG with $\mu_0{=}0$, $\Lambda_0{=}I$, $a_0{=}2$, $b_0{=}1$ in the experiments), posterior-update rules (NIG-conjugate for linear ANM, online Gibbs for nonlinear ANM),

legacy logging policy, RFF-GP-Gibbs hyperparameters (D=128, RBF kernel $\exp(-\|x - x'\|^2)$, λ chosen per bench), and per-bench data-split protocols (e.g., α=0.5 Bernoulli random split for the bound contraction in Section 7.2.3).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Section 7.1 fixes the default: mean $\pm 1\sigma$ across 10 seeds with significance via Welch's t-test; deviations (fewer seeds, standard errors) are stated inline. Headline claims report Cohen's d and p-value (Table 1), with Welch t-statistics in App. H.4 (e.g., $p < 10^{-7}$ on Shift-T, $p < 10^{-10}$ on the recursive-vs-joint contrast in Section 7.2; paired-endpoint test t=31.3, df=9, $p < 10^{-8}$, in the bound-contraction benchmark).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: Compute resources are not explicitly tabulated. Experiments run on a single machine with no GPU required (the two external-data checks run in seconds); the heaviest are the long-K runs of up to 1.5×10^6 certified episodes.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is theoretical and methodological; experiments use simulated environments plus secondary analysis of two public, de-identified datasets released for research (HeartSteps V1, Open Bandit Dataset); no new human-subjects data, no scraped data, no released pre-trained models. The deployment framing (progressive certified handover with practitioner-set risk thresholds) is explicitly oriented towards responsible deployment.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper targets safe deployment in high-stakes sequential domains (agronomic input allocation, process control; Section 1) where mis-deployed learned policies can cause direct harm. *Progressive certified handover* (Section 6.1) is the explicit positive-impact mechanism: control transfers from a legacy controller to the agent only when a $(1-\delta)$-certified gain is met. A false certificate can hand control to a harmful policy: the certificate does not detect wrong propensities (Section 6.2) or within-level endogenous state (Appendix J), so logging and Assumption 2 must be audited before deployment. We see no direct path to deliberate misuse, since the framework adds conservative gating to systems an operator already runs.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases neither pre-trained models, image generators, nor scraped datasets; the contributions are an algorithm (NCTS), a deployment-time wrapper (AEGIS), and a PAC-Bayesian bound (Theorem 1), none of which carry high misuse risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Prior algorithmic frameworks (PSRL, C-MDPs / C-UCBVI, structural causal bandits, online PAC-Bayes, logarithmic smoothing, pessimistic offline policy learning) are cited at first use in Section 2 and throughout the body. Two third-party datasets are used, both cited and both under CC BY 4.0: HeartSteps V1 [12] and the Open Bandit Dataset [28] (Appendix M); all other experiments run on the simulated Γ-SCM in Appendix G; no pre-trained models are loaded.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: A code repository at `https://github.com/PaulDavidZuercher/AEGIS` contains the implementation of NCTS (Algorithm 1), AEGIS (Algorithm 2), the Γ-SCM simulator, and reproduction scripts, with in-repo documentation.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing and collects no data from human subjects; besides simulated SCM data, it performs a secondary analysis of two public, de-identified releases, HeartSteps V1 [12] and the Open Bandit Dataset [28].

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper collects no data from human subjects; its human-derived data are two public, de-identified releases (HeartSteps V1, Open Bandit Dataset), analysed secondarily under their CC BY 4.0 licences; no new IRB review was required for this secondary use.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not part of the core methodology; the algorithms (NCTS in Algorithm 1, AEGIS in Algorithm 2) and the bound (Theorem 1) involve no LLM components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.