WHEN NATURE WHISPERS TO THE OPPONENT: INFORMATION LEAKAGE IN ROBUST MARKOV GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust Markov games contain two adverse entities: a strategic opponent and "nature," which governs uncertain transitions. Fusing them into a single adversary can change the information structure: if nature observes a hidden mode, the fused opponent may condition on information unavailable to the modeled opponent. We formalize the intended counterfactual by relaxing only the opponent's information while preserving the ambiguity family, transition dynamics, timing, and one complete nature plan. A four-value hierarchy separates this information effect from stronger relaxations that reoptimize continuation plans across public actions or private histories. For each fused profile, we construct a legal public-history *shadow* with the same nature plan; the KL divergence from the fused law to its shadow equals the cumulative conditional mutual information carried by opponent actions, yielding information-budget bounds and a zero-leakage exactness criterion. Finite-game, learning, pursuit–evasion, and cyber-defense experiments validate the distinction. In cyber defense, fusion raises a policy's worst-case safety cost from 1.244 to 3.000, reversing safe policy selection at threshold 1.4. For learned attackers, rollout estimates of leakage and public-shadow values match enumeration, while leakage regularization yields a value–leakage tradeoff.

1 INTRODUCTION

Robust MDPs model transition uncertainty through an ambiguity set and evaluate a policy against a worst-case transition model, often represented by an adversary called *nature* (Iyengar, 2005; Nilim & El Ghaoui, 2005; Wiesemann et al., 2013). Rectangular ambiguity supports Bellman recursion because nature's local transition choices can be combined, whereas coupled uncertainty can invalidate such local optimization (Epstein & Schneider, 2003; Goyal & Grand-Clément, 2023; Li & Shapiro, 2025). Robust Markov games add a strategic opponent to this uncertainty, combining adversarial interaction with an uncertain transition model (Shapley, 1953; Littman, 1994; Zhang et al., 2020; McMahan et al., 2024; Shi et al., 2025; Chang et al., 2026).

Because the opponent and nature have aligned objectives, a tempting shortcut is to fuse them into one adversarial policy. This is valid only if the fused entity has the information and feasible commitments intended for both components. Alignment alone does not imply a shared information set.

Running cyber example. Consider three services and a hidden mode identifying the vulnerable one. The defender uses public history, the attacker probes or exploits, and nature selects a stealth or loud transition row after the joint action. Nature sees the mode; the attacker sees only public alerts. A legal attacker therefore uses one action distribution across modes consistent with the same public history, whereas a fused attacker can target the vulnerable service immediately. This is the information error. Separately choosing incompatible nature continuations after hypothetical attacker actions is a plan error. The value hierarchy separates them, the public shadow removes only the former, and the cyber experiment tests their effect on safe policy selection.

At a given public history, a legal opponent must use the same action distribution across private situations it cannot distinguish. We call this *opponent-row sharing*. Fusion removes this restriction and lets actions depend on nature's private history; because actions are public, they may also reveal that information and alter future beliefs. This information relaxation must be distinguished from a separate modeling choice: allowing nature to replace its continuation plan after public actions

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

or across private histories. Such reselection can be invalid under persistent hidden modes, shared budgets, or cross-action-coupled transitions.

Robust-control theory studies when nature's transition choices can be pasted into a global model, while research on asymmetric information and adversarial teams examines how observations and communication shape behavior. We study opponent–nature fusion by changing only the opponent's access to nature-private information, retaining nature's physical authority, and bounding the value change. Without this separation, a conservative Bellman backup can evaluate a different game and even change robust-safety decisions.

We therefore ask: what is the correct counterfactual when the opponent gains access to nature's private history while the physical robust game remains fixed, and how large is the resulting value change? Our answer yields an information-correct robust-safe objective. A fused adversary remains useful as a stress test, but its value should be interpreted as performance against an informed adversary rather than automatically as the value of the original game. Our contributions are threefold.

- **First**, we define fusion as an information-only change, holding fixed the ambiguity set, timing, transitions, and one complete nature plan. A four-value hierarchy separates this change from suffix-plan reselection; every inequality can be strict.
- **Second**, for any fused strategy we construct a *public shadow*, a legal same-nature strategy that averages out private information. Its KL divergence from the fused law equals the cumulative conditional mutual information in opponent actions, yielding a value bound and zero-leakage exactness criterion.
- **Finally**, exhaustive enumeration validates the hierarchy and projection identities; matched interventions isolate private-mode access; two tasks show when revelation closes the gap or fusion reverses safe selection; and mode-labeled rollouts estimate a learned attacker's leakage and shadow value accurately against enumeration.

2 Related Work

Robust control studies which uncertain transition rows nature may combine and when dynamic programming is valid (Iyengar, 2005; Nilim & El Ghaoui, 2005; Epstein & Schneider, 2003; Wiesemann et al., 2013). Coupled uncertainty and partial observation make rectangularity depend on decision timing and available observations (Goyal & Grand-Clément, 2023; Li & Shapiro, 2025; Bovy et al., 2024; Pahlke, 2022), while robust multi-agent work studies equilibria, learning, and safety with an implicit nature player (Zhang et al., 2020; McMahan et al., 2024; Shi et al., 2025; Chang et al., 2026). We ask which of nature's private observations the strategic opponent may use. Rectangularity may justify plan pasting, but it does not justify giving the opponent nature's private observations.

Information relaxations enlarge a player's information set (Brown et al., 2010; Haugh & Wang, 2014); adversarial-team models distinguish communication from shared utility (Celli & Gatti, 2018; Carminati et al., 2022), and common-information methods use private-state prescriptions (Nayyar et al., 2013). Causal information projections minimize KL divergence using conditional marginals (Csiszár, 1975; Massey, 1990; Charalambous & Stavrou, 2016). Our comparison retains one complete nature plan while changing the opponent's information. Theorem 1 separates that change from plan reselection, and Theorem 2 proves that the projected opponent remains feasible under the retained plan, yielding a value bound.

Information-regularized policies control disclosure through actions (Strouse et al., 2018; Cundy et al., 2024). Here the history-conditional leakage bounds the value gap to a feasible same-nature shadow; Q4 evaluates this connection on learned policies. Omniscient-adversary models explicitly allow nature–opponent collusion and optimize regret (Aggarwal & How, 2026). We instead compare NC and fused expected losses under the same physical commitments. These losses supply the reward and safety inputs to a constrained optimizer (Altman, 1999; García & Fernández, 2015; Pinto et al., 2017). Appendix B gives the detailed comparisons.

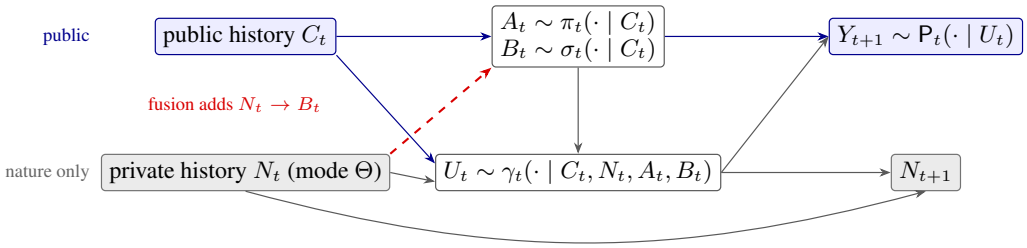


Figure 1: **Stage-t timing and information.** Blue edges are public and grey edges nature-private. The defender and noncommunicating opponent use C_t; nature selects U_t. Fusion adds the dashed edge $N_t \to B_t$; the public shadow removes it while keeping γ fixed.

3 Preliminaries

From robust MDPs to robust Markov games We use a finite-horizon, finite-space robust Markov game with horizon H, state space $\mathcal{S}$, defender actions $\mathcal{A}$, opponent actions $\mathcal{B}$, and stage rewards r_t. The defender uses policy π, the strategic opponent uses policy σ, and nature uses protocol γ to select admissible transition rows U_t from an ambiguity model. With a fixed transition model, it is a zero-sum Markov game (Shapley, 1953; Littman, 1994). All strategies are behavioral with perfect recall. The defender and opponent use public history, whereas nature may additionally observe a latent mode and its own past choices. Nature's ambiguity model may be rectangular or may couple rows through persistent parameters or shared budgets. Although σ and γ maximize the same loss in our convention, they remain distinct decision makers with distinct information and commitments.

Timing and information Consider a finite-horizon game with stages $t = 1, \dots, H$. Let C_t be the public history before the stage-t actions and N_t be nature's private history. At stage t, let $A_t \in \mathcal{A}$ denote the defender's action and $B_t \in \mathcal{B}$ the opponent's action; the two actions are chosen simultaneously. Their joint action becomes public. Nature then chooses an admissible transition row U_t, after which exogenous randomness generates the next state and public observation according to a fixed primitive chance kernel P_t conditional on U_t. Under NC, the opponent does not see N_t; under either information structure, it does not see U_t before acting. Nature sees (C_t, N_t, A_t, B_t) when choosing U_t. The defender uses the fixed public-history row $\pi_t(a \mid C_t)$ throughout each comparison. Figure 1 summarizes the timing and the single edge that fusion adds. Let $\sigma_t(b \mid C_t, N_t)$ be the opponent's stage-t action kernel, where $b \in \mathcal{B}$. A noncommunicating (NC) opponent satisfies

$$\sigma_t(b \mid C_t, N_t) = \sigma_t(b \mid C_t), \tag{1}$$

whereas a fused opponent may use the full displayed kernel. Nature selects row u with kernel $\gamma_t(u \mid C_t, N_t, A_t, B_t)$. This timing and all physical transition rows are identical in the two games. For a finite set $\mathcal{X}$, let $\Delta(\mathcal{X})$ be its probability simplex; for a law μ, let $\mathrm{supp}(\mu) = \{x : \mu(x) > 0\}$. A latent mode Θ in finite set $\mathcal{M}$ is drawn, after the defender commits, from $q \in \Delta(\mathcal{M})$. Nature observes Θ, whereas the opponent does not. The ambiguity law and complete protocols are chosen before Θ is realized. Robust evaluation ranges over a full-support family $\mathfrak{Q}_{\rho,\varepsilon}$ with radius $\rho \geq 0$ and mass floor $\varepsilon \in (0, 1]$; Appendix H gives its experimental instantiation.

Induced trajectory laws and robust values Fix defender policy π and ambiguity law q. Let Σ_o^{nc} and Σ_o^{fuse} be the NC and fused opponent strategy classes, and $\Gamma_n(q)$ be the admissible nature protocols. For $\square \in \{\mathrm{nc}, \mathrm{fuse}\}$, define the induced trajectory-law set

$$\mathcal{L}_\square(\pi; q) = \{\mathrm{Law}_q(\pi, \sigma, \gamma) : \sigma \in \Sigma_o^\square,\ \gamma \in \Gamma_n(q)\}, \tag{2}$$

where $\mathrm{Law}_q(\pi, \sigma, \gamma)$ is the law generated by q, the three protocols, and the primitive chance kernels. For fixed γ, the notation $\mathcal{L}_\square(\pi; q, \gamma)$ denotes the corresponding *same-nature* subfamily; behavioral strategies use independent private randomization. At a reached public history $C_t = c$, an induced law P determines the public belief $\beta_t^P(n \mid c) := P(N_t = n \mid C_t = c)$. This is an analysis object, not an additional observation. For a bounded trajectory loss g, define

$$W_g^{\mathrm{nc}}(\pi; q) = \sup_{P \in \mathcal{L}_{\mathrm{nc}}(\pi; q)} \mathbb{E}_P[g], \qquad W_g^{\mathrm{fuse}}(\pi; q) = \sup_{P \in \mathcal{L}_{\mathrm{fuse}}(\pi; q)} \mathbb{E}_P[g], \tag{3}$$

$$W_g^\square(\pi) = \sup_{q \in \mathfrak{Q}_{\rho,\varepsilon}} W_g^\square(\pi; q), \qquad \square \in \{\mathrm{nc}, \mathrm{fuse}\}. \tag{4}$$

Because $\Sigma_o^{\mathrm{nc}} \subseteq \Sigma_o^{\mathrm{fuse}}$, we have $W_g^{\mathrm{nc}}(\pi) \le W_g^{\mathrm{fuse}}(\pi)$. For trajectory laws P, Q, we use $\mathrm{TV}(P,Q) = \frac{1}{2}\|P - Q\|_1$ for total variation distance and $R_g = \sup g - \inf g$ for the span of g over the finite trajectory space. Let Π be the defender policy class, G_r cumulative reward, and $G_1, \ldots, G_{J_{\mathrm{safe}}}$ cumulative safety costs with thresholds $d_1, \ldots, d_{J_{\mathrm{safe}}}$. Before stating the optimization, define $R^{\square}(\pi) := -W^{\square}_{-G_r}(\pi)$ and $C_k^{\square}(\pi) := W^{\square}_{G_k}(\pi)$. The intended robust-safe problem is

$$\max_{\pi\in\Pi} R^{\mathrm{nc}}(\pi) \quad \text{subject to} \quad C_k^{\mathrm{nc}}(\pi) \le d_k, \qquad k = 1, \ldots, J_{\mathrm{safe}}, \tag{5}$$

Replacing "nc" by "fuse" defines the fused stress test. Reward and each safety cost are evaluated separately because fusion need not affect them equally. For the single-cost finite-library experiments, write $C^{\square} := C_1^{\square}$; each information structure selects the highest-reward policy whose corresponding cost is at most d. A *safe-selection reversal* occurs when the choices differ and the NC-selected policy is unsafe under fused evaluation; reward forfeiture is measured on the NC scale. Appendix H gives the exact selector notation and tie rule.

4 METHODOLOGY

We first give the opponent access to nature's private history, then separate this change from continuation-plan reselection. Finally, we construct a public-history opponent that removes the additional information while retaining nature's complete plan.

4.1 FUSION AS AN INFORMATION RELAXATION

Consider an opponent information set I, and let $\mathcal{R}(I)$ denote the complete histories that the opponent cannot distinguish at I but that may differ in nature's private history. Under the noncommunicating information structure, let $x_I(\cdot \mid r) \in \Delta(\mathcal{B})$ be the opponent's mixed action row at refinement $r \in \mathcal{R}(I)$. The opponent must use the same row at all such histories:

$$x_I(\cdot \mid r) = x_I(\cdot \mid r') \qquad \text{for all } r, r' \in \mathcal{R}(I). \tag{6}$$

Fusion removes only these equality constraints, allowing the opponent to choose different mixed actions at histories that differ in nature's private information. We denote this enlargement of the opponent strategy class by $\mathrm{Lift}^{B}_{n/o}$, where B refers to the opponent action and n/o indicates that nature's private information is made available to the opponent. Write $\widetilde{\sigma}$ for a generic opponent strategy in this lifted class. Importantly, this change does not alter the ambiguity law, primitive chance kernels, timing, feasible nature protocols, or physical transition rows. On the complete history tree,

$$\Sigma_o^{\mathrm{fuse}} = \mathrm{Lift}^{B}_{n/o}(\Sigma_o^{\mathrm{nc}}), \tag{7}$$

$$\mathcal{L}_{\mathrm{fuse}}(\pi; q) = \bigcup\nolimits_{\gamma\in\Gamma_n(q)} \{\mathrm{Law}_q(\pi, \widetilde{\sigma}, \gamma) : \widetilde{\sigma} \in \mathrm{Lift}^{B}_{n/o}(\Sigma_o^{\mathrm{nc}})\}. \tag{8}$$

Appendix C proves these identities. Our default state for recursive analysis is the complete history. A smaller public/private state representation may be used only when histories represented by the same state have identical one-step losses, observations, belief updates, and remaining nature commitments. Appendix F states the condition and proves that such a reduced representation preserves the game.

4.2 PUBLIC ACTIONS AND FOUR CONTINUATION VALUES

Fix a reached public history $C_t = c$. Let z denote a possible value of nature's private history N_t, let $\mathcal{Z}$ be the finite set of such values, and let $\beta \in \Delta(\mathcal{Z})$ be the public conditional belief $\beta(z) = P(N_t = z \mid C_t = c)$. Because the opponent's action B_t is publicly observed, a fused strategy $\sigma_z(b) = P(B_t = b \mid C_t = c, N_t = z)$ both chooses an action and conveys information about z. Its induced action probabilities and posterior beliefs are

$$\lambda_b = \sum\nolimits_{z\in\mathcal{Z}} \beta(z)\sigma_z(b), \qquad \beta_b(z) = \frac{\beta(z)\sigma_z(b)}{\lambda_b}. \tag{9}$$

The posterior β_b is needed only when $\lambda_b > 0$; on a zero-mass branch it may be arbitrary.

Conversely, every action-indexed Bayes-plausible decomposition of β can be implemented by a fused action row on $\mathrm{supp}(\beta)$. Thus fusion is equivalent to permitting a Bayes-plausible, action-indexed

split of the public belief; Appendix D gives the split set and inverse map. The key question is what continuation strategies remain fixed after that action is chosen.

Let $\mathfrak{P}_t$ denote the compact set of admissible complete continuation plans from stage t onward. A plan $p = (\sigma_{t+1:H}, \gamma_{t:H})$ specifies future opponent behavior, nature's future choices, off-path behavior, and any persistent commitments. For $p \in \mathfrak{P}_t$, let $Q_p(z, b)$ be the expected remaining cumulative loss from stages t through H when the private history is z, the current opponent action is fixed to b, and continuation plan p is then followed. We write $\mathfrak{P}_t^{\mathcal{B}}$ for the Cartesian product containing one plan $p_b \in \mathfrak{P}_t$ for each $b \in \mathcal{B}$. The following four values separate the relevant modeling choices:

$$V_{00} = \sup_{p \in \mathfrak{P}_t} \max_{b \in \mathcal{B}} \sum\nolimits_{z \in \mathcal{Z}} \beta(z) Q_p(z, b), \qquad V_{10} = \sup_{p \in \mathfrak{P}_t} \sum\nolimits_{z \in \mathcal{Z}} \beta(z) \max_{b \in \mathcal{B}} Q_p(z, b), \tag{10}$$

$$V_{11} = \sup_{(p_b)_{b \in \mathcal{B}} \in \mathfrak{P}_t^{\mathcal{B}}} \sum\nolimits_{z \in \mathcal{Z}} \beta(z) \max_{b \in \mathcal{B}} Q_{p_b}(z, b), \quad V_{\infty} = \sum\nolimits_{z \in \mathcal{Z}} \beta(z) \sup_{(p,b) \in \mathfrak{P}_t \times \mathcal{B}} Q_p(z, b). \tag{11}$$

Because expected loss is linear in the current mixed action row, optimization reduces to pure actions $b \in \mathcal{B}$. The four expressions use the same continuation-plan class but differ in which choices may depend on private history or public action. This is a local comparison with the future opponent class fixed; global NC and fused values use the classes in Equation (20). In particular, $Q_p(z, b)$ evaluates b under a fixed plan p without re-optimizing nature after observing b.

Value	Interpretation at the current decision point
V_{00}	The opponent uses one public action and one complete continuation plan.
V_{10}	The action may depend on private history z, but the continuation plan remains fixed.
V_{11}	A different complete continuation plan may be chosen for each public action.
V_{∞}	The continuation plan may also depend on private history z.

The step from V_{00} to V_{10} is the information relaxation studied here: the opponent gains access to nature's private history while the continuation plan remains fixed. The steps from V_{10} to V_{11} and from V_{11} to V_{∞} add freedom to reselect continuation plans after the public action or across private histories. When future opponent strategies are fully fused, these relaxations concern the sharing constraints imposed by nature's commitments. Such reselection is valid only when the separately chosen continuation plans can be combined into one admissible global plan. Under standard public-state rectangularity, this condition holds and $V_{10} = V_{11} = V_{\infty}$. Under coupled uncertainty, however, nature may be unable to combine the individually optimal branches, so these values can differ.

Precise causal-pasting conditions—including persistent budgets, cross-action coupling, and zero-probability histories—appear in Appendix D.3.

4.3 Public shadows

We next construct a legal noncommunicating strategy corresponding to any fused strategy while keeping nature fixed. Consider a fused profile (σ, γ) with induced law $P = \mathrm{Law}_q(\pi, \sigma, \gamma)$. At each public history, average the fused opponent's action distribution over nature's private history; in the following display, n ranges over the finite values of N_t:

$$\bar{\sigma}_t(b \mid c) = P(B_t = b \mid C_t = c) = \sum\nolimits_n P(N_t = n \mid C_t = c) \sigma_t(b \mid c, n). \tag{12}$$

We call $\bar{\sigma}$ the *public shadow* of σ. It depends only on public history and is therefore a valid noncommunicating strategy. We retain the same nature protocol γ, so $\bar{P} = \mathrm{Law}_q(\pi, \bar{\sigma}, \gamma)$ is admissible even with persistent or cross-action-coupled nature commitments, provided nature's admissibility does not depend on the opponent's action probabilities.

The public shadow copies the conditional action row $P(B_t \mid C_t = c)$ at each reached history, removing its dependence on nature-private information. Its public-history probabilities follow the resulting shadow law $\bar{P}$. This makes it possible to measure exactly how much information fusion introduces. Using natural logarithms, define

$$\mathcal{I}(P) = \sum\nolimits_{t=1}^{H} I_P(N_t; B_t \mid C_t), \tag{13}$$

where $I_P(N_t; B_t \mid C_t)$ is conditional mutual information under P and $D_{\mathrm{KL}}(P \| Q)$ denotes the Kullback–Leibler divergence between probability laws P and Q.

The quantity $\mathcal{I}(P)$ does not introduce a separate communication mechanism; it measures the dependence created by allowing the public action B_t to depend on nature's private history N_t. If $\mathcal{I}(P) = 0$, the fused opponent uses no private information beyond public history. Larger values indicate greater use of information unavailable to the noncommunicating opponent.

5 Theoretical Analysis

We prove that the three relaxations give an ordered hierarchy and can all change the value. We then bound the effect of private information using a feasible public shadow with the same nature plan.

5.1 What fusion changes

Theorem 1 (Plan-preserving fusion hierarchy)**.** *Fix a public node, a private-history distribution $\beta \in \Delta(\mathcal{Z})$ on a finite set $\mathcal{Z}$, and a finite action set $\mathcal{B}$. Suppose $\mathfrak{P}_t$ is compact and, for every $(z, b) \in \mathcal{Z} \times \mathcal{B}$, the map $p \mapsto Q_p(z, b)$ is upper semicontinuous. Then*

$$V_{00} \le V_{10} \le V_{11} \le V_\infty, \tag{14}$$

and all three inequalities can be strict in one finite game. Equality holds if and only if some optimizer of the larger problem satisfies the smaller problem's sharing constraint. Closure under combining plans across public actions gives $V_{10} = V_{11}$; closure across private histories within each action gives $V_{11} = V_\infty$.

The step from V_{00} to V_{10} is the information relaxation of interest: the opponent gains nature's private history while the continuation plan remains fixed. Later steps allow different plans after the public action or across private histories. All three inequalities can be strict: a solver may use valid transition rows on every branch yet evaluate a different game by combining plans that are not jointly feasible. Appendix D.2 gives equality conditions and a strict example; Appendix D.3 proves the affine result.

Fusion is locally exact iff some informed-optimal continuation plan admits one action that is optimal for every private history in $\mathrm{supp}(\beta)$. The exact optimizer-set characterization is in Appendix D.2. Proposition 3 gives $(V_{00}, V_{10}, V_{11}, V_\infty) = (3/8, 1/2, 3/4, 3/4)$. The first gap reflects private information; the second comes from branchwise plans that violate nature's shared budget. Keeping the budget in the state restores $V_{10} = 1/2$.

5.2 How much fusion changes

Theorem 2 (Same-nature public-shadow certificate)**.** *Fix (π, q) and a bounded trajectory criterion g with span R_g. Suppose the opponent has the same feasible action set at every refinement of an information set and nature's admissible plans do not depend on the numerical probabilities in the opponent strategy. For any fused profile (σ, γ) with $P = \mathrm{Law}_q(\pi, \sigma, \gamma)$ and any public-history opponent strategy τ, let $P^\tau = \mathrm{Law}_q(\pi, \tau, \gamma)$ retain the same q, π, and γ. Then*

$$D_{\mathrm{KL}}(P \| P^\tau) = \mathcal{I}(P) + \sum\nolimits_{t=1}^{H} \mathbb{E}_{P(C_t)} D_{\mathrm{KL}}(\bar{\sigma}_t(\cdot \mid C_t) \| \tau_t(\cdot \mid C_t)) \,. \tag{15}$$

Here $\mathbb{E}_{P(C_t)}$ denotes expectation with respect to the C_t marginal of P. This identity holds in the extended-real sense. Consequently, the public shadow $\bar{P}$ is KL-nearest among $\mathcal{L}_{\mathrm{nc}}(\pi; q, \gamma)$ and $D_{\mathrm{KL}}(P \| \bar{P}) = \mathcal{I}(P)$. In finite games with attained optima, $W_g^{\mathrm{fuse}}(\pi; q) = W_g^{\mathrm{nc}}(\pi; q)$ if and only if there exists an optimal fused profile with $\mathcal{I}(P) = 0$.

The identity follows because the two trajectory laws share every factor except the opponent's action kernels. Appendix E gives the likelihood-ratio proof and its feasibility assumptions. The shadow is a feasible NC opponent under the retained nature plan. Comparing its value with the NC optimum gives the profile bound

$$\mathbb{E}_P[g] - W_g^{\mathrm{nc}}(\pi; q) \le \mathbb{E}_P[g] - \mathbb{E}_{\bar{P}}[g] \le R_g \min\left\{1, \sqrt{\mathcal{I}(P)/2}\right\}. \tag{16}$$

For a nonoptimal fused profile, the first quantity can be negative. For an optimal fused profile, it is exactly the fusion gap. Thus the bound controls the value effect of fusion without requiring the public shadow itself to solve the noncommunicating game.

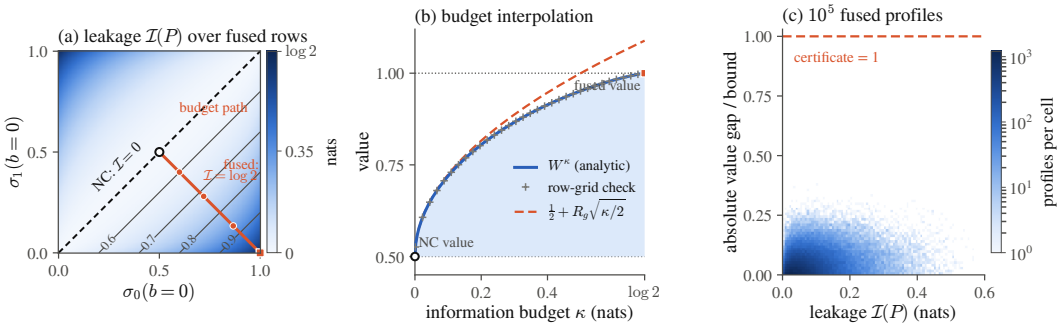


Figure 2: **Leakage certificates.** (a) One-stage matching game: the diagonal is NC, contours show value, and the orange path maximizes value at each leakage budget. (b) Analytic budget curve and a 16.77-million-row grid check against Equation (18). (c) Absolute value gap to the same-nature shadow for 10^5 random two-stage profiles, with $R_g = 2$; colour shows counts on a logarithmic scale. Every profile satisfies Equation (16).

Finally, we can restrict the fused opponent by an explicit information budget. For $\kappa \geq 0$, define

$$W_g^\kappa(\pi; q) = \sup\{\mathbb{E}_P[g] : P \in \mathcal{L}_{\text{fuse}}(\pi; q),\ \mathcal{I}(P) \leq \kappa\}. \tag{17}$$

Then

$$0 \leq W_g^\kappa(\pi; q) - W_g^{\text{nc}}(\pi; q) \leq R_g \min\left\{1, \sqrt{\kappa/2}\right\}. \tag{18}$$

At $\kappa = 0$, the opponent uses no nature-private information beyond public history, recovering the noncommunicating value; a sufficiently large budget recovers fusion. The proof factors the likelihood ratio over the trajectory and does not require nature to reselect its continuation plan after posterior updates. Appendix E gives the projection proof and information-budget result. Figure 2 makes the three certificates concrete: the leakage landscape of a one-stage matching game, its exact budget curve against Equation (18), and the profile bound of Equation (16) over random two-stage profiles.

6 Experiments

We test four questions: whether the theoretical distinctions appear in finite examples (Q1), whether fusion changes robust-safe policy selection (Q2), whether private-mode access causes the observed gap (Q3), and whether leakage and public-shadow quantities can be estimated from rollouts (Q4). Appendix H gives the full environments, protocols, sweeps, and additional controls.

Q1: Do the theoretical distinctions appear in finite games? The five-type construction gives $(V_{00}, V_{10}, V_{11}, V_\infty) = (0.4, 0.6, 0.8, 1.0)$, so every relaxation in Theorem 1 can be strict. Across one million random three-plan, five-type tables, we observe no hierarchy violations, and all three inequalities are strict in 35.0% of cases. The public-shadow identity holds on 10^5 random fused profiles with maximum residual 1.3×10^{-15}. Figure 3 shows that the fusion gap also persists across larger hidden-mode games, while Figure 2 confirms the information-budget bound. Additional scaling, state-compression, and weak-signal checks appear in Appendices H and G.

Q2: Can fusion reverse robust-safe policy selection? We consider a horizon-4 cyber-defense game with three persistent vulnerability modes. Safety is worst-case expected breach count with threshold $d = 1.4$, and reward trades availability against a breach penalty. We evaluate nine declared public-history defender policies by exact robust recursion. Table 1 shows a selection reversal. Under NC, serve-once-then-rotate is feasible and has the highest reward in the library, with safety cost 1.244. Under fusion its cost rises to 3.000, so patch-all is selected and 0.535 units of information-correct reward are lost. Publicly revealing the mode or making breaches mode-independent closes the NC–fused gap to numerical precision. Figure 4 shows that the effect also persists across $1{,}248$ public policies, threshold sweeps, and perturbed channel/reward settings. Further observation-model and Bellman-backup checks appear in Appendices H.6 and H.7.

Q3: Is private regime access the operative mechanism? We fix serve-once-then-rotate, a uniform mode law, and a channel with 0.30 loud-row probability. The same recurrent attacker receives three

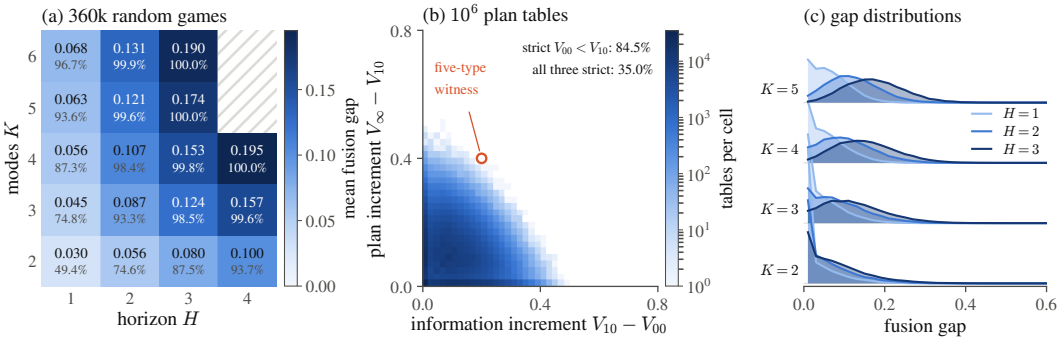


Figure 3: **Exhaustive validation.** (a) Mean fusion gap over 20,000 random games per cell; percentages give positive-gap frequency, and hatching marks unrun cells. (b) Information and plan-reselection increments over 10^6 tables; colour gives counts on a logarithmic scale. (c) Gap distributions for $K \leq 5$ modes and $H \leq 3$ stages.

Table 1: Robust reward and safety values for the nine-policy cyber-defense library. At $d = 1.4$, NC selects serve-once-then-rotate, whereas fusion selects patch-all.

Defender policy	R^{nc}	R^{fuse}	C^{nc}	C^{fuse}	Selected
Always serve	1.480	−2.000	1.680	4.000	–
Patch service 0	−0.080	−2.080	1.653	2.987	–
Patch all	0.400	0.400	0.000	0.000	fused
Rotate patches	0.540	−1.720	1.240	2.747	–
Serve once, then rotate	0.935	−1.700	1.244	3.000	NC
Serve twice, then rotate	0.720	−1.920	1.653	3.413	–
Alert, then patch service 0	0.428	−2.249	1.653	3.581	–
Alert, then rotate	0.709	−1.961	1.509	3.479	–
Alert-count patch	0.471	−2.247	1.638	3.646	–

Table 2: Matched learning intervention over sixteen paired runs. Finite DP is the information-compatible horizon-four upper reference.

Attacker information	Learned cost	Finite DP	Difference from NC
NC (masked regime)	0.9989	1.0027	–
Matched-marginal nuisance	0.9988	1.0027	−0.0001
Fused (true regime)	2.9999	3.0000	2.0010

inputs: no mode information (NC), the true mode (fused), or an independent matched-marginal mode (nuisance). All other training and evaluation conditions are identical. The fused-minus-NC difference is 2.0010 with a paired 95% bootstrap interval $[1.9999, 2.0022]$, whereas the nuisance input has negligible effect. Thus the learned attacker recovers almost the entire advantage predicted from private-mode access rather than from additional input dimension or training variation. Figure 5 gives the matched intervention and horizon sweep; Appendix H.8 gives training details and paired intervals. Appendix H.9 provides the pursuit–evasion revelation control.

Q4: Can rollouts estimate leakage and shadow values? Across four trained attackers and three draws each, 2^{18} mode-labeled episodes give maximum leakage error below 0.0003 nats against enumeration. The empirical shadow's exact value differs from the true shadow value by less than 0.000009; its fresh-rollout evaluation has error below 0.0053. Leakage-penalized training yields a value–leakage tradeoff, and all 56 trained attackers satisfy Equation (16). Appendix H.10 specifies both estimators and all run ranges.



Figure 4: **Fusion changes safe selection.** (a) NC and fused costs for 1,248 policies; marker area gives multiplicity, colour gives best NC reward, and rings mark the nine-policy library. (b) Its selections as the threshold varies. (c) Reversal windows across 192 horizon-three and 32 horizon-four configurations; crosses mark no window.

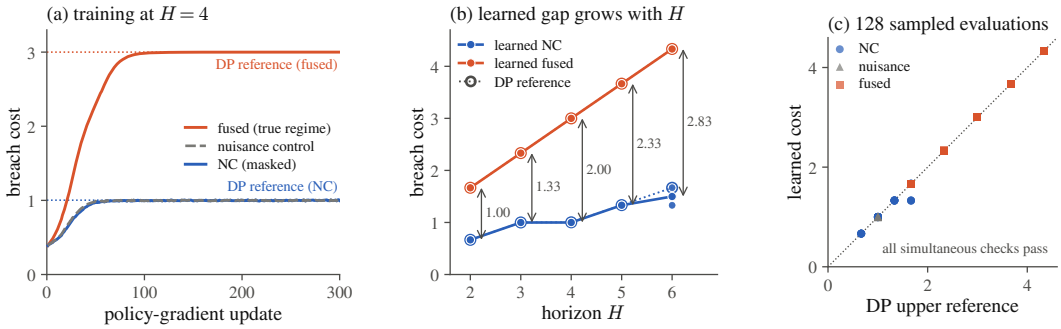


Figure 5: **Matched information intervention.** (a) Mean training curves at $H = 4$. (b) NC and fused costs across horizons with finite-DP references. (c) Simultaneous bounded-return checks.

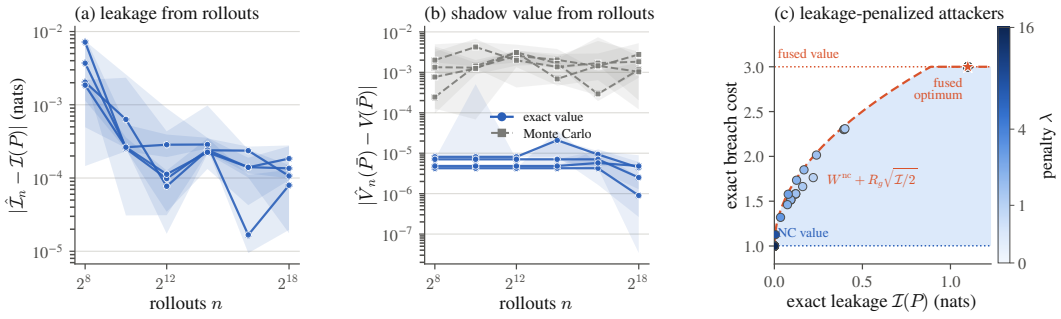


Figure 6: **Estimating certificate quantities.** (a,b) Median and range over three draws per attacker. Panel (b) compares exact empirical-shadow values (solid) and fresh-rollout values (dashed) with the true shadow value. (c) All trained policies; colour gives penalty weight, and the dashed span-3 bound is capped at the fused value (star).

7 Conclusion

The plan hierarchy separates opponent information from nature-plan reselection. The public shadow bounds the resulting value change under the retained plan. Exact and learned evaluations demonstrate safe-selection reversals, including across 1,248 policies, and connect learned attackers' information use to their cost.

STATEMENT OF USE OF AI

The research questions, the main ideas, the theoretical claims, the methodology, and the experimental protocols of this paper were chosen by the authors, who divided the work among themselves, supervised every stage of the project, and made all decisions on claims and conclusions. Under the authors' direction, we used generative AI tools as assistants for the following tasks with required disclosure. For the theoretical component, generative AI was used in part as a "theory attacker": it assisted in implementing computational searches and large-scale simulations over large collections of generated examples in order to identify potential counterexamples, boundary cases, or weaknesses in candidate theorem statements and proof plans. These searches were used to help the authors stress-test and refine the theoretical claims. The absence of a counterexample in such searches does not constitute a proof, and no theorem in this paper is considered established on the basis of AI-assisted simulation, search, or testing. The authors reviewed and independently assessed all theorem statements, proof plans, and formal proofs.

Generative AI tools were also used to discuss the experiment design and assist with implementing the methods and experiments. The authors reviewed all AI-assisted work. No experimental data or results were produced by generative AI tools themselves: every reported number comes from experiments that were actually executed, and translation and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools for literature-search triage, drafting and editing the text and LaTeX for readability, and checking notation and the consistency of reported numbers. The authors take full responsibility for every claim, proof, citation, and experimental result in this paper; no AI system is an author.

REPRODUCIBILITY STATEMENT

Appendix H specifies the games, policy classes, solvers, and evaluation protocols of all experiments. The supplementary material provides the source code for all experiments, together with the recorded outputs behind all reported results.

REFERENCES

Naman Aggarwal and Jonathan P. How. Beyond bayesian Nash: Learning minimax-regret equilibria for adversarial team games under asymmetric information. *arXiv preprint arXiv:2607.09993*, 2026. URL https://arxiv.org/abs/2607.09993. Submitted to Transactions on Machine Learning Research.

Eitan Altman. *Constrained Markov Decision Processes*. Chapman and Hall/CRC, 1999.

Eline M. Bovy, Marnix Suilen, Sebastian Junges, and Nils Jansen. Imprecise probabilities meet partial observability: Game semantics for robust POMDPs. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 6697–6706. International Joint Conferences on Artificial Intelligence Organization, 2024. doi: 10.24963/ijcai.2024/740.

David B. Brown, James E. Smith, and Peng Sun. Information relaxations and duality in stochastic dynamic programs. *Operations Research*, 58(4-part-1):785–801, 2010. doi: 10.1287/opre.1090.0796.

Luca Carminati, Federico Cacciamani, Marco Ciccone, and Nicola Gatti. A marriage between adversarial team games and 2-player games: Enabling abstractions, no-regret learning, and subgame solving. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2638–2657. PMLR, 2022. URL https://proceedings.mlr.press/v162/carminati22a.html.

Andrea Celli and Nicola Gatti. Computational results for extensive-form adversarial team games. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 965–972, 2018. doi: 10.1609/aaai.v32i1.11462.

Ningkang Chang, Chenyu Xu, Ziying Jia, Yue Wang, and Sihong He. Robust constrained Markov games: Multi-agent decision-making under model uncertainty and constraints. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 1003–1033. PMLR, 2026. URL https://proceedings.mlr.press/v337/chang26a.html.

Charalambos D. Charalambous and Photios A. Stavrou. Directed information on abstract spaces: Properties and variational equalities. *IEEE Transactions on Information Theory*, 62(11):6019–6052, November 2016. doi: 10.1109/TIT.2016.2604846. URL https://doi.org/10.1109/TIT.2016.2604846.

Imre Csiszár. i-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975. doi: 10.1214/aop/1176996454.

Chris J. Cundy, Rishi Desai, and Stefano Ermon. Privacy-constrained policies via mutual information regularized policy gradients. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2809–2817. PMLR, 2024. URL https://proceedings.mlr.press/v238/cundy24a.html.

Larry G. Epstein and Martin Schneider. Recursive multiple-priors. *Journal of Economic Theory*, 113 (1):1–31, 2003. doi: 10.1016/S0022-0531(03)00097-8.

Javier García and Fernando Fernández. A Comprehensive Survey on Safe Reinforcement Learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015. URL https://jmlr.org/papers/v16/garcia15a.html.

Vineet Goyal and Julien Grand-Clément. Robust Markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 48(1):203–226, 2023. doi: 10.1287/moor.2022.1259.

Martin Haugh and Chun Wang. Information relaxations and dynamic zero-sum games. *arXiv preprint arXiv:1405.4347*, 2014. URL https://arxiv.org/abs/1405.4347.

Angel Y. He and David Parker. Robust verification of concurrent stochastic games. In *Tools and Algorithms for the Construction and Analysis of Systems*, volume 16505 of *Lecture Notes in Computer Science*, pp. 505–525. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-22752-2_26.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005. doi: 10.1287/moor.1040.0129.

Dhruva Kartik, Ashutosh Nayyar, and Urbashi Mitra. Common information belief based dynamic programs for stochastic zero-sum games with competing teams. *arXiv preprint arXiv:2102.05838*, 2021. URL https://arxiv.org/abs/2102.05838.

Merlijn Krale, Eline M. Bovy, Maris F. L. Galesloot, Thiago Simão, and Nils Jansen. On evaluating policies for robust POMDPs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1919.

Yan Li and Alexander Shapiro. Rectangularity and duality of distributionally robust Markov decision processes. *Mathematical Programming*, 2025. doi: 10.1007/s10107-025-02297-y.

Michael L. Littman. Markov Games as a Framework for Multi-Agent Reinforcement Learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 157–163. Morgan Kaufmann, 1994. doi: 10.1016/B978-1-55860-335-6.50027-1.

James L. Massey. Causality, feedback and directed information. In *Proceedings of the International Symposium on Information Theory and Its Applications*, pp. 303–305, 1990.

Jeremy McMahan, Giovanni Artiglio, and Qiaomin Xie. Roping in uncertainty: Robustness and regularization in markov games. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35267–35295, 2024.

Ashutosh Nayyar, Aditya Mahajan, and Demosthenis Teneketzis. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, 58(7):1644–1658, 2013. doi: 10.1109/TAC.2013.2239000.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. doi: 10.1287/opre.1050.0216.

Marieke Pahlke. Dynamic consistency in incomplete information games with multiple priors. *Games and Economic Behavior*, 133:85–108, 2022. doi: 10.1016/j.geb.2022.02.004.

Haim H. Permuter, Young-Han Kim, and Tsachy Weissman. Interpretations of directed information in portfolio theory, data compression, and hypothesis testing. *IEEE Transactions on Information Theory*, 57(6):3248–3259, 2011. doi: 10.1109/TIT.2011.2136270.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust Adversarial Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2817–2826. PMLR, 2017. URL https://proceedings.mlr.press/v70/pinto17a.html.

Jérôme Renault. The value of Markov chain games with lack of information on one side. *Mathematics of Operations Research*, 31(3):490–512, 2006. doi: 10.1287/moor.1060.0199.

Lloyd S. Shapley. Stochastic Games. *Proceedings of the National Academy of Sciences*, 39(10): 1095–1100, 1953. doi: 10.1073/pnas.39.10.1095.

Laixi Shi, Jingchu Gai, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Breaking the curse of multiagency in robust multi-agent reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54904–54918. PMLR, 2025. URL https://proceedings.mlr.press/v267/shi25c.html.

DJ Strouse, Max Kleiman-Weiner, Josh Tenenbaum, Matthew Botvinick, and David J. Schwab. Learning to share and hide intentions using information regularization. In *Advances in Neural Information Processing Systems*, volume 31, pp. 10270–10281, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/1ef03ed0cd5863c550128836b28ec3e9-Abstract.html.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov Decision Processes. *Mathematics of Operations Research*, 38(1):153–183, 2013. doi: 10.1287/moor.1120.0566.

Kaiqing Zhang, Tao Sun, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Basar. Robust multi-agent reinforcement learning with model uncertainty. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10571–10583, 2020.

Wei Zheng, Taeho Jung, and Hai Lin. The Stackelberg equilibrium for one-sided zero-sum partially observable stochastic games. *Automatica*, 140:110231, 2022. doi: 10.1016/j.automatica.2022.110231.

A Scope and Proof Map

This appendix proves the two main results and records the assumptions needed to pass from complete histories to a smaller causal state. Readers interested primarily in the paper's claims can go directly to Appendix D.2 for the plan-preserving hierarchy and Appendix E for the public-shadow and leakage results. Appendices C and F justify the underlying history and reduced-state representations, while Appendix G is an optional example. Neither main proof relies on an unverified state compression.

The causal scope is fixed throughout:

- all primitive spaces and the horizon are finite;
- behavioral strategies have perfect recall;
- the defender and opponent act simultaneously, their joint action is then public, and nature chooses a transition row afterward;
- the defender uses a fixed public-history behavioral kernel; private defender information is outside the stated recursion unless it is included explicitly in the exact and quotient states;
- the exact state is the public and nature-private history; a smaller state is used only under the row-compatible quotient conditions in Equation (57);
- the affine-collapse regime of the plan hierarchy assumes that complete suffix plans are recursively pasteable after augmenting the state by every persistent parameter or resource constraint;
- the causal-shadow theorem instead keeps one complete nature plan fixed and therefore permits cross-action coupling and persistent commitments;
- there is no private opponent–nature correlation seed; and
- the ambiguity law is fixed before each raw causal strategy fiber is formed, with the robust supremum taken afterward.

A public randomization variable can be included in C_t. We exclude a private opponent–nature correlation seed because the public-shadow identity requires separate behavioral randomization. Such a seed does not change the scalar NC or fused support value when it is drawn ex ante, independently of the mode and primitive chance, and every seed-conditioned profile is admissible: conditioning on the seed expresses its value as a mixture of admissible profile values. If nature-plan admissibility depends numerically on the opponent row, the same-nature shadow need not be feasible; that case lies outside Theorem 2.

B Extended Related Work

Robust control and games. Rectangular ambiguity enables dynamic programming in robust control (Iyengar, 2005; Nilim & El Ghaoui, 2005; Epstein & Schneider, 2003; Wiesemann et al., 2013), while coupled uncertainty and partial observation make rectangularity depend on timing and filtration (Goyal & Grand-Clément, 2023; Li & Shapiro, 2025; Bovy et al., 2024; Pahlke, 2022). Recent robust multi-agent models introduce nature and study equilibrium learning and sample complexity (Zhang et al., 2020; Shi et al., 2025); state-rectangular games connect robust equilibria to regularization (McMahan et al., 2024), while constrained and verification-oriented models study safety under uncertainty (Chang et al., 2026; He & Parker, 2026). Robust POMDP evaluation stresses an expressive nature-policy class (Krale et al., 2025). This literature asks which transitions nature may combine and how to compute robust value. We instead ask which private observations of nature a strategic opponent may use. Standard public-state, state–action-rectangular models satisfy

our plan-pasting condition and hence justify continuation-plan reselection across public-action and private-history branches, but rectangularity does not justify giving the opponent nature's private information.

Safety and adversarial robustness. Constrained MDPs and safe RL optimize performance subject to expected-cost or other safety requirements (Altman, 1999; García & Fernández, 2015), while adversarial RL trains policies against a learned disturbance player (Pinto et al., 2017). These formulations motivate worst-case training and constraint evaluation, but do not resolve the information boundary when a strategic opponent and transition nature are distinct entities. We show that the adversary's information set changes the robust reward and cost supplied to such an optimizer, potentially changing feasibility and policy selection.

Equation-level solver audit. Equation (2.1) of Zhang et al. (2020) selects one full public-state model before action averaging, corresponding to a shared action and a shared transition plan. Chang et al. (2026) assumes state–action-rectangular kernels, so continuation plans can be reselected row by row. On its printed grid, however, a fixed-policy check gives $4/9$ for the stated global-kernel family and $68/189$ for its local-row projections, isolating global-plan from rowwise quantifier order rather than reassessing its equilibrium (Appendix H.1).

Asymmetric information, teams, and public actions. Adversarial-team games separate common utility from communication: Celli & Gatti (2018) compare no, ex-ante, and in-play communication, while the team-public-information coordinator of Carminati et al. (2022) observes common information and issues private-state prescriptions. These prescription and public-action signaling mechanisms motivate our shared-row and posterior-split construction. One-sided games and common-information methods provide related Bayesian tools (Renault, 2006; Zheng et al., 2022; Nayyar et al., 2013; Kartik et al., 2021). These works distinguish information from communication within strategic teams, but not their interaction with robust ambiguity or nature's coupled transition commitments. We bring this shared-prescription logic into robust Markov games while retaining one complete post-action nature plan and separating communication from action-wise and private-history-wise plan reselection. The omniscient opponent in Aggarwal & How (2026) already observes the hidden type and can collude with nature's type-distribution choice. Their regret objective compares an uninformed team's policy with type-specific best responses. Here we compare expected losses before and after revealing nature's private history to the opponent, retaining the physical nature plan.

Information relaxations. Row cloning is a nonanticipativity relaxation (Brown et al., 2010; Haugh & Wang, 2014). Information projection, directed information, and KL control provide tools (Csiszár, 1975; Massey, 1990; Permuter et al., 2011), while sequential variational equalities identify the causal marginal as a KL minimizer (Charalambous & Stavrou, 2016). Our KL algebra is a finite specialization of their Theorem IV.1(A)(ii). The game-specific step is proving this marginal row remains admissible under the same nature plan, so other causal factors cancel even with persistent or cross-action-coupled transitions. Existing information-relaxation identities therefore provide the divergence algebra, but not the robust-game feasibility needed for a valid same-nature comparison.

Information regularization and the retained-plan comparison. Strouse et al. (2018) express $I(A; G \mid S)$ as expected KL between goal-conditioned and state-marginal action policies (their Equations (1)–(3)). Cundy et al. (2024) optimize disclosure of sensitive state through actions, including trajectory-level alternatives. Our diagnostic conditions on complete public history and compares the induced law with a legal opponent under the same nature plan. The contribution is this feasible game comparison and its value bound; the Q4 regularizer tests the resulting information–cost relation.

C Causal Measure Sets and Finite-Tree Lifting

For fixed q and defender policy π, let $\widehat{\mathcal{K}}^{\pi}_{\square,\mathrm{hist}}$ denote the globally compatible complete-history selectors for opponent class $\square \in \{\mathrm{nc}, \mathrm{fuse}\}$. Each selector records all opponent and nature rows, including off-path rows and persistent commitments. If Rec_q composes these rows with q, π, and

primitive chance, then

$$\mathcal{L}_{\square}(\pi;q) = \mathrm{Rec}_q(\widehat{\mathcal{K}}^{\pi}_{\square,\mathrm{hist}}), \qquad W_g^{\square}(\pi;q) = \sup_{P\in\mathrm{Rec}_q(\widehat{\mathcal{K}}^{\pi}_{\square,\mathrm{hist}})} \mathbb{E}_P[g]. \tag{19}$$

At an opponent information set I, nature's private history partitions I into refinements $r \in \mathcal{R}(I)$. NC imposes one action row x_I across all refinements; fusion replaces it by $(x_{I,r})_r$. Copying rows between a fused information set and its refinements proves both strategy inclusions. With q and one complete nature plan fixed, the common finite product kernel then proves the corresponding law identity. Thus fusion removes only the opponent's equal-row constraints. The reduced-state version follows from the row-compatible quotient conditions in Appendix F.

To connect these global values to the local ladder, let $\mathfrak{P}_{1,\square}$ contain the complete root suffixes with future opponent rows in class $\square$. For clarity, $V_{ij}[\mathfrak{P}]$ denotes the expression V_{ij} in Equations (10)–(11) evaluated with suffix-plan class $\mathfrak{P}$. Then

$$W_g^{\mathrm{nc}}(\pi;q) = V_{00}[\mathfrak{P}_{1,\mathrm{nc}}], \qquad W_g^{\mathrm{fuse}}(\pi;q) = V_{10}[\mathfrak{P}_{1,\mathrm{fuse}}]. \tag{20}$$

The brackets specify the suffix class in each support calculation.

D Action Rows, Complete Plans, and Rectangularity

D.1 Rows and labeled posterior decompositions

At a public history with private-history posterior β, a fused opponent row $\sigma_z(b)$ induces the public-action probability and posterior

$$\lambda_b = \sum_z \beta(z)\sigma_z(b), \qquad \beta_b(z) = \frac{\beta(z)\sigma_z(b)}{\lambda_b} \quad (\lambda_b > 0). \tag{21}$$

Conversely, every labeled Bayes-plausible split $\lambda \in \Delta(\mathcal{B})$ with $\sum_b \lambda_b\beta_b = \beta$ is implemented on $\mathrm{supp}(\beta)$ by $\sigma_z(b) = \lambda_b\beta_b(z)/\beta(z)$. Zero-mass branches are arbitrary. Hence fusion permits a Bayes-plausible posterior split, whereas NC restricts every reached action branch to the original posterior β.

This row–split equivalence concerns the current opponent action. The continuation still uses one complete suffix plan. Allowing a separate suffix for every public action gives V_{11}; selecting suffixes separately by private history also gives V_∞. With future opponent rows fully fused and pasteable, these are relaxations of nature-plan commitments. This distinction is formalized next.

D.2 Plan-preserving quantifier ladder

This section proves Theorem 1. The ordered values identify which maximization implements the intended information relaxation and which ones also relax the admissible complete adversarial suffix plan.

Why Q_p is a causal object. Fixing p fixes all current and future nature rows, future opponent rows, off-path prescriptions, and persistent commitments before the current action is forced. Together with the fixed defender, these objects define a unique finite product law on every (z,b) branch, so $Q_p(z,b)$ is its remaining expected loss. The same p must be used wherever a value is called plan-preserving. Recomputing the continuation after seeing b, or after seeing (z,b), changes the quantifier order to V_{11} or V_∞.

Write $\mathcal{Z}_+ = \{z : \beta(z) > 0\}$ and define

$$M_p(z) = \max_{b\in\mathcal{B}} Q_p(z,b), \qquad M(z) = \sup_{p\in\mathfrak{P}_t,\, b\in\mathcal{B}} Q_p(z,b). \tag{22}$$

Compactness and upper semicontinuity ensure attainment.

For the first equality, define the informed-optimal plan set and each plan's maximizing-action set by

$$\mathfrak{P}^\star_{10} = \arg\max_{p\in\mathfrak{P}_t} \sum\nolimits_{z\in\mathcal{Z}} \beta(z)\max_{b\in\mathcal{B}} Q_p(z,b), \qquad \mathcal{B}_p(z) = \arg\max_{b\in\mathcal{B}} Q_p(z,b). \tag{23}$$

Then, under attainment,

$$V_{00} = V_{10} \quad \Longleftrightarrow \quad \exists p^\star \in \mathfrak{P}_{10}^\star : \bigcap\nolimits_{z \in \mathrm{supp}(\beta)} \mathcal{B}_{p^\star}(z) \neq \varnothing. \tag{24}$$

Proof of Theorem 1. For fixed (p, b), $Q_p(z, b) \leq M_p(z)$ on every supported private history; averaging and optimizing gives $V_{00} \leq V_{10}$. A single plan p embeds in the action-indexed problem as the diagonal tuple $p_b = p$, which gives $V_{10} \leq V_{11}$. Finally, $\max_b Q_{p_b}(z, b) \leq M(z)$ pointwise, so $V_{11} \leq V_\infty$.

The same nonnegative-deficit argument gives the first equality condition in Equation (24): equality forces one V_{10}-optimal plan to admit a common maximizing action on every $z \in \mathcal{Z}_+$, and such a pair attains V_{00} conversely. For the second equality, let

$$\mathcal{H}((p_b)_b) = \sum_z \beta(z) \max_b Q_{p_b}(z, b), \qquad \mathfrak{D} = \{(p_b)_b : p_b = p \text{ for some } p\}. \tag{25}$$

Then

$$V_{10} = V_{11} \quad \Longleftrightarrow \quad \arg\max_{\mathfrak{P}_t^{\mathcal{B}}} \mathcal{H} \cap \mathfrak{D} \neq \varnothing. \tag{26}$$

This is an optimizer condition, not a claim that $\mathfrak{P}_t$ is action-rectangular. Similarly,

$$V_{11} = V_\infty \quad \Longleftrightarrow \quad \exists (p_b^\star)_b : \max_b Q_{p_b^\star}(z, b) = M(z) \quad \forall z \in \mathcal{Z}_+. \tag{27}$$

Indeed, the weighted sum of the pointwise nonnegative deficits vanishes if and only if every supported deficit vanishes.

It remains to show that all inclusions can be strict simultaneously. Let five private types be uniform, let $\mathcal{B} = \{L, R\}$, and let $\mathfrak{P}_t = \{p_0, p_1, p_2\}$. Define the success set $S_{p,b} = \{z : Q_p(z, b) = 1\}$ for each plan–action pair by

$$\begin{aligned} S_{p_0,L} &= \{1, 5\}, & S_{p_0,R} &= \{1, 2\}, \\ S_{p_1,L} &= \{3, 4\}, & S_{p_1,R} &= \{3, 4\}, \\ S_{p_2,L} &= \{1, 3\}, & S_{p_2,R} &= \{3, 4\}, \end{aligned} \tag{28}$$

and set $Q_p(z, b) = 0$ outside these sets. Every fixed plan–action pair covers two types; one plan with type-dependent actions covers at most three; one plan per action covers four, for example $S_{p_0,L} \cup S_{p_1,R}$; and the full envelope covers all five. Therefore

$$(V_{00}, V_{10}, V_{11}, V_\infty) = \left(\frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1\right). \tag{29}$$

The construction is a one-stage zero-sum game whose plan index is a persistent commitment shared by all counterfactual branches, so each successive value changes only the displayed quantifier order. □

Proposition 3 (A two-stage coupled-budget gap)**.** *Let Z be uniform on $\{0, 1\}$ and let the defender have one action. Before Z is realized, nature selects a complete-plan index $\iota \in \{0, 1\}$ and hence the plan p^ι. Its two action-branch masses $w^\iota = (w_0^\iota, w_1^\iota)$ are*

$$w^0 = (3/4, 1/4), \qquad w^1 = (1/4, 3/4). \tag{30}$$

At stage one the opponent chooses public action $B \in \{0, 1\}$. The selected plan then generates $X \in \{0, 1\}$. Writing $\mathbf{1}\{E\}$ for the indicator of an event E, its transition law is

$$\Pr_{p^\iota}(X = 1 \mid Z = z, B = b) = \mathbf{1}\{z = b\} w_b^\iota. \tag{31}$$

At stage two, after observing X, the opponent chooses $D_2 \in \{0, 1\}$ and incurs loss $D_2(2X - 1)$. Within each optimized profile, ι is an ex-ante deterministic complete-plan index, not a random variable or correlation seed. The information structures differ only in Z: the NC opponent does not observe it, whereas the fused opponent does. Then

$$(V_{00}, V_{10}, V_{11}, V_\infty) = \left(\frac{3}{8}, \frac{1}{2}, \frac{3}{4}, \frac{3}{4}\right). \tag{32}$$

Retaining the persistent allocation ι in the recursive state computes V_{10} exactly; dropping it and maximizing nature's support separately on each public-action branch computes V_{11}.

Proof. Let $J_2(x)$ denote the optimal stage-two continuation loss after observing $X = x$, and abbreviate $Q_\iota = Q_{p^\iota}$. Backward induction gives $J_2(x) = x$ and therefore $Q_\iota(z, b) = \mathbf{1}\{z = b\}w_b^\iota$. A shared opponent row achieves $V_{00} = 3/8$. A fused opponent can choose $b = z$, but the same allocation w^ι must serve both branches, so

$$V_{10} = \max_\iota \frac{1}{2}(w_0^\iota + w_1^\iota) = \frac{1}{2}. \tag{33}$$

The action-indexed relaxation instead combines the high rows of p^0 and p^1, giving $V_{11} = \frac{1}{2}(w_0^0 + w_1^1) = 3/4$; no branch can exceed $3/4$, so $V_\infty = 3/4$. The pasted allocation $(3/4, 3/4)$ has total mass $3/2$ and is infeasible. Retaining ι as a persistent state variable therefore recovers the exact value $V_{10} = 1/2$, whereas omitting it implements the V_{11} relaxation. □

D.3 Causal pasting and affine exact suffixes

Call a suffix-plan class *action-pasteable* if plans selected separately after each public action can be combined into one legal complete plan. Call it *type-within-action pasteable* if the same is true across private histories at a fixed action. These properties respectively imply $V_{10} = V_{11}$ and $V_{11} = V_\infty$; they are sufficient structural conditions rather than necessary conditions for equality under one criterion.

When both properties hold after every persistent parameter and remaining resource have been included in the state, define $Q^\star(z, b) = \sup_{p \in \mathfrak{P}_t} Q_p(z, b)$. Rowwise admissibility and perfect recall permit the optimizing suffix rows to be pasted on disjoint history subtrees. Let $\nu \in \Delta(\mathrm{supp}(\beta))$ be an interim private-history distribution. Fix (π, q), the public node, and a suffix class that is independent of ν, with future opponent rows fully fused. At the terminal layer the continuation is zero. Let $J_{t+1}(c', n')$ denote the complete-history continuation value at next public history c' and next private history n', and define its belief average by $v_{t+1}(c', \beta') = \sum_{n'} \beta'(n') J_{t+1}(c', n')$ for a next-stage belief β'. For a complete plan p and forced action b, write $\mathbb{E}^b_{\nu,p}$ for expectation under the resulting interventional branch law. Then

$$\mathbb{E}^b_{\nu,p}[v_{t+1}(C_{t+1}, \beta_{t+1})] = \mathbb{E}^b_{\nu,p}[J_{t+1}(C_{t+1}, N_{t+1})].$$

The tower property removes posterior denominators after averaging, while pasting attains the typewise envelope with one admissible plan. Denoting the forced-action value by $\Psi_b(\nu)$, induction gives

$$\Psi_b(\nu) = \sum_z \nu(z) Q^\star(z, b). \tag{34}$$

Consequently,

$$V_{00} = \max_b \sum_z \beta(z) Q^\star(z, b), \qquad V_{10} = V_{11} = V_\infty = \sum_z \beta(z) \max_b Q^\star(z, b). \tag{35}$$

The information gap vanishes exactly when one action maximizes $Q^\star(z, \cdot)$ for every $z \in \mathrm{supp}(\beta)$. A pathwise budget satisfies the pasting premise only when its residual resource is retained in the state. If admissible counterfactual rows remain coupled across actions, they must be optimized jointly; separate branch supports would instead enlarge nature's feasible complete-plan class.

E Causal Shadows and Leakage Budgets

This section proves Theorem 2. Unlike the local split operator, the argument works with complete raw-history profiles. It therefore does not require branchwise nature rows, a Markov state, or a separately optimized continuation.

E.1 The public shadow

Fix (π, q) and an annotated fused profile (σ, γ). Write

$$P = \mathrm{Law}_q(\pi, \sigma, \gamma). \tag{36}$$

For every public history c with $P(C_t = c) > 0$, define

$$\bar{\sigma}_t(b \mid c) = P(B_t = b \mid C_t = c). \tag{37}$$

Because the opponent action set is common to every nature-private refinement of c, this conditional marginal is a valid public opponent row. At a P-null public history, choose any fixed valid row. Retain the complete nature plan γ, including all off-path rows and every persistent commitment, and set

$$\bar{P} = \mathrm{Law}_q(\pi, \bar{\sigma}, \gamma). \tag{38}$$

The profile $(\bar{\sigma}, \gamma)$ is admissible in the noncommunicating game. No nature row has been reoptimized or pasted between action branches.

Same-nature likelihood-ratio identity. Under fresh entity-specific behavioral randomization and no private opponent–nature correlation seed, let τ be any complete public-history opponent profile and write

$$P^\tau = \mathrm{Law}_q(\pi, \tau, \gamma). \tag{39}$$

Then, as an extended-real identity,

$$D_{\mathrm{KL}}(P\|P^\tau) = \sum_{t=1}^{H} I_P(N_t; B_t \mid C_t) + \sum_{t=1}^{H} \mathbb{E}_{P(C_t)} D_{\mathrm{KL}}(\bar{\sigma}_t(\cdot \mid C_t)\|\tau_t(\cdot \mid C_t))\,. \tag{40}$$

In particular, $P \ll \bar{P}$ (that is, P is absolutely continuous with respect to $\bar{P}$) and

$$D_{\mathrm{KL}}(P\|\bar{P}) = \sum_{t=1}^{H} I_P(N_t; B_t \mid C_t). \tag{41}$$

The shadow is therefore KL-nearest over $\mathcal{L}_{\mathrm{nc}}(\pi; q, \gamma)$, uniquely on P-positive public histories. The logarithms are natural.

Derivation of the projection identity. The canonical product kernels of P and P^τ have the same initial law, defender rows, complete nature plan, and chance kernels. Only the opponent factor differs. Whenever $P \ll P^\tau$, their likelihood ratio is

$$\frac{\mathrm{d}P}{\mathrm{d}P^\tau} = \prod_{t=1}^{H} \frac{\sigma_t(B_t \mid C_t, N_t)}{\tau_t(B_t \mid C_t)}. \tag{42}$$

Insert $\bar{\sigma}_t(B_t \mid C_t)$ into each logarithm and take expectation under P. The first part is

$$\begin{aligned} &\mathbb{E}_P\left[\log \frac{\sigma_t(B_t \mid C_t, N_t)}{\bar{\sigma}_t(B_t \mid C_t)}\right] \\ &\quad = \mathbb{E}_{C_t, N_t}\left[D_{\mathrm{KL}}(P(B_t \mid C_t, N_t) \,\|\, P(B_t \mid C_t))\right] \\ &\quad = I_P(N_t; B_t \mid C_t). \end{aligned} \tag{43}$$

The remaining part is

$$\begin{aligned} &\mathbb{E}_P\left[\log \frac{\bar{\sigma}_t(B_t \mid C_t)}{\tau_t(B_t \mid C_t)}\right] \\ &\quad = \mathbb{E}_{P(C_t)} D_{\mathrm{KL}}(\bar{\sigma}_t(\cdot \mid C_t)\|\tau_t(\cdot \mid C_t))\,. \end{aligned} \tag{44}$$

Summing proves Equation (40). If τ assigns zero probability to a positive-mass action, both sides are infinite, so the extended-real identity still holds.

For $\tau = \bar{\sigma}$, a positive numerator at a reached (C_t, N_t) has a positive conditional average in Equation (37). Thus $P \ll \bar{P}$, and the second sum vanishes. Nonnegativity of KL proves minimality; equality forces $\tau_t = \bar{\sigma}_t$ on every P-positive public history. Zero-probability histories contribute neither a likelihood ratio nor mutual information. □

E.2 Why the causal assumptions matter

The projection uses two substantive assumptions. First, opponent and nature use separate behavioral randomization; a private shared seed could create a hidden correlation between their rows that is destroyed by marginalizing the opponent action. Second, nature-plan feasibility is independent of the numerical probabilities in the opponent row; otherwise replacing that row by its public marginal could make the retained nature plan infeasible. Dependence of nature's row on the realized public action and coupling within one complete plan remain allowed. Models violating either assumption require a different projection theorem.

E.3 Proof of the leakage-budget interpolation

For a complete fused law P, recall

$$\mathcal{I}(P) = \sum_{t=1}^{H} I_P(N_t; B_t \mid C_t). \tag{45}$$

Every noncommunicating profile satisfies $\mathcal{I}(P) = 0$. Conversely, if a fused profile has zero leakage, Equation (41) gives

$$D_{\mathrm{KL}}(P \| \bar{P}) = 0. \tag{46}$$

Hence $P = \bar{P}$ as raw trajectory laws, while $\bar{P}$ is generated by an admissible noncommunicating profile. Taking support values proves

$$W_g^0(\pi; q) = W_g^{\mathrm{nc}}(\pi; q). \tag{47}$$

The leakage-feasible sets are nested, so W_g^κ is nondecreasing. Let $H_P(B_t \mid C_t)$ denote conditional Shannon entropy under P, and let $\mathcal{B}_t(c)$ be the feasible opponent action set at public history c. At each stage,

$$I_P(N_t; B_t \mid C_t) \le H_P(B_t \mid C_t) \tag{48}$$

$$\le \log \max_c |\mathcal{B}_t(c)|. \tag{49}$$

Thus every fused profile is feasible at

$$\kappa_{\max} = \sum_{t=1}^{H} \log \max_c |\mathcal{B}_t(c)|, \tag{50}$$

and

$$W_g^{\kappa_{\max}}(\pi; q) = W_g^{\mathrm{fuse}}(\pi; q). \tag{51}$$

Now let g take values in an interval of width R_g. Total variation controls the expectation of any such criterion:

$$\mathbb{E}_P[g] \le \mathbb{E}_{\bar{P}}[g] + R_g \, \mathrm{TV}(P, \bar{P}). \tag{52}$$

The shadow is noncommunicating, so

$$\mathbb{E}_{\bar{P}}[g] \le W_g^{\mathrm{nc}}(\pi; q). \tag{53}$$

Pinsker's inequality and Equation (41) yield

$$\mathrm{TV}(P, \bar{P}) \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P \| \bar{P})} = \sqrt{\frac{1}{2} \mathcal{I}(P)}. \tag{54}$$

Combining the last three displays and taking the supremum over $\mathcal{I}(P) \le \kappa$ gives

$$W_g^\kappa(\pi; q) \le W_g^{\mathrm{nc}}(\pi; q) + R_g \sqrt{\frac{\kappa}{2}}. \tag{55}$$

The trivial span bound supplies the minimum with one in the main-text interpolation. This argument uses suprema and does not require an attained optimizer.

Suppose now that the finite-game optima are attained. If the fused and noncommunicating values agree, an NC optimizer is also fused-optimal and has zero leakage. Conversely, a zero-leakage fused optimizer has an NC shadow with the same raw law and value. This proves the final assertion of Theorem 2. The quantifier is existential: another fused optimizer can reveal information without improving the value.

The same proof applies after the robust outer supremum. The shadow retains the optimizer's own q and complete γ. Define $W_g^\kappa(\pi) = \sup_{q \in \mathfrak{Q}_{\rho,\varepsilon}} W_g^\kappa(\pi; q)$ over the same ambiguity family. Then

$$0 \le W_g^\kappa(\pi) - W_g^{\mathrm{nc}}(\pi) \le R_g \min\left\{1, \sqrt{\kappa/2}\right\}, \tag{56}$$

where the budget is imposed on each candidate raw law before the outer optimization. Under attainment, robust equality is equivalent to the existence of a robust fused optimizer with zero leakage.

F Full-History Realization and Causal Quotients

The complete-history construction above is always valid but may be too large for computation. The following sufficient conditions justify the smaller belief-state recursions used in the experiments.

For a candidate quotient $\eta_t(C_t, N_t) = (\chi_t, Z_t)$, let χ_t denote the public quotient state and Z_t the nature-private quotient state. Let Y_{t+1} collect the public innovation released before the next decision, and let ℓ_t denote the stage-t loss. At a reached public history, write $\beta_t = \mathrm{Law}(Z_t \mid C_t)$ for the conditional law of the private quotient state. Histories in the same quotient fiber must share a kernel K_t and public update ϕ_t satisfying

$$\begin{aligned}\mathrm{Law}(\ell_t, Y_{t+1}, Z_{t+1} \mid C_t, N_t, a, b, u) &= K_t(\cdot \mid \chi_t, Z_t, a, b, u),\\ \chi_{t+1} &= \phi_t(\chi_t, a, b, Y_{t+1}).\end{aligned} \tag{57}$$

Definition 4 (Row-compatible causal quotient)**.** *The map η_t is row-compatible for fixed π and an additive criterion when: (i) χ_t is measurable with respect to C_t, records the stage, and histories in one public quotient state have the same defender row; (ii) histories in one complete quotient state have the same action sets, nature-row set, stage-loss law, and kernel in Equation* (57)*; (iii) for every feasible belief and complete prescription, conditioning the joint kernel on each positive-mass public innovation gives the full-history posterior, and the law of the next $(\chi_{t+1}, \beta_{t+1})$ is independent of the representative public history; and (iv) every persistent nature commitment is recorded in the quotient state, after which feasible nature rows are recursively pasteable and convex under conditional mixing, with loss and transition expectations affine in the rows.*

Proposition 5 (Value preservation under a row-compatible quotient)**.** *Under Definition 4, define the quotient information state $\mathsf{I}_t = (\chi_t, \mathrm{Law}(Z_t \mid C_t))$. Backward recursion on I_t has the same NC and fused values as recursion on the complete history (C_t, N_t).*

Proof. At the terminal layer both recursions have value zero. Suppose their remaining additive values agree at stage $t+1$. Conditions (i)–(iii) make the conditional expected stage loss plus that continuation depend only on $\mathsf{I}_t = (\chi_t, \beta_t)$ and the selected opponent and nature rows. A quotient selector lifts by copying these rows to the corresponding complete histories; condition (iv) makes this a single admissible complete plan. Conversely, fix each reached public history $C_t = c$. Average forgotten private histories conditional on (c, Z_t) for the opponent row and on (c, Z_t, A_t, B_t) for nature's row. The NC row remains independent of Z_t because it was already shared at c. Condition (iv) preserves feasibility and the conditional one-step loss–transition expectations; condition (iii) preserves the next posterior law. The resulting prescription is feasible at $\mathsf{I}_t(c)$, so the conditional objective at each c is bounded by the quotient optimum there. Taking suprema gives equality by backward induction for NC and fused rows separately. Null branches have zero weight and can be assigned compatible rows. The argument preserves the additive optimal value. □

For an applied model, the audit is correspondingly short: verify within each candidate fiber the defender row, action alphabets, nature-row labels, stage loss, joint public-innovation/next-state kernel, and posterior update; then verify globally that every remaining commitment is stored in the state. The cyber model passes this audit for every ambiguity law used in the experiments.

G Sticky Weak-Signal Example

This optional example shows that accumulating public evidence need not remove the fusion gap. Let Θ be uniform on $\{-1, +1\}$. Conditional on Θ, the signals $Y_1, \ldots, Y_H \in \{-1, +1\}$ are independent and, before action t, signal Y_t is revealed with

$$\Pr(Y_t = \Theta \mid \Theta) = \frac{1 + \delta_H}{2}, \qquad \delta_H = \frac{1}{\sqrt{32H}}. \tag{58}$$

The opponent chooses $B_t \in \{-1, +1\}$ and the evaluator records $\ell_t = \mathbf{1}\{B_t = \Theta\}$ without revealing earlier losses. Define the cumulative loss $g = \sum_{t=1}^{H} \ell_t$. A fused opponent obtains H. The NC opponent is the Bayes classifier based on $Y_{1:t} = (Y_1, \ldots, Y_t)$. Let $P_+^t = \mathrm{Law}(Y_{1:t} \mid \Theta = +1)$

and $P_-^t = \mathrm{Law}(Y_{1:t} \mid \Theta = -1)$; its Bayes error is $e_t = (1 - \|P_+^t - P_-^t\|_{\mathrm{TV}})/2$. Conditional independence and Pinsker's inequality give

$$D_{\mathrm{KL}}(P_+^t \| P_-^t) \le 4t\delta_H^2 \le \frac{1}{8}, \qquad W_g^{\mathrm{fuse}} - W_g^{\mathrm{nc}} = \sum_{t=1}^{H} e_t \ge \frac{3H}{8}. \tag{59}$$

Thus each public signal is informative, yet the cumulative fusion gap remains linear in the horizon because each signal's strength is proportional to $H^{-1/2}$.

H Experimental Details and Additional Results

This section specifies the finite games and the matched learning intervention used in Section 6. The finite-game comparisons enumerate all relevant opponent prescriptions, nature plans, and extreme ambiguity laws. The learned comparison is intentionally separate: it asks whether the private information edge is discoverable by the same recurrent policy class under one fixed transition channel.

H.1 A fixed-policy audit of a published backup

We isolate the quantifier order in the nine-state grid of Chang et al. (2026). A state is the ordered pair $s = (i, j) \in \{1, 2, 3\}^2$ of player locations, and s' denotes the next state. Both players choose left, stay, or right, motion is clipped at the boundary, the discount factor is $\delta_{\mathrm{disc}} = 0.9$, and the initial state is uniform. Player 2 receives one when the next locations differ. The printed uncertainty family is $\{\mathsf{P}^{(0)}, \mathsf{P}^{(1)}, \mathsf{P}^{(2)}, \mathsf{P}^{(3)}\}$, where $\mathsf{P}^{(0)}$ is nominal. Under $\mathsf{P}^{(k)}$ for $k \in \{1, 2, 3\}$, the executed action of every player currently in cell k is uniform over the three moves, while other actions execute nominally. We use independent execution coins when both players occupy that cell.

Consider the deterministic policy in which both players always choose right. It is a complete realization of moving to cell 3 and remaining there, since a right move is blocked at the boundary. Let Q_k be the rational transition matrix induced by this policy and kernel $\mathsf{P}^{(k)}$, let r be the nine-vector whose entry $r(s')$ is the separation reward at next state s', let I_9 be the 9×9 identity matrix, and let $\mathbf{e}_9 \in \mathbb{R}^9$ be the all-ones column vector. Committing to one kernel for the whole discounted return gives

$$\underline{J}_{\mathrm{global}} = \min_{k \in \{0,1,2,3\}} \frac{1}{9} \mathbf{e}_9^\top (I_9 - \delta_{\mathrm{disc}} Q_k)^{-1} Q_k r = \frac{4}{9}. \tag{60}$$

The four quantities inside the minimum are $4/9$, $335/243$, $1463980/588627$, and $36160/9477$.

Local row minimization instead solves

$$v(s) = \min_{k \in \{0,1,2,3\}} \sum_{s'} Q_k(s, s') \big(r(s') + \delta_{\mathrm{disc}} v(s') \big). \tag{61}$$

Exact rational policy iteration returns the selector $(\mathsf{P}^{(0)}, \mathsf{P}^{(2)}, \mathsf{P}^{(3)}, \mathsf{P}^{(2)}, \mathsf{P}^{(0)}, \mathsf{P}^{(0)}, \mathsf{P}^{(3)}, \mathsf{P}^{(0)}, \mathsf{P}^{(0)})$ in lexicographic state order and

$$v = \left(0, \frac{20}{21}, \frac{2}{3}, \frac{20}{21}, 0, 0, \frac{2}{3}, 0, 0\right), \qquad \underline{J}_{\mathrm{row}} = \frac{68}{189}. \tag{62}$$

Thus the rectangular closure changes this fixed-policy value by exactly $16/189$. The gap also survives if simultaneous overrides share one common uniform move rather than independent coins.

As a specification check, evaluating the nine transition rows printed with the study reproduces its reported reward $85730/28509$ and cost $190/387$ exactly. Those displayed rows, however, combine cell-specific uncertainty patterns and do not identify a unique intended policy together with one member of the four-kernel family. We therefore use Equations (60)–(62) only to establish a strict fixed-policy quantifier gap; they make no claim about the reported equilibrium policy or value.

H.2 AMBIGUITY FAMILY AND NUMERICAL CONVENTIONS

Let $f : (0, \infty) \to \mathbb{R}$ be a convex generator with $f(1) = 0$. For full-support probability laws q and q_0 on $\mathcal{M}$, define the f-divergence

$$D_f(q\|q_0) = \sum_{\theta \in \mathcal{M}} q_0(\theta) f\left(\frac{q(\theta)}{q_0(\theta)}\right).$$

For a radius $\rho \geq 0$ and floor parameter $\varepsilon \in (0, 1]$, the experiments use the full-support family

$$\mathfrak{Q}_{\rho,\varepsilon} = \{q \in \Delta(\mathcal{M}) : D_f(q\|q_0) \leq \rho, \quad q(\theta) \geq \varepsilon q_0(\theta) \text{ for every } \theta\}, \tag{63}$$

where $\Delta(\mathcal{M})$ is the probability simplex on the finite mode space and q_0 is a full-support nominal law. The cyber and pursuit tasks specialize the divergence constraint to total variation. Finite recursions exhaustively enumerate the declared finite classes in double precision and compare independently constructed expressions at a tolerance of 10^{-9}. We reserve *exact* for analytic or rational identities; exhaustive numerical values are reported to their displayed precision, and learned averages and hardware timings are empirical measurements.

H.3 REPRODUCIBILITY AND COMPUTE AUDIT

The learning runs use an NVIDIA RTX 4080 SUPER, PyTorch 2.13.0+cu130 with CUDA build 13.0, `float64` tensors, and 16 intra/inter-op CPU threads. The parallel robustness runs use Python 3.12.3 under Linux/WSL2 and 16 workers. Table 3 reports end-to-end execution times. The numerical evaluations behind Figures 2, 3, and 8 use the same solvers and sampling distributions as the corresponding experiments.

Table 3: Recorded end-to-end wall times for the principal experiments, including training and evaluation where applicable. Each time measures one execution of the stated experiment.

Run	Wall time	Execution path
Exact semantic suite	11.27 s	CPU
Cyber policy table/audits	807.72 s	CPU
Channel robustness sweeps	3872.62 s	16 CPU workers
Learning and horizon sweep	693.37 s	RTX 4080 SUPER
Pursuit–evasion task	2.44 s	CPU
Learned-attacker certificates	518.27 s	RTX 4080 SUPER
Observation-model grid	2318.36 s	10 CPU workers

H.4 PLAN HIERARCHY AND PUBLIC-SHADOW CHECKS

The strict construction has five equally likely private types, three complete plans, and two opponent actions, as defined in Equation (28). It realizes

$$(V_{00}, V_{10}, V_{11}, V_\infty) = \left(\frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1\right). \tag{64}$$

Each value is computed under its own quantifier order. A stress test draws one million tables with conditional values uniform on $\{0, 1/3, 2/3, 1\}$ and a five-type belief from normalized positive weights. It evaluates all four orders, the three inequalities, and the diagonal optimizer condition.

The shadow study samples one hundred thousand two-stage profiles with two modes, actions, nature rows, and public outcomes. For each profile it retains the complete nature plan, constructs the public opponent marginal, and evaluates both sides of Equation (15). One independently sampled alternative public profile per fused profile tests the full Pythagorean identity. Simplex rows normalize positive uniform weights, and losses are uniform on $[0, 1]$; Table 4 reports the checks.

The three ladder inequalities are simultaneously strict on 35.0% of the random tables. This frequency is descriptive of the sampled family; the formal non-vacuity claim follows from the explicit strict construction.

Table 4: Finite-game semantic checks. A residual is the largest absolute difference between independently evaluated sides of an asserted identity.

Check	Tested units	Violations	Maximum residual
Plan hierarchy	1,000,000	0	–
Diagonal optimizer test	1,000,000	0	0.0
Public-shadow KL identity	100,000	0	1.3×10^{-15}
Horizon-four quotient histories	20,268	0	0.0

The criterion-dependent fixture is a one-shot game with a uniform mode $\Theta \in \{0, 1\}$ and one opponent action $b \in \{0, 1\}$. Its safety loss is the mode-matching indicator $\mathbf{1}\{b = \Theta\}$, whereas its mode-independent availability criterion is $\mathbf{1}\{b = 1\}$. Exact evaluation gives a fused-minus-NC gap of $1/2$ for safety and 0 for reward on the same game, and the safety-optimal fused profile carries $\log 2$ nats of action leakage against none for the reward-optimal profile. This realizes the criterion-specific exactness of Section 3: the same information relaxation is exact for one trajectory criterion and strictly conservative for another.

For the uniform one-stage matching game, the analytic budget curve is $W^\kappa = w$, where $w \in [1/2, 1]$ solves $\log 2 - h_2(w) = \kappa$ for $0 \le \kappa \le \log 2$, and $W^\kappa = 1$ thereafter; $h_2(w) = -w \log w - (1 - w) \log(1 - w)$ is binary entropy, with $0 \log 0 = 0$. Averaging a channel with its simultaneous mode/action relabeling preserves matching probability and cannot increase mutual information, so a symmetric channel attains the optimum. For the discrete check, let W denote the matching probability. The two Bernoulli opponent rows are evaluated on a 4096×4096 grid and grouped into 512 information-budget bins over $[0, \log 2]$. The prefix maximum in each bin gives the reported discrete frontier. The first point is $(\kappa, W) = (0.001354, 0.526007)$, compared with the theoretical envelope 0.526017; the last point reaches the fused value 1.0. A double-precision calculation on the same grid agrees to 7.5×10^{-8}.

For scalars $l < u$, let $\mathrm{Unif}(l, u)$ denote the continuous uniform distribution on $[l, u]$. The scaling family has two actions, nature rows, and public outcomes. Transitions and the mode law normalize independent $\mathrm{Unif}(0.05, 1.05)$ weights, while losses are $\mathrm{Unif}(0, 1)$. Write $K_{\mathrm{mode}} = |\mathcal{M}|$ for the number of nature modes. For $K_{\mathrm{mode}} = 2$, horizons $H = 1, 2, 3, 4$ use $200, 200, 100, 20$ thousand games, respectively. For $K_{\mathrm{mode}} = 3$, horizons $H = 1, 2, 3$ use $200, 100, 8$ thousand games; for $K_{\mathrm{mode}} = 4$, horizons $H = 1, 2$ use $100, 20$ thousand games; and $(K_{\mathrm{mode}}, H) = (5, 1)$ uses 50 thousand games. A batching scan repeats $10, 50, 100, 200$, and 400 thousand $(2, 2)$ games. Thus the controlled grid contains 998,000 games and the fixed-size scan contains 760,000 additional evaluations.

H.5 Sticky weak signals

Appendix G defines a uniform binary mode and, before each decision, a binary public signal of accuracy $(1 + \delta_H)/2$ with $\delta_H = 1/\sqrt{32H}$. The fused opponent knows the mode, whereas the NC opponent uses the Bayes classifier based on all signals observed so far. We evaluate horizons from 4 to 16384. The per-signal log-likelihood ratio falls from 0.1772 to 0.0028, while the normalized gap increases from 0.4448 to 0.4531 and remains above the analytic $3/8$ lower bound. The experiment illustrates that many individually weak public signals need not eliminate the value of direct mode access. Figure 8(c) plots both quantities against the horizon.

H.6 Robust cyber defense

The cyber environment has three services and a persistent vulnerability mode $\Theta \in \{0, 1, 2\}$. At each of four stages the defender may serve, patch one service, or patch all services. The attacker chooses from idle, probe, and one exploit per service. A burned attacker is restricted to idle thereafter. Patches protect only the current stage. Nature sees the regime and may choose either a stealth or loud post-action row after every joint action. The opponent sees public actions and outcomes under NC and additionally sees the regime under fusion.

For every complete nature information set $J = (t, C_t, N_t, A_t, B_t)$ in the primary task, let $\mathcal{U}_J = \{\text{stealth}, \text{loud}\}$. Let $\mathfrak{I}_n^{\text{cyb}}$ denote the set of all such nature information sets. Then

$$\Gamma_n(q) = \prod_{J \in \mathfrak{I}_n^{\text{cyb}}} \Delta(\mathcal{U}_J). \tag{65}$$

At the added public-revelation stage, $\mathcal{U}_J$ is a singleton. There is no residual budget or shared row parameter. Conditional mixtures therefore remain feasible, and loss and transition expectations are affine in the selected row. This analytically verifies the global product and convexity clauses in condition (iv) of Definition 4 for every ambiguity law.

The nominal regime law is uniform, the primary total-variation radius is 0.08, and each regime retains at least 30% of its nominal mass. Loud-row probe alerts have probabilities $(0.75, 0.50, 0.25)$; stealth moves each by 0.10 toward 0.50. A matching exploit breaches unless patched and alerts with probability 0.35 (stealth) or 0.60 (loud). A mismatched exploit never breaches: it burns with probability 0.80 and, conditional on surviving, alerts with probability 0.60 or 0.90. Idle never alerts. Public outcomes are no-alert, alert, and burn; breach is a stage cost that may co-occur with the first two. Thus a patched matching exploit can alert despite zero breach cost. Breach costs are recorded by the evaluator and are not an additional within-episode observation for either policy. Serving, patching one service, and patching all services yield availability rewards 1.0, 0.6, and 0.1 before a 1.5 breach penalty. Safety is expected breach count, with threshold 1.4.

Let S denote serve, P_{all} patch-all, P_k patch service $k \in \{0, 1, 2\}$, and m_t the public alert count clipped at three, with a burn counted as an alert. The nine policies are S, P_0, P_{all}, $P_{(t-1) \bmod 3}$, serve-once-then-rotate (S at $t = 1$ and $P_{(t-2) \bmod 3}$ thereafter), and serve-twice-then-rotate (S through $t = 2$ and $P_{(t-3) \bmod 3}$ thereafter), and three alert-triggered rules: use S while $m_t = 0$, then use respectively P_0, $P_{(t-1) \bmod 3}$, or $P_{\min\{m_t - 1, 2\}}$.

The recursive state is the clipped alert count, burned status, and, when present, the revealed mode. We exhaustively check its horizon-four causal signature for all nine policies and both reward and safety criteria. Across $20{,}268$ reachable pre-action histories grouped into 288 fibers, $5{,}036{,}040$ mode–action–row–outcome comparisons agree on action sets, defender rows, transition probabilities, losses, next quotient states, and residual commitments, with zero violations. The robust evaluation then maximizes over every ambiguity vertex and compatible adversarial continuation. Separately, at horizon three and the fixed audit vertex $q = (0.2533333333, 0.3333333333, 0.4133333333)$, the quotient recursion matches an uncompressed public-history recursion retaining Θ to zero error for all nine policies, both criteria, and both information semantics; their recorded wall times are 5.32 and 24.86 seconds, respectively.

For serve-once-then-rotate, the worst fused profile has value 3.000 and action leakage 1.079 nats. Its opponent-row marginal is a feasible same-nature NC witness of value 1.038, below the separately optimized NC value 1.244. Publicly announcing the regime, or making breach transitions regime-independent, closes the gap. At radii 0.00, 0.04, 0.08, and 0.16, the NC costs are 1.008, 1.124, 1.244, and 1.483; the fused cost remains 3.000. In the mode-independent control, probe alerts have probability $1/2$, every exploit is treated as matching in every mode, and its alert probability is 0.35 or 0.60 under stealth or loud. Consequently no mismatch burn occurs. Patch protection, availability rewards, and the defender policy are unchanged.

Finite-library selector. For a declared finite policy library Π_0, one safety cost $C^{\square} = C_1^{\square}$, and threshold d, define

$$\pi_{\square}(d) \in \arg\max_{\pi \in \Pi_0 :\, C^{\square}(\pi) \le d} R^{\square}(\pi), \qquad \square \in \{\text{nc}, \text{fuse}\}. \tag{66}$$

An empty feasible set leaves $\pi_{\square}(d)$ undefined. Otherwise, ties are resolved first by lower same-semantics safety cost and then by one fixed policy order. A counted reversal requires both selector disagreement and fused infeasibility of $\pi_{\text{nc}}(d)$; reward forfeiture compares the two choices using R^{nc}.

Robustness of the reversal. Because the robust cost/reward table does not depend on the safety threshold, the two selectors in Equation (66) are read off for every d. Sweeping d over the attainable cost range, the information-correct selector is NC-feasible but fused-infeasible for a contiguous band

$d \in [1.240, 4.000)$, which is 69% of that range; the reward it forfeits ranges from 0.140 to 1.080 across the band and equals 0.535 at the primary threshold $d = 1.4$ (Table 1).

For the channel/reward audit, let (m_s, m_l) be the stealth and loud alert probabilities of a matching exploit, and (x_s, x_l) those of a mismatched exploit. We sample

$$\begin{aligned} m_s &\sim \mathrm{Unif}(0.20, 0.50), & m_l &\sim \mathrm{Unif}(\max\{m_s + 0.05, 0.50\}, 0.75), \\ x_s &\sim \mathrm{Unif}(\max\{m_s, 0.50\}, 0.75), & x_l &\sim \mathrm{Unif}(\max\{x_s + 0.05, 0.75\}, 0.95). \end{aligned} \tag{67}$$

The three probe-alert probabilities are sampled independently from $[0.55, 0.85]$, $[0.40, 0.60]$, and $[0.15, 0.45]$; burn probability from $[0.50, 0.95]$; probe garbling from $[0.05, 0.15]$; and breach reward penalty from $[1.00, 2.50]$. Draws are rounded to four decimals, all remaining task parameters stay at their primary values, and each robust table evaluates the same nine policies exhaustively. This authored generator is a controlled stress-test distribution, not an estimate of operational prevalence, and the declared library is not the complete defender policy class.

At horizon three, both the any-threshold and fixed-$d = 1.4$ joint reversal rates are $163/192 = 84.9\%$, with Wilson 95% interval $[79.1\%, 89.3\%]$. We reevaluate the same first 32 configurations at horizon four, changing no sampled parameter: $29/32 = 90.6\%$ admit a reversal threshold and $23/32 = 71.9\%$ reverse at $d = 1.4$, with intervals $[75.8\%, 96.8\%]$ and $[54.6\%, 84.4\%]$, respectively. Every observed window is contiguous. Conditional on the 163 and 29 any-threshold reversals, the median normalized widths are 0.686 and 0.688; their respective 5–95% ranges are $[0.587, 0.723]$ and $[0.579, 0.690]$. The corresponding median widths are 2.057 and 2.753 cost units. Thus the fully matched horizon-four audit preserves the broad threshold phenomenon while separating it from the stricter fixed-threshold estimand.

Observation-model sensitivity. Two parameters change how quickly the public history identifies the mode. A breach-detection probability φ publicly flags an unshielded matching exploit with probability φ; the flag counts as an alert and identifies the mode. A probe-informativeness scale s replaces the loud-row probe pattern p_θ by $0.5 + s(p_\theta - 0.5)$, so $s = 0$ makes probes uninformative and $s = 2$ gives the largest probe contrast in this sweep; the stealth row still moves the pattern toward 0.5. Every other primary setting is unchanged, and $(\varphi, s) = (0, 1)$ reproduces the primary table exactly. For each of the 25 cells we evaluate the nine declared policies exactly, apply both selectors at $d = 1.4$, and recompute the reversal window; the quotient recursion is checked against the full-history recursion at horizon three in every cell, with maximum discrepancy 1.1×10^{-16}. Figure 7 shows the result. The NC cost of serve-once-then-rotate rises from 1.244 to 1.595 as the public history becomes more informative, whereas its fused cost stays at 3.000, so the safety gap ranges from 1.405 to 1.756. The selection at $d = 1.4$ reverses in 18 of 25 cells with forfeited reward between 0.338 and 0.535, and the reversal window covers between 0.351 and 0.690 of the attainable cost range. The cells without a reversal at $d = 1.4$ are $(0, 2)$, $(0.25, 2)$, $(0.5, 2)$, $(0.75, 2)$, $(1, 1)$, $(1, 1.5)$, $(1, 2)$ in (φ, s) coordinates; there the NC cost of serve-once-then-rotate is at least 1.414, so both selectors choose patch-all at this threshold, while the reversal window still covers at least 0.351 of the range.

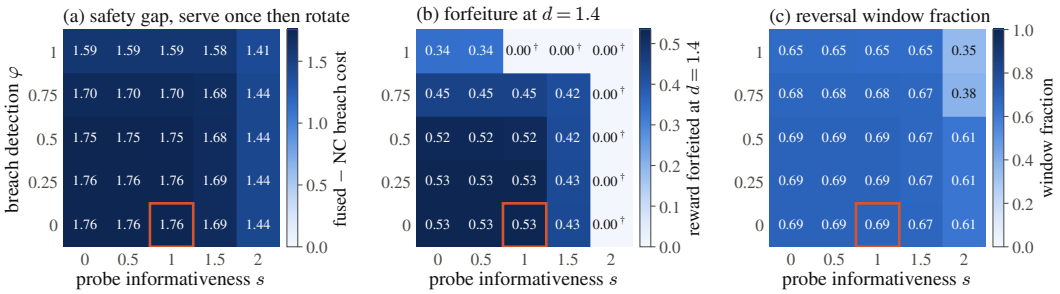


Figure 7: **Observation-model sensitivity.** Exact cyber results over the public breach-detection probability φ and the probe-informativeness scale s; the outlined cell is the primary task. (a) Fused-minus-NC safety cost of serve-once-then-rotate. (b) Reward forfeited by fused selection at $d = 1.4$ (a dagger marks cells without a reversal at that threshold). (c) Fraction of the attainable cost range on which the two selectors disagree.

H.7 EXPANDED AND COUPLED POLICY SELECTION

Does reversal persist in a larger defender class? We evaluate every four-stage schedule over the five defender actions, its variant that serves after a public burn, and the nine original policies. Deduplicating their reachable public action tables gives $1{,}248$ policies. At the same $d = 1.4$ and ambiguity radius 0.08, 742 are NC-feasible and 34 are fused-feasible. Nominal planning uses the uniform prior and a fixed $50/50$ stealth/loud mixture with an NC opponent. Nominal and robust NC select the same policy: serve twice, then patch all, switching to serve after burn. Its robust NC cost is 0.82667, whereas its fused cost is 2.00000. Table 5 shows that fused selection forfeits 1.57376 NC reward in this larger class.

Table 5: Selection over $1{,}248$ public defender policies at $d = 1.4$. The last three columns evaluate each selected policy. The informed rowwise backup is an independently implemented check of the fused information model. Each feasible count uses its row's evaluation semantics at the common threshold.

Selector	Feasible	C^{nc}	C^{fuse}	R^{nc}
Nominal	1248	0.82667	2.00000	1.97376
Robust NC	742	0.82667	2.00000	1.97376
Fused robust	34	0.00000	0.00000	0.40000
Informed rowwise robust	34	0.00000	0.00000	0.40000

Can plan reselection itself change the selected policy? In the two-stage shared-budget game, the defender protects neither action branch, one branch, or both. At threshold $d = 3/5$, the plan-preserving fused value selects no protection, with cost $1/2$ and availability 1. Action-wise plan reselection raises that cost to $3/4$ and selects one-branch protection, whose availability is $17/20$. The forfeiture is $3/20$ and persists throughout $d \in [1/2, 3/4)$. Public revelation removes the information gap while retaining this plan-reselection gap (Appendix H.7).

The expanded cyber class is the union of all $5^4 = 625$ deterministic action schedules, their 625 serve-after-burn variants, and the nine policies in Table 1. Equality of action tables on the 16 reachable $(t, \text{alert count}, \text{burned})$ summaries removes duplicates, leaving $1{,}248$ policies. Every policy uses public information. Nominal evaluation fixes $q = q_0$ and mixes stealth/loud rows equally; robust NC and fused evaluation optimize over the original ambiguity vertices and complete row assignments. Reward and safety retain separate adversarial optimizations. Expanded-class ties use canonical action-table order within a 10^{-8} tolerance.

Since observations are independent of the defender action, the solver shares a likelihood graph across policies. Each node stores the public alert count and an integer likelihood vector reduced by its greatest common divisor; the current prior then recovers the full posterior. The live layers contain 1, 57, $1{,}441$, and $22{,}029$ nodes. Burned suffixes have only the idle opponent action and are evaluated directly. The NC backup maximizes over one shared opponent action and one joint mode-indexed nature prescription before summing over public outcomes. A separate scalar Bellman implementation maximizes over opponent-action/nature-row pairs conditional on the persistent mode, reproducing fused values to 8.89×10^{-16}.

All 114 comparisons against the original engine and archived H4 values pass with maximum discrepancy 1.97×10^{-10}, including a fresh nonuniform-prior check of the new burn-responsive policy. The revealed-mode control closes the NC–fused gap exactly, and the mode-independent control has maximum residual 4.01×10^{-10}. The full execution, including these checks, takes 55.29 seconds with Python 3.12.3 and NumPy 2.5.2 on CPU. The nominal and robust NC optima coincide; the fused optimum ties patch-all with its serve-after-burn variant. Exactly 708 policies are NC-feasible and fused-infeasible.

For the coupled selection experiment, extend Proposition 3 by an ex-ante protection set $\mathcal{D} \in \{\varnothing, \{0\}, \{1\}, \{0, 1\}\}$. Protection changes the success probability to $\Pr(X = 1 \mid z, b, p^{\iota}, \mathcal{D}) = \mathbf{1}\{z = b,\ b \notin \mathcal{D}\} w_b^{\iota}$; the complete allocations remain $(3/4, 1/4)$ and $(1/4, 3/4)$. Availability is 1, $17/20$, $17/20$, and $1/5$, respectively, independent of the adversarial actions. Rational enumeration gives Table 6; ties use the displayed policy order.

Table 6: Exact safety costs under a shared transition budget. At $d = 3/5$, V_{00} and V_{10} select no protection; V_{11} and V_∞ select one-branch protection.

Protected branches	Availability	V_{00}	V_{10}	V_{11}	V_∞
Neither	1	$3/8$	$1/2$	$3/4$	$3/4$
Branch 0	$17/20$	$3/8$	$3/8$	$3/8$	$3/8$
Branch 1	$17/20$	$3/8$	$3/8$	$3/8$	$3/8$
Both	$1/5$	0	0	0	0

The sensitivity grid replaces $3/4$ by $h \in \{1/2, 5/8, 3/4, 7/8, 1\}$ and $1/4$ by $1 - h$. For no protection, $(V_{00}, V_{10}, V_{11}, V_\infty) = (h/2, 1/2, h, h)$; either single protection has cost $h/2$ in every order. The plan-only selection band is $[1/2, h)$, empty at $h = 1/2$. Revealing the mode gives $V_{00} = V_{10}$ while preserving the shared budget; legally replacing the coupled complete-allocation family by its rectangular closure $\{1 - h, h\}^2$ instead closes the plan-reselection gap. Across the primary and both controls, $41{,}952$ profile objectives and $1{,}064$ assertions are checked with rational arithmetic in 0.20 seconds.

H.8 Matched learning intervention

The learning study freezes the serve-once-then-rotate defender and one post-action channel shared by every condition. Independently at each stage, it selects the loud row with probability 0.30 and stealth otherwise; neither the mixture nor primitive randomness depends on attacker information. The public input contains the stage, clipped alert count, burned status, previous public action, and resulting public outcome—enough to update the finite-program belief.

A single-layer GRU receives this input and an additional three-dimensional coordinate: zero under NC, the true regime one-hot under fusion, or an independent uniform one-hot sampled once per episode under the nuisance control. All three actors have $16{,}261$ parameters and the same optimization budget. Training and final evaluation pair primitive streams across conditions; evaluation rollouts are fresh and disjoint from training.

The 64-unit GRU is optimized by policy gradient with Adam (learning rate 0.003, betas $(0.9, 0.999)$, epsilon 10^{-8}, no weight decay or gradient clipping), a 0.01 entropy bonus, and PyTorch's default linear-layer initialization. Let G be an episode return and let $\bar{G}$ and s_G be its batch mean and batch standard deviation, respectively. The standardized batch return is $(G - \bar{G})/(s_G + 10^{-6})$. At the primary horizon four, each of 300 updates uses 8192 parallel games in one compiled CUDA graph. Each policy is evaluated on $262{,}144$ fresh episodes and compared in Table 7 with the unrestricted public-history dynamic program under the same channel.

Table 7: Matched learning intervention at horizon four. Differences use unrounded paired means. The upper reference is the value of the corresponding unrestricted public-history dynamic program, not a fitted baseline.

Attacker information	Learned cost	Upper reference	Difference from NC
NC (masked regime)	0.9989	1.0027	–
Matched-marginal nuisance	0.9988	1.0027	−0.0001
Fused (true regime)	2.9999	3.0000	2.0010

Means use sixteen paired runs. A 95% paired percentile bootstrap with 5000 resamples gives $[1.9999, 2.0022]$ for fused minus NC and $[-0.00056, 0.00031]$ for nuisance minus NC. As a secondary post-run analysis, we resample the eight paired run differences at each horizon without retraining; Table 8 gives the resulting intervals. The horizon-four row reuses the first eight primary pairs.

We also compare returns in $[0, H]$ with their dynamic-program upper references using $n = 262{,}144$ evaluation episodes per comparison and simultaneous radius $H\sqrt{\log(128/0.001)/(2n)}$: all 128 labeled comparisons pass across the three primary conditions and five-horizon NC/fused sweep,

Table 8: Paired-run uncertainty for the learned horizon sweep. Intervals are 95% percentile intervals from 10,000 deterministic resamples of eight paired differences per horizon.

H	Mean fused–NC gap	Paired interval
2	0.99971	[0.99864, 1.00081]
3	1.33292	[1.33150, 1.33415]
4	2.00052	[1.99932, 2.00169]
5	2.33447	[2.33207, 2.33743]
6	2.83342	[2.70709, 2.95930]

including the reused horizon-four subset. Only a regime-predictive coordinate produces the change; equal-dimensional nuisance input does not.

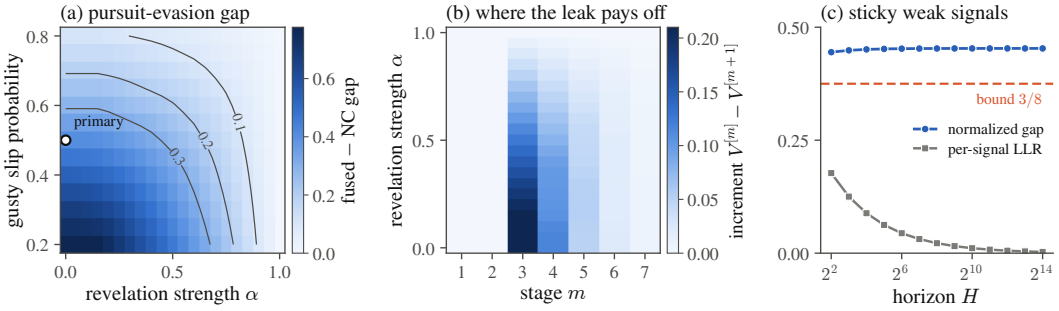


Figure 8: **Secondary tasks.** (a) Exact fused–NC capture-time gap in the pursuit–evasion task as the revelation strength α and the gusty slip probability vary; the circle is the primary setting, and revelation closes the gap at every slip level. (b) Stagewise hybrid increments $V^{[m]} - V^{[m+1]}$ at the mode law $q(+1) = 0.58$, where stage 1 releases the public signal; the fused evader's advantage is concentrated in the early movement stages and decays along the horizon. (c) Sticky weak-signal family: the normalized gap stays above $3/8$ while the per-signal log-likelihood ratio vanishes.

H.9 Pursuit–evasion secondary task

The secondary task uses a seven-position ring, initial positions $(0, 3)$, six stages, and a wind mode $\Theta \in \{-1, +1\}$. After the public signal, the evader announces $b \in \{-1, 0, 1\}$ and nature selects a calm or gusty row; displacement is b without a slip and $b + \Theta$ after one, while the pursuer moves toward the evader's pre-move position. Slip probabilities are 0.20 and 0.50. Let d_{ring} denote the current shortest-arc distance. The pursuer traverses the shorter arc by $\min\{2, d_{\text{ring}}\}$, with clockwise tie-breaking. Coincident updates cause capture and absorbing zero loss; otherwise the loss is one. For $q(+1) := q(\Theta = +1) \in [0.42, 0.58]$, convexity places the robust maximum at an endpoint; $q(+1) = 0.50$ is also checked.

Let $\alpha \in [0, 1]$ denote the public revelation strength. One public signal is drawn before the six movement stages and identifies the wind with probability $(1+\alpha)/2$. At $\alpha = 0, 0.25, 0.50, 0.75, 1$, the NC values are $5.594, 5.629, 5.725, 5.856, 6.000$, while every fused value is 6.000; the gap therefore decreases monotonically from 0.406 to zero. A regime-independent control that splits slips equally between directions gives a gap below 10^{-15}. Figure 8(a,b) repeats the exact recursion over a grid of revelation strengths and gusty slip probabilities and reports the stagewise hybrid increments: the gap closes at every slip level as α increases, and the fused evader's advantage is concentrated in the early movement stages.

For these hybrids, $V^{[m]}$ withholds the private mode from the evader before stage m and reveals it from stage m onward; stage one is the public-signal stage. Each value optimizes opponent behavior and nature's complete plan within the same admissible family. The increments $V^{[m]} - V^{[m+1]}$ therefore compare successive revelation times.

H.10 SAMPLE-BASED EVALUATION OF A LEARNED ATTACKER

The reference game is the fixed-channel serve-once-then-rotate game of Appendix H.8 at horizon four with the uniform mode law. A trained attacker defines an exact fused profile: its action row at every public history and mode is obtained by running the recurrent network along that history, and the full-history recursion then gives the trajectory law, the leakage of Equation (13), the public shadow of Equation (12), and all values. The four unpenalized attackers have exact leakage between 1.0982 and 1.0984 nats, exact cost at least 2.9998, and exact shadow value at most 1.0000; the NC attacker, which never receives the mode, has exact leakage zero. The KL identity holds for all four attackers to 6.0×10^{-15}. Their action rows depend on nature's private history only through Θ, so $I_P(N_t; B_t \mid C_t) = I_P(\Theta; B_t \mid C_t)$. The fixed defender patches each service once during stages two through four, so every trajectory has at most three breaches. We therefore use $R_g = 3$ for this policy's certificate. The exact fused optimum has leakage $\log 3 = 1.099$ nats to three decimals.

Table 9: Leakage-penalized attackers: exact leakage and exact breach cost, averaged over the trained runs at each penalty weight, with the range across runs in brackets.

λ	Leakage (nats)	Breach cost	λ	Leakage (nats)	Breach cost
0	1.098 [1.098, 1.098]	3.000 [3.000, 3.000]	2.5	0.020 [0.000, 0.079]	1.144 [1.000, 1.578]
0.5	1.098 [1.098, 1.098]	3.000 [3.000, 3.000]	3	0.009 [0.000, 0.035]	1.080 [1.000, 1.320]
1	0.879 [0.221, 1.098]	2.691 [1.763, 3.000]	4	0.000 [0.000, 0.000]	1.000 [1.000, 1.000]
1.25	0.338 [0.161, 0.401]	2.143 [1.663, 2.308]	6	0.000 [0.000, 0.000]	1.000 [1.000, 1.000]
1.5	0.150 [0.120, 0.239]	1.690 [1.579, 2.015]	8	0.001 [0.000, 0.005]	1.031 [1.000, 1.127]
1.75	0.112 [0.093, 0.168]	1.598 [1.514, 1.850]	12	0.000 [0.000, 0.000]	1.000 [1.000, 1.000]
2	0.050 [0.000, 0.126]	1.299 [1.000, 1.734]	16	0.000 [0.000, 0.000]	1.000 [1.000, 1.000]

The rollout estimates use $n \in \{2^8, 2^{10}, 2^{12}, 2^{14}, 2^{16}, 2^{18}\}$ episodes, with three independent draws for each attacker and sample size. Each episode records its mode label for evaluation. The leakage estimator is the plug-in conditional mutual information $\sum_t \hat{I}(\Theta; B_t \mid C_t)$ computed from empirical counts over the finite public histories. The tabular shadow sets $\hat{\hat{\sigma}}_t(b \mid c)$ to the empirical frequency of action b among episodes that reach history c, uniform over feasible actions at unseen histories (idle after burn), and is evaluated both exactly and on 2^{18} fresh episodes. With 2^8 episodes the leakage error is at most 0.0084 nats and the tabular shadow's value error at most 8.2×10^{-6}; with 2^{18} episodes the respective maxima are below 0.0003 and 9.0×10^{-6}. These trained attackers are nearly deterministic given the public history and mode, concentrating most trajectory probability on a small set of histories.

Leakage-penalized attackers add $\lambda \widehat{L}$ to the Q3 training loss, where a batch of M episodes gives

$$\widehat{L} = \sum_{t=1}^{H} \frac{1}{M} \sum_{i=1}^{M} D_{\mathrm{KL}}\Big(\sigma_t(\cdot \mid C_t^{(i)}, \Theta^{(i)}) \| \widehat{m}_t(\cdot \mid C_t^{(i)})\Big). \tag{68}$$

Here $\widehat{m}_t(\cdot \mid c)$ averages the network's action rows over batch episodes with public history c. Gradients pass through both action rows in each KL term, with the sampled histories held fixed. This estimates leakage from policy probabilities; the separate rollout estimator above uses observed action counts. We train four attackers for each $\lambda \in \{0, 0.5, 1, 1.25, 1.5, 1.75, 2, 2.5, 3, 4, 6, 8, 12, 16\}$ with the Q3 optimizer and budget and report the exact enumerated leakage and cost of every trained attacker (Table 9). Larger penalties move the attackers toward the NC value; the intermediate weights yield the run ranges in Table 9.