LEARNING CONTROLLED DYNAMICS THROUGH OBEDIENT RECOMMENDATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning controlled dynamics for model-based reinforcement learning requires informative actions. When strategic receivers choose whether to execute those actions, observability also depends on which experiments recommendations can implement. We study this dependence on recommendation visibility. In a transfer-free two-stage system with informed, far-sighted receivers, public and receiver-private recommendations have identical dynamics, utilities, sender objectives, learner feedback, and worst-equilibrium known-model values. Nevertheless, minimax regret under worst-case sequential-equilibrium selection is linear publicly and logarithmic privately for fixed model alternatives. Private signals uniquely implement actions that reach and activate an informative transition, whereas public signaling admits a common uninformative equilibrium selection. Under prescribed obedience, we characterize how learning changes experiment access through the obedient observability closure: evidence eliminates candidate models, relaxes robust obedience constraints, and unlocks further experiments. For finite model classes and compact mechanism classes with exact diagnostic and oracle access, sublinear regret with all-round obedience at vanishing failure probability is possible exactly when each terminal closure admits a common optimal obedient mechanism. These results characterize how recommendation visibility and model uncertainty jointly constrain access to experiments on controlled dynamics.

1 INTRODUCTION

Learning a predictive model of controlled dynamics requires visiting informative transitions (Kim et al., 2020; Sekar et al., 2020; Tarbouriech et al., 2020; Gu et al., 2026). When strategic receivers control actions, the learner must induce those visits through recommendations. A physically informative action may then be unavailable because a receiver prefers to deviate. We study how recommendation visibility and uncertainty about the dynamics determine which experiments can be implemented.

Incentivized exploration already studies explorable actions and unknown MDPs (Mansour et al., 2022; Simchowitz & Slivkins, 2024), and selective historical disclosure can change learning rates (Immorlica et al., 2020). Our first question isolates a different intervention: can changing who sees whose current recommendation alter learnability when the dynamics, learner feedback, and known-model worst-equilibrium values remain fixed? Matching these values separates access to informative experiments from the decision power available when the model is known.

Our second question asks how access evolves during learning. Certifying obedience requires satisfying every retained model's incentive constraints, so a candidate is both a hypothesis to test and a constraint on available tests. Eliminating it can unlock an experiment that distinguishes other models. We establish the visibility separation against adverse sequential-equilibrium selection, then characterize evolving access under prescribed obedience. These two levels distinguish guaranteed execution in the separation instance from learning with a given obedience protocol.

Our main contributions are as follows.

Matched-oracle visibility separation. In a transfer-free two-stage system, private signals uniquely implement actions that reach and activate an informative transition; public signaling admits a common uninformative sequential-equilibrium selection. With identical dynamics, utilities, sender objective, learner feedback, and worst-equilibrium oracle values, minimax regret is linear publicly

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.

and logarithmic privately for fixed alternatives (Theorem 1). The separation also holds when the unknown transition affects receiver continuation payoffs (Theorem 2).

Decision-sufficient obedient closure. Under prescribed obedience, the *obedient observability closure* tracks elimination through currently accessible experiments. For finite model classes and compact mechanism classes with exact diagnostic and oracle access, sublinear regret with all-round obedience at vanishing failure probability is possible exactly when every terminal closure admits a common optimal obedient mechanism (Theorem 7). A common optimum permits stopping before full identification or closure stabilization.

When elimination unlocks new information. In the convex affine setting, radial slack makes closure stabilize after one step. Genuine later unlocking therefore requires a binding candidate-model incentive constraint that blocks a future informative experiment (Proposition 4 and Corollary 5). A three-model construction realizes this obstruction over the full convex direct class, even with strict obedience on every used recommendation.

2 Related Work

Persuasion and Bayes correlated equilibrium formalize commitment and obedience (Kamenica & Gentzkow, 2011; Bergemann & Morris, 2016), including private histories and sequential deviations (Makris & Renou, 2023). Unknown dynamics appear in sequential persuasion (Wu et al., 2022; Bernasconi et al., 2022; Zu et al., 2025; Bacchiocchi et al., 2025) and far-sighted principal–agent learning (Gan et al., 2025). Markov Information Processes combines dynamic obedience with receivers' transition learning under persistent excitation (Sezer, 2026). Our question concerns the learner's access to informative transitions: whether visibility and model uncertainty permit their obedient execution.

Bayesian exploration recursively discovers explorable actions (Mansour et al., 2022; Immorlica et al., 2019); incentivized RL explores unknown MDPs through agents' episode-policy choices (Simchowitz & Slivkins, 2024). Related work studies prior-agnostic incentives (Ramalingam et al., 2026), exploration episodes (Banihashem et al., 2026), and principal–agent regret (Liu & Ratliff, 2025; Lin & Chen, 2025). Safe exploration and identification likewise expand robustly admissible experiments as uncertainty shrinks (Sui et al., 2015; Amani et al., 2019; Ahmadi et al., 2021), including controlled dynamics (**?**). Our closure specializes this recursive access structure to model-wise dynamic obedience, characterizes the incentive boundaries needed for new identifying directions, and stops at a common optimum rather than requiring full identification.

Selective historical disclosure changes learning rates for sequential myopic agents (Immorlica et al., 2020). External information and the choice between incentive-compatible advice and arbitrary undominated behavior can also separate linear from sublinear regret (Chakarov et al., 2026). Our separation holds the worst-sequential-equilibrium response criterion fixed and varies current recommendation visibility among simultaneous, far-sighted receivers. Matching learner telemetry and worst-equilibrium oracle values isolates this intervention from changes in feedback or the benchmark.

Multi-channel persuasion orders implementable posterior laws (Babichenko et al., 2021); related models study externalities, interacting receivers, and multi-agent RL (Shaki et al., 2025; Zeng, 2026; Lin et al., 2023). This implementation advantage need not guarantee informative execution under every equilibrium. Perturbations support sender-favored selection (Lipnowski et al., 2024), transfers expand implementation (Liu et al., 2026), and strategic click-bandits bound regret across Nash equilibria (Kleine Buening et al., 2024). Our transfer-free construction combines a selector-independent private diagnostic with a common uninformative public selection while preserving oracle values.

Active model estimation and experimental design optimize transition visits or policy-induced measurements (Tarbouriech et al., 2020; Mutny et al., 2023; Gu et al., 2026; Blanke & Lelarge, 2023). We restrict those experiments by strategic implementability (Equation (2)). Partial monitoring and DEC supply information–decision frameworks (Lattimore & Szepesvári, 2019; Kirschner et al., 2020; Foster et al., 2021; 2023b;a); structured Stackelberg learning defines learnability dimensions (Balcan et al., 2026). Controlled testing supplies the diagnostic rates once obedient experiments are available (Nitinawarat et al., 2013; Deshmukh et al., 2021).

3 Model: Recommendations, Actions, and Feedback

A learner recommends actions to n strategic receivers, who control execution. The receivers know the fixed model $\theta^\star \in \Theta$; only the learner is uncertain which member of the finite class Θ is true. Interaction consists of T reset episodes of H stages, with fresh physical randomness and episode-local utilities. These assumptions separate learning across episodes from strategic continuation within an episode. Receiver i chooses $A_{i,h}$, and the joint action $A_h = (A_{i,h})_{i=1}^n$ determines the next state and physical observation:

$$(S_{h+1}, O_{h+1}) \sim P_{\theta^\star}(\cdot \mid S_h, A_h).$$

The learner's recorded feedback is specified below. At episode t, it uses past observations to select $\pi_t \in \mathcal{P}_{\mathsf{A}}$, publicly announces the mechanism, and commits to it. Architecture $\mathsf{A} \in \{\mathrm{pub}, \mathrm{priv}\}$ specifies visibility; the mechanism specifies signaling kernels, auxiliary messages M_h, recommendations R_h, and causal continuations after every action history, including deviations. Actions are finite; state, observation, and message spaces are standard Borel; receivers have perfect recall. A motivating idealization is a device fleet whose operators know a shared calibrated response law while a new coordinator learns it from telemetry. Heterogeneous private models or learning receivers require an enlarged information state.

Public messages and recommendation rules are seen by all receivers; private receivers see only their own, possibly correlated channels. The private class includes direct recommendations and replicated-public messages with common knowledge of replication, permitting public simulation. Only visibility changes across architectures. Receivers act but cannot report the parameter. Reports would enlarge both the feedback channel and the strategic game: extending our lower bound would require an uninformative equilibrium for every admissible report-dependent mechanism, not merely the possibility of babbling.

To decide whether to follow, a receiver compares the remaining payoff under obedience with that under deviation, using its own information. Write its pre-action history, including $R_{i,h}$, as $I_{i,h}^{\mathsf{A}}$, with $\mathcal{F}_{i,h}^{\mathsf{A}} = \sigma(I_{i,h}^{\mathsf{A}})$. Standard Borelness ensures regular conditional values. For stage utility $u_{i,h}(S_h, A_h, O_{h+1})$, episode utility is $U_i = \sum_{h=1}^H \delta^{h-1} u_{i,h}(S_h, A_h, O_{h+1})$. A deviation changes future states and recommendations, so the comparison includes its best causal continuation. Write $\mathbb{E}_\theta^{\pi, a_i', d_i^+}$ for the law when receiver i takes a_i', then deviates according to d_i^+ using its future private information, while others obey. Its remaining value is

$$Q_{i,h}^{\pi,\theta}(I_{i,h}^{\mathsf{A}}, a_i'; d_i^+) = \mathbb{E}_\theta^{\pi, a_i', d_i^+}\left[\sum_{\ell=h}^H \delta^{\ell-h} u_{i,\ell} \,\middle|\, \mathcal{F}_{i,h}^{\mathsf{A}}\right].$$

Utilities are measurable and stagewise integrable under every model and mechanism, under obedience and unilateral causal deviations, with finite conditional deviation-value essential suprema almost surely. Following throughout gives $V_{i,h}^{\pi,\theta}(I_{i,h}^{\mathsf{A}}) = Q_{i,h}^{\pi,\theta}(I_{i,h}^{\mathsf{A}}, R_{i,h}; \mathrm{obey})$. Dynamic γ-obedience requires this value to exceed every deviation by at least γ:

$$V_{i,h}^{\pi,\theta}(I_{i,h}^{\mathsf{A}}) \geq \operatorname*{ess\,sup}_{a_i' \neq R_{i,h},\, d_i^+} Q_{i,h}^{\pi,\theta}(I_{i,h}^{\mathsf{A}}, a_i'; d_i^+) + \gamma.$$

This holds $\mathbb{P}_\theta^\pi$-almost surely on the obedient path for every receiver and stage; null histories impose no constraint. Receivers know θ, so the supremum is model-wise. It rules out causal deviations starting on path, without imposing off-path sequential rationality. The margin applies to pure first deviations, not arbitrarily small mixtures with obedience.

The learner cannot check obedience only at the unknown truth. It must satisfy every model in its retained set $C \subseteq \Theta$, giving the model-wise and robust mechanism sets

$$\Pi_{\mathsf{A}}^\gamma(\theta) = \{\pi \in \mathcal{P}_{\mathsf{A}} : \pi \text{ is dynamically } \gamma\text{-obedient under } \theta\},$$
$$\Pi_{\mathsf{A}}^\gamma(C) = \bigcap_{\theta \in C} \Pi_{\mathsf{A}}^\gamma(\theta).$$

Identification uses the learner's log, not the entire physical trajectory. The log contains messages, recommendations, actions, and declared state or telemetry feedback F:

$$Z = G_{\mathsf{A}}(X) = ((M, R)_{1:H}, A_{1:H}, F_{1:H}),$$

a fixed projection of $X = (S_{1:H+1}, O_{1:H+1}, M_{1:H}, R_{1:H}, A_{1:H})$. The law $K_\theta^\pi = \mathcal{L}_\theta^{\pi,\mathrm{obey}}(Z)$ describes this log under following, even when π is nonobedient. Holding the response rule fixed lets us compare laws across mechanisms in the affine analysis.

Statistical analysis also needs a law when a learning error selects a nonobedient mechanism. A fixed measurable response completion defines $\widetilde{K}_\theta^\pi$, equal to K_θ^π whenever $\pi \in \Pi_{\mathsf{A}}^\gamma(\theta)$. Resets keep these episode kernels fixed while learning changes mechanism choice. Geometry uses the prescribed law K; the selection-robust separation instead allows every sequential-equilibrium law in $\mathcal{Q}_\theta^\pi$.

Unlisted rewards are not observed, so identical feedback laws can coexist with different sender values. Recording the sender payoff would rule this out for a fixed mechanism and response rule. For bounded expected payoff $J_\theta(\pi)$, assume $\Pi_{\mathsf{A}}^\gamma(\theta) \neq \varnothing$ and set $J_\theta^\star = \sup_{\pi \in \Pi_{\mathsf{A}}^\gamma(\theta)} J_\theta(\pi)$. Prescribed pseudo-regret is $R_T(\theta) = \sum_{t=1}^T [J_\theta^\star - J_\theta(\pi_t)]$. The compactness assumptions below ensure oracle attainment. Payoffs remain bounded on safety-failure events, and regret is not observed. Section 4 allows arbitrary sequential-equilibrium selection; Sections 5–6 prescribe obedience. Table 1 distinguishes their regret criteria from the guarantee criterion in Appendix E.1.

4 Architecture-Induced Dynamic Separation

The construction must make informative actions unavoidable under private recommendations while leaving an uninformative equilibrium under public recommendations, without raising the private known-model value. We first achieve this when the unknown transition affects information but not receiver payoffs, then extend it to payoff-relevant dynamics.

Two successive disagreements are needed to expose the unknown transition: the first can reach an informative state, and the second activates its model-dependent row. Two receivers choose binary actions at $h \in \{1, 2\}$; write disagreement as $D_h = \mathbf{1}\{A_{1,h} \neq A_{2,h}\}$. The parameter $\theta \in \{-\Delta, +\Delta\}$, with $0 < \Delta < 1/2$, determines the transition shift. With gate probability $q \in (0, 1]$, the state evolves as

$$\begin{aligned} S_2 \mid D_1 &\sim \mathrm{Bernoulli}(qD_1), \\ S_3 \mid S_2, D_2 &\sim \mathrm{Bernoulli}\big(\tfrac{1}{2} + \theta S_2 D_2\big)\,. \end{aligned} \tag{1}$$

The sender prefers opposite matched terminal actions under the two models. Encode this preference by $x(11) = 1$, $x(00) = -1$, and $x(01) = x(10) = 0$, giving the unobserved criterion $J_\theta(\pi) = \theta \mathbb{E}_{\theta,\pi}[x(A_2)]$. The learner logs recommendations, actions, and states, but not the payoff. Coordination utility alone permits a message-ignoring equilibrium (Appendix E.2); the next construction removes this obstruction to private diagnosis.

To make private diagnosis work under every equilibrium selection, recommendations convey payoff information that receivers cannot ignore. Keep (1) and introduce a fresh target $W_h \sim \mathrm{Bernoulli}(1/2)$, independent across stages and of θ and the physical history. The learner observes and logs the target before signaling; receivers learn it only through their signals. They value both matching the target and coordinating with each other:

$$u_{i,h} = \mathbf{1}\{A_{i,h} = W_h\} + \kappa \mathbf{1}\{A_{i,h} = A_{j,h}\}.$$

Actions are simultaneous, with no pre-action communication. Under both architectures, receivers publicly observe each completed stage's actions and physical state; logged W_h belongs to F_h. Mechanisms are causal, type-blind action-label kernels without reports or receiver-evaluated model menus. Signals are finite and the discount is one. Appendix A.1 establishes sequential selectability of every dynamically γ-obedient prescribed law in this instance.

Figure 1 contrasts a common uninformative public selection with a uniquely implemented private diagnostic. In the public row, disagreement is nonobedient; the closed gate means no model-dependent transition information, not that S_3 is unreachable.

Theorem 1 (Transfer-free selection-robust separation)**.** *Fix $0 < \kappa < \gamma/2 < 1/2$, $0 < \Delta < 1/2$, and $q \in (0, 1]$. Put*

$$\mu_\gamma = \gamma - \kappa, \qquad b_\star = \frac{1 - \mu_\gamma}{1 + \mu_\gamma}, \qquad c_\kappa = \frac{2\kappa}{1 + \mu_\gamma}.$$

Let $\mathcal{Q}_\theta^\pi$ collect the laws induced by sequential equilibria of π, let $j_\theta(\nu) = \theta \mathbb{E}_\nu x(A_2)$, and define $V_{\theta,\mathsf{A}}^{\mathrm{rob}} = \sup_{\pi \in \Pi_{\mathsf{A}}^\gamma(\theta)} \inf_{\nu \in \mathcal{Q}_\theta^\pi} j_\theta(\nu)$. Under both architectures, $V_{\theta,\mathsf{A}}^{\mathrm{rob}} = \Delta b_\star$, equal to the prescribed

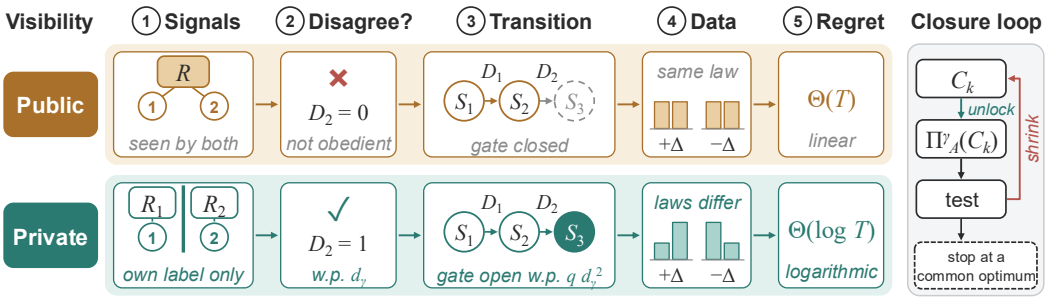

Figure 1: Visibility changes informative access under matched oracle values (left). Under prescribed obedience, evidence can unlock experiments until a common optimum is certified (right).

and unique-action-outcome oracle values. Regret is $R_T = \sum_t [V^{\mathrm{rob}}_{\theta,\mathsf{A}} - \theta \mathbb{E}_{\theta,\pi_t,\sigma_t} x(A_{2,t})]_+$. *If an adaptive randomized selector may choose any sequential equilibrium* σ_t *after each mechanism, then, among uniformly and deterministically dynamically* γ*-obedient algorithms,*

$$\inf_{\mathfrak{A}_{\mathrm{pub}}} \sup_{\theta,\sigma} \mathbb{E}_{\theta,\sigma} R_T = \Delta b_\star T,$$
$$\inf_{\mathfrak{A}_{\mathrm{priv}}} \sup_{\theta,\sigma} \mathbb{E}_{\theta,\sigma} R_T = \Theta(\log T),$$

where $\theta \in \{-\Delta, +\Delta\}$ *and* σ *ranges over selectors. The private upper bound uses unique-action-outcome mechanisms. Its diagnostic has stagewise disagreement probability* $d_\gamma = (1-\gamma^2)/2$ *and squared Hellinger separation* $q d_\gamma^2 (1 - \sqrt{1 - 4\Delta^2})$.

Public visibility leaves an uninformative equilibrium available to the selector. At a terminal public history, mismatch makes the two follow-minus-switch gaps sum to $-2\kappa < 2\gamma$, so the prescribed path has $D_2 = 0$. The completion lemma makes this path sequentially selectable even when other equilibria are informative.

Private labels can disagree while each receiver strictly prefers its own recommendation. Knowing the dynamics does not reveal the fresh target W_h. Conditionally independent binary-symmetric labels of accuracy $(1+\gamma)/2$ give prescribed gap $\gamma + \kappa\gamma^2$ and gap at least $\gamma - \kappa > 0$ against any opponent action. Independent, history-independent continuation channels permit backward induction. Every sequential selector therefore implements the diagnostic, opening the informative gate with probability $q d_\gamma^2$.

Learning the correct oracle requires costly informative actions. Aggregating terminal obedience over all private signals gives the oracle envelope

$$\pm \mathbb{E}[x(A_2)] + c_\kappa \Pr(D_2 = 1) \le b_\star.$$

Uniquely implemented matched labels attain the bias bound in both architectures. Disagreement reduces attainable bias, charging each informative visit a positive regret cost. The KL chain rule and change of measure then force logarithmic loss to distinguish oracle choices; diagnostic repetition followed by a likelihood test attains this rate (Section A.1). At $\kappa = 0$, this opportunity cost vanishes and the private rate need not be logarithmic.

4.1 Payoff-Relevant Unknown Dynamics

The same access gap persists when learning concerns a transition that receivers value. Simply adding εS_3 to utility can also improve the private oracle (Proposition 12). The following one-sided transition condition preserves matched values while making obedience depend on the unknown dynamics.

Theorem 2 (Payoff-relevant matched-oracle separation)**.** *Replace the terminal transition in* (1) *by*

$$S_3 \mid S_2, D_2 \sim \mathrm{Bernoulli}\big(\tfrac{1}{2} + (\beta + \theta) S_2 D_2\big), \qquad \theta \in \{-\Delta, \Delta\},$$

and add terminal utility εS_3 *to both receivers. Assume* $0 < \Delta \le \beta$, $\beta + \Delta < 1/2$, *and* $0 < \varepsilon(\beta + \Delta) < \kappa < \gamma/2 < 1/2$. *Keep the sender payoff, mechanism class, and feedback convention*

unchanged. Both architectures have prescribed, robust, and unique-action-outcome oracle value $\Delta b_\star$. Under the deterministic all-round obedience and adaptive sequential-selection criteria of Theorem 1, public minimax regret is $\Delta b_\star T$ and private minimax regret is $\Theta(\log T)$. The same private diagnostic remains dynamically γ-obedient with unique action outcomes.

The one-sided condition makes reaching the informative state weaken coordination. The effective coefficient is $\kappa_s^\theta = \kappa - \varepsilon(\beta+\theta)s \in (0, \kappa]$, and optimal sender bias increases with it. Both architectures attain the common benchmark at $S_2 = 0$, while private labels activate both diagnostic gates uniquely. Appendix A.2 derives the full-deviation gaps, oracle envelope, and regret bounds.

5 Obedient Experiment Access and Closure

The separation establishes that visibility can change access even under adverse equilibrium selection. We next ask how access evolves as models are eliminated, under prescribed obedience. The learner's log then has law K_θ^π. To connect this law to controlled dynamics, consider a fully logged finite Markov system with a common initial-state law and type-blind causal mechanisms based on logged histories. Writing $d_{\theta,h}^\pi(s,a) = \Pr_\theta^\pi(S_h = s, A_h = a)$ for a row's visitation probability and $Q_{\theta,h}$ for its next-state/feedback kernel, trajectory information decomposes as

$$D_{\mathrm{KL}}(K_\theta^\pi \| K_{\theta'}^\pi) = \sum_{h,s,a} d_{\theta,h}^\pi(s,a) D_{\mathrm{KL}}(Q_{\theta,h}(\cdot \mid s,a) \| Q_{\theta',h}(\cdot \mid s,a)) \,. \tag{2}$$

This standard decomposition (Tarbouriech et al., 2020; Mutny et al., 2023; Gu et al., 2026) separates access from information per visit. Positive occupancy on a differing row permits model distinction; visitation probability and row divergence jointly determine its KL contribution. Define the *obediently reachable rows* at candidate set C:

$$\mathsf{X}_\mathsf{A}(C,\theta) = \{(h,s,a) : \exists \pi \in \Pi_\mathsf{A}^\gamma(C),\ d_{\theta,h}^\pi(s,a) > 0\}. \tag{3}$$

We use *strategic observability* for observability through currently robustly obedient mechanisms. Closure retains the models that none of these experiments distinguish from the truth, then recomputes access. Using full observable laws rather than physical rows also covers partial feedback.

Definition 3 (Closure operator)**.** For $C \subseteq \Theta$ containing $\theta^\star$, let

$$\mathcal{T}_{\mathsf{A},\theta^\star}(C) = \{\theta \in C : K_\theta^\pi = K_{\theta^\star}^\pi \ \forall \pi \in \Pi_\mathsf{A}^\gamma(C)\}.$$

Starting from $C_0 = \Theta$, set $C_{k+1} = \mathcal{T}_{\mathsf{A},\theta^\star}(C_k)$ and denote the stabilized set by $C_\infty^\mathsf{A}(\theta^\star)$.

The operator is indexed by the truth to describe which distinctions are possible; it does not give the learner access to the truth or to exact laws. Section 6 implements reductions by statistical tests. Certified access depends on membership in C, not posterior weights: an unlikely retained candidate can still block a probe, whereas a confidence set can exclude it without its posterior mass becoming zero.

For the fully logged Markov system, Proposition 31 gives the rowwise form

$$\mathcal{T}_{\mathsf{A},\theta}(C) = \{\theta' \in C : Q_{\theta',h}(\cdot \mid s,a) = Q_{\theta,h}(\cdot \mid s,a)\ \forall (h,s,a) \in \mathsf{X}_\mathsf{A}(C,\theta)\}.$$

Shrinking C can enlarge $\mathsf{X}_\mathsf{A}(C,\theta)$. The operator is monotone and contractive, with at most $|\Theta| - 1$ strict reductions. At a fixed point, surviving models either agree on all ambient experiments or their joint incentive constraints exclude every separating experiment.

Remaining uncertainty need not affect the decision. Write the model-wise oracle set as $\Pi_\mathsf{A}^\star(\theta) = \arg\max_{\pi \in \Pi_\mathsf{A}^\gamma(\theta)} J_\theta(\pi)$. Call C *decision-compatible* if $\bigcap_{\theta \in C} \Pi_\mathsf{A}^\star(\theta) \neq \varnothing$: one mechanism is optimal and obedient throughout C, so retaining the truth suffices to stop learning.

In Theorem 1, the two models require opposite oracle biases. Public closure leaves both unresolved; private closure is a singleton. Equal oracle *values* across architectures thus coexist with decision-incompatible public ambiguity.

A three-model construction requires two reductions (Proposition 16). A receiver chooses a safe action S or probe P; the learner observes noisy telemetry, not payoffs. Candidate ζ makes P strictly dominated, so the only initially robust experiment always recommends S. Its feedback separates

ζ from α, β without resolving their opposing oracle choices. Removing ζ permits a probe on a favorable signal, separating α from β. At either truth, closure follows $\{\alpha, \beta, \zeta\} \to \{\alpha, \beta\} \to \{\theta^\star\}$. The first experiment has *experiment-access value*: it enables the diagnostic needed to resolve the decision.

5.1 The Geometry of Obedient Experiments

To determine whether removing an incentive constraint creates new information, we express feasibility and observable laws in common coordinates. For finite mediator and receiver pre-action histories, messages, and actions, let $\mathfrak{D}_{\mathsf{A}}$ collect deterministic causal direct plans, including deviation continuations. A hidden draw from $p \in \mathcal{P}_{\mathsf{A}} \subseteq \Delta(\mathfrak{D}_{\mathsf{A}})$ pre-samples mechanism randomness. Multiplying each conditional obedience inequality by its recommendation-event probability gives an affine row $g_{\theta j}(p) \geq 0$, with $0 \geq 0$ at null events (Lemma 15). Thus the robust feasible set is $\mathsf{P}^{\gamma}_{\mathsf{A}}(C) = \{p \in \mathcal{P}_{\mathsf{A}} : g_{\theta j}(p) \geq 0 \; \forall \theta \in C, j\}$.

Revealing the sampled plan would change what receivers condition on and hence the obedience constraints. With the plan hidden, $K^p_\theta = \mathsf{B}_{\mathsf{A},\theta} p$ is linear into finite signed measures. Differences between robust mechanisms define the *obedient experimental directions*:

$$\mathsf{L}_{\mathsf{A}}(C) = \operatorname{span}\{p - q : p, q \in \mathsf{P}^{\gamma}_{\mathsf{A}}(C)\}.$$

Checking directions alone can miss a constant nonzero law difference. Agreement under every obedient experiment requires equality at one robust base point and on $\mathsf{L}_{\mathsf{A}}(C)$ (Proposition 14). A new direction adds information only if it escapes the observation kernel.

Can a small probability of probing already reveal every available model difference? Radial mixing also appears in safe linear exploration (Amani et al., 2019). By affinity, agreement at a robust base p^0 and a feasible mixture $(1-\varepsilon)p^0 + \varepsilon q$, $\varepsilon > 0$, implies agreement at q. No identifying direction remains to unlock if every probe permits such mixing. The condition below ensures this by allowing binding rows only when no ambient probe violates them.

Proposition 4 (Interior dilution precludes iterative unlocking)**.** *Suppose $\mathcal{P}_{\mathsf{A}}$ is compact and convex, there are finitely many affine obedience rows, and the observable-law map is affine. Assume some $p^0 \in \mathsf{P}^{\gamma}_{\mathsf{A}}(C)$ satisfies the radial-slack condition*

$$g_{\theta j}(p^0) = 0 \quad \Longrightarrow \quad \min_{q \in \mathcal{P}_{\mathsf{A}}} g_{\theta j}(q) \geq 0 \quad \forall \theta \in C, j. \tag{RS}$$

Then, for every truth $\theta \in C$,

$$\mathcal{T}_{\mathsf{A},\theta}(C) = C \cap \{\theta' : K^q_{\theta'} = K^q_\theta \; \forall q \in \mathcal{P}_{\mathsf{A}}\},$$

and applying $\mathcal{T}_{\mathsf{A},\theta}$ again leaves this set unchanged.

Corollary 5 (Necessary boundary for later unlocking)**.** *Under the same affine assumptions, suppose θ' survives $\mathcal{T}_{\mathsf{A},\theta}(C)$, but an ambient experiment q satisfies $K^q_{\theta'} \neq K^q_\theta$. Then for every currently robust p^0 there is a candidate-model obedience row such that*

$$g_{\vartheta j}(p^0) = 0, \qquad g_{\vartheta j}(q) < 0.$$

Thus every informative future direction is blocked at every current base point by a binding, ambient-violable incentive constraint.

Iterative unlocking of new identifying directions is an incentive-boundary phenomenon in this affine setting. If q becomes robust after elimination, its blocking row belongs to a removed model; the truth's constraints remain enforced. For (RS), it suffices that p^0 lies in the interior of $\mathsf{P}^{\gamma}_{\mathsf{A}}(C)$ relative to $\operatorname{aff} \mathcal{P}_{\mathsf{A}}$; relative interior in the robust set's own affine hull does not suffice.

The three-model construction realizes this obstruction on a support boundary: all used recommendations have strict IC, but ζ prohibits any positive probe probability. Hidden mixing cannot bypass that constraint. Removing ζ releases a new identifying direction over the full convex direct class.

Proposition 18 gives a bounded-error counterpart for thresholded finite-library reductions.

Table 1: Regret criteria. Each sums the displayed loss over episodes. Here $j_\theta(\nu_t)$ is the selected equilibrium's expected payoff, $\underline{J}_\theta(\pi) = \inf_\nu j_\theta(\nu)$ its worst-equilibrium value, and $V^{\mathrm{rob}}_{\theta,\mathsf{A}}$ the corresponding oracle.

Criterion	Episode loss	Used in
Prescribed pseudo-regret	$J^\star_\theta - J_\theta(\pi_t)$	Section 6
Selector-realized regret	$[V^{\mathrm{rob}}_{\theta,\mathsf{A}} - j_\theta(\nu_t)]_+$	Section 4
Guarantee regret	$V^{\mathrm{rob}}_{\theta,\mathsf{A}} - \underline{J}_\theta(\pi_t)$	Appendix E.1

5.2 Architecture Refinement

Private channels reproduce public information through common knowledge of message replication. Let $E : \mathcal{P}_{\mathrm{pub}} \to \mathcal{P}_{\mathrm{priv}}$ preserve recommendations, continuations, and learner feedback, with, for every θ,

$$\pi \in \Pi^\gamma_{\mathrm{pub}}(\theta) \Rightarrow E\pi \in \Pi^\gamma_{\mathrm{priv}}(\theta), \qquad K^{E\pi}_\theta = K^\pi_\theta.$$

Proposition 6 (Private recommendations refine the closure)**.** *Under this obedience- and observation-preserving embedding,*

$$C^{\mathrm{priv}}_\infty(\theta^\star) \subseteq C^{\mathrm{pub}}_\infty(\theta^\star) \quad \textit{for every } \theta^\star.$$

The induction combines public simulation with constraint relaxation (Section B.4). Strict refinement requires a private experiment that separates models at the *terminal* public closure. With a robust public base in the affine setting, this requires a direction escaping the common observation kernel (Proposition 19). More access alone does not order regret, since it may also raise the oracle benchmark. Matching oracle values in Theorem 1 removes this benchmark effect.

6 Decision Learning under Prescribed Obedience

The learner needs enough evidence to certify a common optimum, not necessarily to distinguish every accessible model pair. Statistical elimination must also retain the truth: otherwise it can unlock a mechanism the receivers reject. An algorithm is uniformly all-round (δ_T, γ)-obedient if $\inf_{\theta\in\Theta} \mathbb{P}_\theta\{\pi_t \in \Pi^\gamma_{\mathsf{A}}(\theta)\ \forall t \le T\} \ge 1 - \delta_T$, controlling failure over the entire run.

Theorem 7 (Obedient learnability for compact classes)**.** *Let Θ be finite and $\mathcal{P}_{\mathsf{A}}$ be a compact metric space. Assume every $\Pi^\gamma_{\mathsf{A}}(\theta)$ is nonempty and closed, J_θ is continuous and bounded, and the observable kernels are measurable with evaluable likelihoods. Under exact separation, feasibility, and common-oracle access, there exists a sequence of uniformly all-round (δ_T, γ)-obedient algorithms with $\delta_T \to 0$ and $\mathbb{E}_\theta R_T(\theta) = o(T)$ for every $\theta \in \Theta$ if and only if every $C^{\mathsf{A}}_\infty(\theta)$ is decision-compatible. In the compatible case, a finite witness library preserves every one-step closure and every oracle intersection and yields $O(\log T)$ expected regret for each fixed instance.*

For sufficiency, one robust witness per distinguishable pair and candidate set, plus each available common oracle, gives at most $(m^2+1)2^m$ witnesses, where $m = |\Theta|$ (Lemma 24). This finite library preserves the continuous class's elimination and stopping conditions. The learner tests with currently robust witnesses, removes rejected candidates, and recomputes access. Retaining the truth makes each updated experiment obedient; reaching a common oracle ends exploration.

Conversely, terminal models induce identical stopped history laws under adaptive robust experiments. Adaptivity cannot separate them by choosing different future tests, because those choices depend on histories with the same law. Without a common oracle, compactness gives a positive aggregate payoff gap, forcing linear maximum expected regret (Appendix C.2). Exact witness and oracle access need not be computationally efficient.

The remaining question is the cost of each reduction. Within library $\mathcal{L}$, write $\mathcal{L}_C = \mathcal{L} \cap \Pi^\gamma_{\mathsf{A}}(C)$ for available experiments and $D_C(\theta) = C \setminus \mathcal{T}_{\mathsf{A},\theta}(C)$ for removable candidates. A randomized design ρ must distinguish these pairs without knowing the truth. Using squared Hellinger distance $\mathsf{h}^2(P,Q) = 1 - \int \sqrt{dP\,dQ}$, the *obedient escape modulus* maximizes its weakest pairwise separation:

$$\underline{\chi}_{\mathsf{A}}(C) = \max_{\rho\in\Delta(\mathcal{L}_C)} \min_{\theta\in C,\ q\in D_C(\theta)} \sum_{\pi\in\mathcal{L}_C} \rho(\pi)\mathsf{h}^2(K^\pi_\theta, K^\pi_q).$$

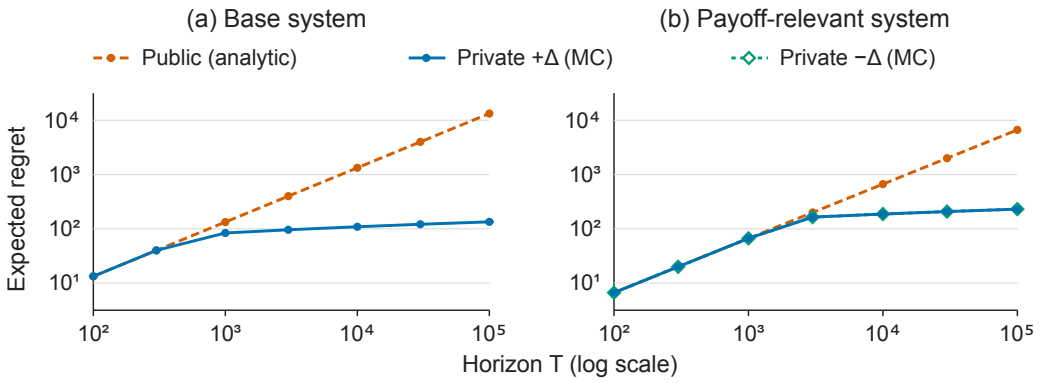

Figure 2: Matched-oracle regret: analytic public curves and private means with pointwise 95% intervals from 10^5 trials per horizon. Base: $+\Delta$; payoff-relevant: both signs.

The value is zero for an empty pair set. Restricting to one truth gives $\chi_{\mathsf{A}}(C,\theta) > 0$ exactly for strict closure steps. Under terminal compatibility, every incompatible C has positive $\underline{\chi}_{\mathsf{A}}(C)$. Hellinger testing retains the truth and removes $D_C(\theta)$ with probability $1-\beta$ in $O(\log(m/\beta)/\underline{\chi}_{\mathsf{A}}(C))$ episodes (Proposition 25). At most $m-1$ successful blocks reach a common oracle; total failure T^{-2} yields logarithmic regret.

7 Computational Validation

The experiments examine the regret separation and the evolution of experiment access separately. For the separation, we simulate diagnosis followed by the likelihood-selected oracle, without observing rewards. The base instance uses $(\gamma,\kappa,q,\Delta) = (0.3,0.1,0.7,0.2)$; the payoff-relevant instance retains (γ,κ,q) with $(\beta,\varepsilon,\Delta) = (0.2,0.2,0.1)$. Each horizon uses 10^5 independent trials with the Hellinger budget and exact informative-count/Bernoulli sampling. Public regret is analytic for evenly randomized oracles.

At $T = 10^5$, private mean regret is 134.53 versus public 13,333.33 in the base system, and 230.73 versus 6,666.67 under either payoff-relevant model (Figure 2). The latter private 95% interval is $[230.7333, 231.2082]$. Binomial Chernoff intervals retain uncertainty even with zero testing errors (Section F.2).

To test evolving access, the three-model experiment compares online elimination with a policy restricted to the initially available safe experiment, using 10^4 trials per model and noise level (Section F.1). With unit Gaussian noise and truth β, by episode 100, 99.95% of runs have unlocked the probe, yet 98.73% remain decision-incompatible: access precedes certification. Median unlock/certification times are 38/191 episodes under both α,β. At $T = 10^4$, adaptive regret is 9.91 under β, versus 2500 with frozen initial access. The probe is already optimal in β; under α, diagnosis costs 39.54 before returning to the safe optimum. Unlocking therefore enables the next decision-relevant test without necessarily resolving which action is optimal.

8 Discussion

The matched-oracle separation shows that recommendation visibility can change learnability through experiment access rather than known-model value. Under prescribed obedience, closure describes how this access evolves and when remaining ambiguity is harmless for decisions. Dynamic access also depends on counterfactual continuations, even when the prescribed trajectory is unchanged (Appendix D.2). Extending the characterization to learning receivers or cross-episode incentives requires tracking their evolving information and continuation values.

REFERENCES

Amir Ali Ahmadi, Abraar Chaudhry, Vikas Sindhwani, and Stephen Tu. Safely learning dynamical systems from short trajectories. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, volume 144 of *Proceedings of Machine Learning Research*, pp. 498–509, 2021. URL https://proceedings.mlr.press/v144/ahmadi21a.html.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. *arXiv preprint arXiv:1908.05814*, 2019. URL https://arxiv.org/abs/1908.05814.

Yakov Babichenko, Inbal Talgam-Cohen, Haifeng Xu, and Konstantin Zabarnyi. Multi-channel bayesian persuasion. *arXiv preprint arXiv:2111.09789*, 2021. URL https://arxiv.org/abs/2111.09789.

Francesco Bacchiocchi, Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Markov persuasion processes: Learning to persuade from scratch. In *Advances in Neural Information Processing Systems*, volume 38, pp. 70123–70160, 2025. doi: 10.52202/085713-2357. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/65798a76cc176c29b6bfefe84b0a03ff-Abstract-Conference.html.

Maria-Florina Balcan, Kiriaki Fragkia, and Keegan Harris. Learning in structured stackelberg games. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/63352.

Kiarash Banihashem, Natalie Collina, and Aleksandrs Slivkins. Bandit social learning with exploration episodes. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.05835v2.

Dirk Bergemann and Stephen Morris. Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Economics*, 11(2):487–522, 2016. doi: 10.3982/TE1808.

Martino Bernasconi, Matteo Castiglioni, Alberto Marchesi, Nicola Gatti, and Francesco Trovò. Sequential information design: Learning to persuade in the dark. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Matthieu Blanke and Marc Lelarge. FLEX: An adaptive exploration algorithm for nonlinear systems. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2577–2591, 2023. URL https://proceedings.mlr.press/v202/blanke23a.html.

Dimitar Chakarov, Lee Cohen, and Nathan Srebro. On incentivized exploration beyond bayesianism and full-information. *arXiv preprint arXiv:2607.18300*, 2026. URL https://arxiv.org/abs/2607.18300.

Krishnendu Chatterjee and Laurent Doyen. Partial-observation stochastic games: How to win when belief fails. In *Proceedings of the 27th Annual ACM/IEEE Symposium on Logic in Computer Science*, pp. 175–184, 2012. doi: 10.1109/LICS.2012.28.

Aditya Deshmukh, Venugopal V. Veeravalli, and Srikrishna Bhashyam. Sequential controlled sensing for composite multihypothesis testing. *Sequential Analysis*, 40(2):259–289, 2021. doi: 10.1080/07474946.2021.1912525.

Dean Foster, Dylan J. Foster, Noah Golowich, and Alexander Rakhlin. On the complexity of multi-agent decision making: From learning in games to partial monitoring. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2678–2792, 2023a. URL https://proceedings.mlr.press/v195/foster23a.html.

Dylan J. Foster, Sham M. Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021. URL https://arxiv.org/abs/2112.13487.

Dylan J. Foster, Noah Golowich, and Yanjun Han. Tight guarantees for interactive decision making with the decision-estimation coefficient. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 3969–4043, 2023b. URL https://proceedings.mlr.press/v195/foster23b.html.

Jiarui Gan, Rupak Majumdar, Debmalya Mandal, and Goran Radanovic. Stochastic principal-agent problems: Computing and learning optimal history-dependent policies. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/e5fefcfcb7288181562b67162c7d5e3a-Abstract-Conference.html.

Xihe Gu, Urbashi Mitra, and Tara Javidi. κ-Explorer: A unified framework for active model estimation in MDPs. *arXiv preprint arXiv:2602.20404*, 2026. URL https://arxiv.org/abs/2602.20404.

Nicole Immorlica, Jieming Mao, Aleksandrs Slivkins, and Zhiwei Steven Wu. Bayesian exploration with heterogeneous agents. *arXiv preprint arXiv:1902.07119*, 2019. URL https://arxiv.org/abs/1902.07119.

Nicole Immorlica, Jieming Mao, Aleksandrs Slivkins, and Zhiwei Steven Wu. Incentivizing exploration with selective data disclosure. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 647–648, 2020. doi: 10.1145/3391403.3399487. URL https://arxiv.org/abs/1811.06026v8. Extended version revised April 2026.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6): 2590–2615, 2011. doi: 10.1257/aer.101.6.2590.

Kuno Kim, Megumi Sano, Julian De Freitas, Nick Haber, and Daniel Yamins. Active world model learning with progress curiosity. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5306–5315, 2020. URL https://proceedings.mlr.press/v119/kim20e.html.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information directed sampling for linear partial monitoring. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2328–2369, 2020.

Thomas Kleine Buening, Aadirupa Saha, Christos Dimitrakakis, and Haifeng Xu. Bandits meet mechanism design to combat clickbait in online recommendation. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2ac879d1865475a7abc8dfc7a9c15c27-Abstract-Conference.html.

Tor Lattimore and Csaba Szepesvári. Cleaning up the neighborhood: A full classification for adversarial partial monitoring. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pp. 529–556, 2019.

Tao Lin and Yiling Chen. Generalized principal-agent problem with a learning agent. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/pdf?id=LqTz13JS2P.

Yue Lin, Wenhao Li, Hongyuan Zha, and Baoxiang Wang. Information design in multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Elliot Lipnowski, Doron Ravid, and Denis Shishkin. Perfect bayesian persuasion. *arXiv preprint arXiv:2402.06765*, 2024. URL https://arxiv.org/abs/2402.06765.

Junyan Liu and Lillian J. Ratliff. Principal-agent bandit games with self-interested and exploratory learning agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39318–39365, 2025. URL https://proceedings.mlr.press/v267/liu25bf.html.

Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. Equilibrium with internal transfers. *arXiv preprint arXiv:2606.20960*, 2026. URL https://arxiv.org/abs/2606.20960.

Miltiadis Makris and Ludovic Renou. Information design in multistage games. *Theoretical Economics*, 18(4):1475–1509, 2023. doi: 10.3982/TE4769.

Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Zhiwei Steven Wu. Bayesian exploration: Incentivizing exploration in bayesian games. *Operations Research*, 70(2):1105–1127, 2022. doi: 10.1287/opre.2021.2205.

Mojmir Mutny, Tadeusz Janik, and Andreas Krause. Active exploration via experiment design in markov chains. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7349–7374, 2023. URL https://proceedings.mlr.press/v206/mutny23a.html.

Sirin Nitinawarat, George K. Atia, and Venugopal V. Veeravalli. Controlled sensing for multihypothesis testing. *IEEE Transactions on Automatic Control*, 58(10):2451–2464, 2013. doi: 10.1109/TAC.2013.2261188.

Ramya Ramalingam, Osbert Bastani, and Aaron Roth. Prior-agnostic incentive-compatible exploration. *arXiv preprint arXiv:2602.20465*, 2026. URL https://arxiv.org/abs/2602.20465.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8583–8592, 2020. URL https://proceedings.mlr.press/v119/sekar20a.html.

Furkan Sezer. Markov information processes. *arXiv preprint arXiv:2607.04308*, 2026. URL https://arxiv.org/abs/2607.04308.

Jonathan Shaki, Jiarui Gan, and Sarit Kraus. Bayesian persuasion with externalities: Exploiting agent types. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14095–14102, 2025. doi: 10.1609/aaai.v39i13.33543. URL https://ojs.aaai.org/index.php/AAAI/article/view/33543.

Max Simchowitz and Aleksandrs Slivkins. Exploration and incentives in reinforcement learning. *Operations Research*, 72(3):983–998, 2024. doi: 10.1287/opre.2022.0495. URL https://doi.org/10.1287/opre.2022.0495.

Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 997–1005, 2015.

Jean Tarbouriech, Shubhanshu Shekhar, Matteo Pirotta, Mohammad Ghavamzadeh, and Alessandro Lazaric. Active model estimation in markov decision processes. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pp. 1019–1028, 2020. URL https://proceedings.mlr.press/v124/tarbouriech20a.html.

Jibang Wu, Zixuan Zhang, Zhe Feng, Zhaoran Wang, Zhuoran Yang, Michael I. Jordan, and Haifeng Xu. Sequential information design: Markov persuasion process and its efficient reinforcement learning. *arXiv preprint arXiv:2202.10678*, 2022. URL https://arxiv.org/abs/2202.10678.

Yishu Zeng. Persuasion of interacting receivers. *Games and Economic Behavior*, 158:563–583, 2026. doi: 10.1016/j.geb.2026.04.009.

You Zu, Krishnamurthy Iyer, and Haifeng Xu. Learning to persuade on the fly: Robustness against ignorance. *Operations Research*, 73(1):194–208, 2025. doi: 10.1287/opre.2021.0529.

Appendix Contents

Notation **14**

A Selection-Robust Dynamic Separation **15**

A.1 Transfer-Free Selection-Robust Separation 15

A.2 Oracle-Preserving Incentive Attenuation 18

A.3 Unknown Dynamics in Receiver Continuation Values 20

A.4 An Exact Boundary for Equal Oracle Values 21

B Obedient Geometry and Information Architecture **22**

B.1 Set-Theoretic Preliminaries 22

B.2 Obedient Experiment Geometry 22

B.3 Stability of a Thresholded Finite-Library Recursion 26

B.4 Architecture Nesting 27

C Obedient Learnability and Diagnostic Complexity **28**

C.1 Closure Characterization 28

C.2 Finite Witness Extraction and Compact-Class Learnability 31

C.3 Finite-Sample Diagnostic Bounds 32

D Controlled-Markov Support and Observability **33**

D.1 Pure Nash versus Correlated-Obedient Support 33

D.2 Sequential Markov Support and Continuation Games 35

D.3 Dynamic Continuation-Punishment Separation 36

D.4 Controlled-Markov Occupancy Coverage 37

E Equilibrium Selection and Alternative Mechanisms **39**

E.1 Adversarial Equilibrium Feedback 39

E.2 State-Gated Dynamic Separation 43

F Computational Verification **45**

F.1 Online Learning with Endogenous Experiment Access 45

F.2 Uncertainty for Diagnostic-Then-Commit Regret 46

AI Use Statement **46**

NOTATION

The table collects notation used across the main results. Symbols specific to individual extensions are defined locally.

Symbol	Meaning
Controlled dynamics and recommendations	
$\theta^\star, \Theta, C;\ m = \lvert\Theta\rvert$	True model, finite model class, candidate set, and model count.
$t, T;\ h, H$	Episode index and learning horizon; stage index and episode horizon.
$\mathsf{A} \in \{\mathrm{pub}, \mathrm{priv}\}$	Public or receiver-private information architecture.
S_h, A_h, O_h	Physical state, joint receiver action, and physical observation.
P_θ	Controlled transition and observation kernel.
$\pi, \mathcal{P}_{\mathsf{A}}$	Committed causal mechanism and its admissible class.
M_h, R_h, F_h	Auxiliary messages, action recommendations, and logged physical feedback.
$I^{\mathsf{A}}_{i,h}, \mathcal{F}^{\mathsf{A}}_{i,h}$	Receiver information history and its generated sigma-field.
$u_{i,h}, U_i, \delta$	Stage utility, episode utility, and receiver discount factor.
$Q^{\pi,\theta}_{i,h}, V^{\pi,\theta}_{i,h}$	Deviation continuation value and obedient continuation value.
$\gamma, \Pi^{\gamma}_{\mathsf{A}}(C)$	Obedience margin and mechanisms obedient under every model in C.
Observation laws, closure, and learning	
Z, G_{A}	Logged episode trajectory and the learner's feedback projection.
$K^{\pi}_{\theta}, \widetilde{K}^{\pi}_{\theta}$	Prescribed-following law and actual law under the fixed response completion.
$\mathcal{T}_{\mathsf{A},\theta}(C)$	Models indistinguishable from θ under every currently robust experiment.
$C_k, C^{\mathsf{A}}_{\infty}(\theta)$	Iterated candidate set and terminal obedient observability closure.
$J_\theta, J^\star_\theta, \Pi^\star_{\mathsf{A}}(\theta)$	Sender payoff, optimal obedient value, and set of optimal obedient mechanisms.
R_T, δ_T	Cumulative regret and all-round obedience failure probability; the regret benchmark is specified in each result.
$\mathcal{Q}^{\pi}_{\theta}, V^{\mathrm{rob}}_{\theta,\mathsf{A}}$	Sequential-equilibrium trajectory laws and worst-equilibrium oracle value.
$\mathsf{P}^{\gamma}_{\mathsf{A}}(C), \mathsf{L}_{\mathsf{A}}(C)$	Robust obedient polytope and its affine direction space.
$d^{\pi}_{\theta,h}, Q_{\theta,h}, \mathsf{X}_{\mathsf{A}}(C,\theta)$	State-action occupancy, next-state/feedback kernel, and obediently reachable rows.
$\mathcal{L}, \mathcal{L}_C, D_C(\theta)$	Finite witness library, its robust subset, and models removed by one closure step.
$\chi_{\mathsf{A}}(C,\theta), \underline{\chi}_{\mathsf{A}}(C)$	Truth-specific and truth-independent obedient escape moduli on the finite library.
α, β, ζ	Model labels in the three-model unlocking construction; ζ blocks the probe.
Two-stage separation	
W_h, D_h	Hidden receiver payoff state and indicator of receiver disagreement.
Δ, q, κ	Model magnitude, first-gate transition probability, and coordination utility coefficient.
$\mu_\gamma = \gamma - \kappa$	Weak-label posterior advantage in the matched oracle construction.
β, ε	Transition offset and terminal-utility weight in the payoff-relevant instance.

A Selection-Robust Dynamic Separation

A.1 Transfer-Free Selection-Robust Separation

The separation must compare the best available mechanisms, not only the diagnostic and exploitation rules used by our learner. Both theorems therefore optimize over all finite causal, type-blind signaling kernels in the stated interface, including auxiliary messages, correlated hidden randomization, and action-history-dependent continuations. Only the constructed learning mechanisms use history-independent channels. Per-episode regret conditions on the learner's preceding history, its mechanism, and the current selector choice; the outer expectation averages these adaptive choices.

Obedience alone does not specify rational play after deviations. To use prescribed following as an adverse sequential-equilibrium selection, we must complete the off-path strategies without overturning on-path incentives. The next lemma does so by bounding the resulting change in a deviator's continuation payoff.

Lemma 8 (Prescribed-path completion)**.** *Fix a model and a finite, causal direct mechanism in the two-stage payoff-state instance. If it is dynamically γ-obedient on its prescribed path and $\delta\kappa < \gamma$, then its prescribed-path law is induced by some sequential equilibrium. In particular, this holds under the parameters of Theorem 1, where $\delta = 1$.*

Proof. We first construct a rational terminal continuation, then check that its effect on first-stage deviations is smaller than the obedience margin. Prune structurally zero-probability chance branches, so the induced extensive game is finite. Take fully mixed first-stage strategies converging to prescribed following. Along a subsequence, Bayes' rule induces limiting beliefs μ at every terminal information set.

At each terminal information set reached with positive probability under prescribed following, dynamic obedience makes the label a strict best response when the other receiver follows. The limiting belief at such a set assigns probability one to prescribed-path nodes, where the other receiver's information set is also on path. Fixing following at all these sets therefore preserves their best responses regardless of play at the remaining sets. At those remaining sets, a mixed best-response fixed point exists in the finite agent-normal-form continuation game, treating the fixed on-path actions as committed. This gives a terminal strategy s_2 that is sequentially rational under μ and follows every on-path label. Adding vanishing full-support trembles to s_2 does not change any pre-action terminal belief, since terminal actions are simultaneous. Combined with the first-stage trembles, it proves consistency of (s_2, μ).

It remains to check a first-stage deviation. Hold fixed any receiver's deviating action and causal continuation rule. Replacing the other receiver's declared terminal following rule by s_2 leaves the pre-action signal law and the deviator's state-matching payoff unchanged; the ensuing physical transition enters no later receiver utility. The replacement changes only $\delta\kappa\mathbf{1}\{A_{i,2} = A_{j,2}\}$, so it can increase the deviation payoff by at most $\delta\kappa$. Taking the supremum over the deviator's continuation rules preserves this bound. On-path dynamic obedience therefore leaves first-stage gap at least $\gamma - \delta\kappa > 0$. Prescribed following at stage one, together with (s_2, μ), is consequently a sequential equilibrium and induces exactly the prescribed-path law. □

Proof of Theorem 1. We first identify a model-independent public selection. At a prescribed-path terminal public information set, both receivers have the same posterior over W_2 and observe the joint label. If the labels mismatch, their two follow-minus-switch gaps sum to -2κ, so they cannot both be at least $\gamma > 0$. Every obedient public mechanism therefore has $D_2 = 0$ on its prescribed path. By Lemma 8, that path is a sequential-equilibrium selection. Since θ enters only through $S_2 D_2$, its complete feedback law is common to $\theta = \pm\Delta$. Applying this selection after every common adaptive history preserves that equality inductively. This constructs one common observable law; other public equilibria may be informative.

For the private diagnostic, conditionally independently pass W_h through a binary symmetric channel of accuracy $p = (1 + \gamma)/2$ to form the two labels. Use independent channels across stages and make the second-stage kernel independent of S_2 and the preceding history. Conditional on its label, a receiver's state-matching advantage from following is $2p - 1 = \gamma$. When the other receiver follows, the match-minus-mismatch probability is $(2p - 1)^2 = \gamma^2$, so the exact prescribed gap is

$\gamma + \kappa\gamma^2 > \gamma$. Against an arbitrary action of the other receiver, the coordination contribution is at least $-\kappa$; following therefore has gap at least $\gamma - \kappa > 0$. It is strictly dominant at stage two. The open-loop continuation removes any continuation effect of a first-stage deviation, so the same argument and backward induction make the action outcome unique at stage one.

The mismatch probability of either channel pair is $d_\gamma = 2p(1-p) = (1-\gamma^2)/2$. Thus the informative gate $G = \{D_1 = 1, S_2 = 1, D_2 = 1\}$ has probability qd_γ^2. Conditional on G, S_3 is Bernoulli with parameter $1/2 + \theta$; off G, the feedback law is common. Decomposing Hellinger affinity over these two events gives

$$H^2_{\mathrm{lat}} = qd_\gamma^2 \left(1 - \sqrt{1 - 4\Delta^2}\right) > 0. \tag{PS1}$$

To compare oracle values and charge for information, we relate terminal action bias to the probability of disagreement. Fix a learner history and an arbitrary obedient private mechanism, including fine auxiliary signals, and evaluate its prescribed action law. For receiver i, record how often it is recommended one and how often that recommendation disagrees with the other label:

$$q_i = \Pr(R_i = 1), \qquad d_i = \Pr(R_i = 1, R_j = 0).$$

Multiply each fine-signal obedience inequality recommending one by that signal's probability and sum. This gives

$$\mathbb{E}[(2W_2 - 1)\mathbf{1}\{R_i = 1\}] + \kappa(q_i - 2d_i) \geq \gamma q_i. \tag{PS2}$$

Let $\ell_i \geq 0$ be the left side of (PS2) minus γq_i. Because W_2 is conditionally fair, $\mathbb{E}[(2W_2 - 1)\mathbf{1}\{R_i = 1\}] \leq 1 - q_i$. With $\mu_\gamma = \gamma - \kappa$, (PS2) therefore implies

$$q_i \leq \frac{1 - 2\kappa d_i}{1 + \mu_\gamma}.$$

Since $x(R_2) = q_1 + q_2 - 1$ in expectation and $d_1 + d_2 = \Pr(D_2 = 1)$, summing over receivers yields

$$\mathbb{E}[x(R_2)] + c_\kappa \Pr(D_2 = 1) \leq b_\star. \tag{PS3}$$

More explicitly, its slack has the nonnegative decomposition

$$b_\star - \mathbb{E}[x(R_2)] - c_\kappa \Pr(D_2 = 1) = \frac{\sum_{i=1}^2 [\ell_i + 2\Pr(W_2 = 1, R_i = 0)]}{1 + \mu_\gamma}.$$

Here ℓ_i is the sum of the unnormalized obedience slacks over all fine information histories recommending one, including retained private histories. This certificate does not bound their number or require independent private signals. Applying the same argument to recommendations of zero gives

$$-\mathbb{E}[x(R_2)] + c_\kappa \Pr(D_2 = 1) \leq b_\star. \tag{PS4}$$

The aggregation proves these inequalities even when the private signal is finer than its action label. It also applies publicly. At each prescribed- path fine terminal signal, obedience and the upper bound κ on the coordination contribution imply that the recommended action's state-posterior advantage is at least $\mu_\gamma = \gamma - \kappa$. Against any opponent action its gap is therefore at least $\mu_\gamma - \kappa = \gamma - 2\kappa > 0$. Thus (PS3)–(PS4) apply to the executed actions under the prescribed-path selection supplied by Lemma 8.

The positive oracle attains (PS3) by sending 11 when $W_2 = 1$, and, when $W_2 = 0$, sending 11 with probability $b_\star$ and 00 otherwise. The negative oracle is its bit complement. At the weak signal, the posterior state advantage is μ_γ; following the matched label has gap $\mu_\gamma + \kappa = \gamma$. Its gap even against the opposite receiver action is $\mu_\gamma - \kappa = \gamma - 2\kappa > 0$, so the terminal action outcome is unique. A fully revealing matched first-stage signal, followed by a second-stage kernel independent of S_2 and the first-stage history, makes the dynamic outcome unique. Both public and private architectures can use these mechanisms.

The benchmark comparison distinguishes prescribed following, the worst sequential selection, and mechanisms whose terminal action distribution is invariant to selection. Let $\nu_\theta^{\pi,\mathrm{obey}}$ denote the prescribed-following law and $\Pi_{\mathsf{A}}^{\mathrm{uao}}(\theta) \subseteq \Pi_{\mathsf{A}}^{\gamma}(\theta)$ the mechanisms for which every law in $\mathcal{Q}_\theta^\pi$ induces

the same distribution of A_2. Their oracle values are

$$V^{\mathrm{pres}}_{\theta,\mathsf{A}} = \sup_{\pi\in\Pi^{\gamma}_{\mathsf{A}}(\theta)} j_\theta(\nu^{\pi,\mathrm{obey}}_\theta),$$
$$V^{\mathrm{rob}}_{\theta,\mathsf{A}} = \sup_{\pi\in\Pi^{\gamma}_{\mathsf{A}}(\theta)} \inf_{\nu\in\mathcal{Q}^{\pi}_\theta} j_\theta(\nu),$$
$$V^{\mathrm{uao}}_{\theta,\mathsf{A}} = \sup_{\pi\in\Pi^{\mathrm{uao}}_{\mathsf{A}}(\theta)} j_\theta(\nu), \qquad \nu\in\mathcal{Q}^{\pi}_\theta.$$

The last value is well-defined because j_θ depends only on A_2. The completion lemma and the restriction to unique action outcomes give

$$V^{\mathrm{uao}}_{\theta,\mathsf{A}} \le V^{\mathrm{rob}}_{\theta,\mathsf{A}} \le V^{\mathrm{pres}}_{\theta,\mathsf{A}}.$$

Equations (PS3)–(PS4) bound the last term by $\Delta b_\star$, while the constructed unique oracle attains this value. Hence all three benchmarks equal $\Delta b_\star$ under both architectures.

For the public regret lower bound, let the selector use the common law just constructed after every adaptively chosen public mechanism. The learner's history then has the same law under both models. If v_t is the common conditional expectation of $x(A_{2,t})$, the two one-period regrets sum to $2\Delta b_\star$. Hence the larger total regret is at least $\Delta b_\star T$. Independently randomizing between the two unique oracles attains this value under every selector, proving the public equality.

Privately, repeat the unique diagnostic for

$$N = \left\lceil H_{\mathrm{lat}}^{-2}\log(2T)\right\rceil \wedge T$$

episodes, apply the likelihood-ratio test, and then use the selected unique oracle. Product Hellinger affinity bounds the sum of testing errors by $(1-H^2_{\mathrm{lat}})^N \le e^{-NH^2_{\mathrm{lat}}}$. The diagnostic has zero sender value, so this gives $O(\log T)$ regret uniformly over adaptive selectors.

For the converse, distinguishing the two oracle choices requires informative gate visits, each incurring the opportunity cost in (PS3). We quantify this tradeoff under the prescribed-path selection, which is admissible by Lemma 8. Let R be the larger expected regret under the two models. Conditional versions of (PS3) and the adaptive KL chain rule bound the accumulated information by this regret:

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{P}^{H_T}_{+}\|\mathbb{P}^{H_T}_{-}) &= I_\Delta \mathbb{E}_+ \sum_{t=1}^{T} S_{2,t}D_{2,t} \\ &\le I_\Delta \mathbb{E}_+ \sum_{t=1}^{T} D_{2,t} \le \frac{I_\Delta}{\Delta c_\kappa} R, \\ I_\Delta &= 2\Delta \log\frac{1/2+\Delta}{1/2-\Delta}. \end{aligned} \tag{PS5}$$

Here H_T includes the learner's mechanism choices and all logged episode observations. For each fixed causal, type-blind mechanism, every observation before S_3 has a common prescribed law across models. Only $S_2D_2 = 1$ contributes conditional KL; no independence between the gates is assumed. The same factorization makes v_t, the prescribed expectation of $x(A_{2,t})$ conditional on the learner history and chosen mechanism, the same predictable function under both models. Set $E = \{\sum_t v_t \ge 0\}$. Oracle regret accounting gives

$$\Pr_{+}(E^c) + \Pr_{-}(E) \le \frac{2R}{\Delta b_\star T}. \tag{PS6}$$

Bretagnolle–Huber with (PS5)–(PS6) yields

$$\frac{2R}{\Delta b_\star T} \ge \frac{1}{2}\exp\left(-\frac{I_\Delta}{\Delta c_\kappa}R\right).$$

If $R \ge \sqrt{T}$, the logarithmic claim is immediate; otherwise taking logarithms gives $R = \Omega(\log T)$ for fixed $(\Delta, q, \gamma, \kappa)$. This completes the private lower bound. □

Remark 9 (The private lower bound is not selector-wise). Take $q = 1$, $\kappa = 0.1$, and $\gamma = 0.3$. At stage one always send 00 and reveal no information about W_1. If $S_2 = 0$, reveal W_2 by the matched label W_2W_2; if $S_2 = 1$, send independent binary-symmetric labels of accuracy $p = (1 + \gamma)/2 = 0.65$. The two prescribed continuation values are

$$V_0 = 1 + \kappa = 1.1, \qquad V_1 = p + \kappa[p^2 + (1 - p)^2] = 0.7045.$$

The binary-symmetric branch has terminal obedience gap $\gamma+\kappa\gamma^2$, and prescribed stage-one following has gap $\kappa + V_0 - V_1 = 0.4955 > \gamma$. The mechanism is therefore dynamically γ-obedient. Yet 00 and 11 are both strict first-stage equilibrium outcomes: a unilateral deviation loses coordination and changes the continuation from V_0 to V_1.

A model-aware selector can choose 00 under $+\Delta$ and 11 under $-\Delta$. Executed first-stage actions then reveal the sign in one episode; the matched fair terminal action has zero expected sender payoff. The learner can subsequently use the correct unique oracle and incur only $O(1)$ regret. Thus the lower bound in Theorem 1 uses the worst prescribed-path selector, whereas its private upper bound is uniform over selectors.

Remark 10 (The optimistic equilibrium-law oracle can be larger). The robust qualifier cannot be dropped. Take $\kappa = 0.1$, $\gamma = 0.22$, and $\mu_\gamma = \gamma - \kappa = 0.12$; more generally, the construction below works for $2\kappa < \gamma < 5\kappa/2$. At stage one, send a matched public binary label whose posterior advantage toward the label is μ_γ. If both receivers follow, run the sign-matched unique oracle at stage two. If exactly one receiver flips its first-stage label, stage two is the no-information coordination game and the continuation equilibrium has both receivers mix uniformly. If both flip, use the same game and a pure continuation equilibrium.

The prescribed first-stage gap is at least γ. Indeed, the oracle continuation gives each receiver $1 - b_\star/2 + \kappa > 1/2 + \kappa$, the largest payoff after a unilateral first-stage flip when the other receiver obeys. There is also a sequential equilibrium in which both receivers flip. Relative to returning unilaterally to its label, a receiver loses $\mu_\gamma - \kappa = \gamma - 2\kappa$ immediately and gains $\kappa/2$ in continuation payoff. The net gain $5\kappa/2 - \gamma$ is positive. The selector can choose terminal action 11 under $\theta = +\Delta$ and 00 under $\theta = -\Delta$. Consequently

$$\sup_{\pi \in \Pi_{\mathrm{A}}^{\gamma}(\theta)} \sup_{\nu \in \mathcal{Q}_\theta^\pi} j_\theta(\nu) = \Delta > \Delta b_\star$$

for both architectures in this parameter subregion. The physical bound $|x| \le 1$ proves equality. The positive-part regret in Theorem 1 assigns zero loss to this favorable selection and takes the supremum over selectors.

A.2 Oracle-Preserving Incentive Attenuation

When receivers value the terminal state, the unknown transition also changes incentives. The one-sided perturbation in Theorem 2 weakens effective coordination at the informative state. We use this monotonicity to preserve the oracle bound while retaining the private diagnostic.

Proof of Theorem 2. Separate the terminal payoff into an action-independent term, target matching, and an effective coordination reward. Write $\mu_\theta = \beta + \theta \ge 0$, $k_s = \kappa - \varepsilon\mu_\theta s$, and $k_{\min} = \kappa - \varepsilon(\beta + \Delta) > 0$. Conditionally on $S_2 = s$, the expected receiver terminal payoff is

$$\varepsilon/2 + \varepsilon\mu_\theta s + \mathbf{1}\{A_i = W_2\} + k_s\mathbf{1}\{A_i = A_j\}. \tag{PR1}$$

The first two terms are action-independent at that history. Thus terminal incentive inequalities are those of the payoff-state game with coordination coefficient k_s. Replacing the opponent's off-path terminal strategy can change the deviator's expected continuation value by at most $k_s \le \kappa$. The tremble construction of Lemma 8 consequently applies without change, since $\gamma > \kappa$. Every obedient prescribed law is sequentially selectable.

For public signals, the terminal mismatch gaps sum to $-2k_s < 0$, so the prescribed path has $D_2 = 0$. Its entire feedback law is common to both models. This common selection can be made after every adaptive history. Privately, the binary-symmetric diagnostic has terminal gap $\gamma + k_s\gamma^2$, and following against any opponent strategy has gap at least $\gamma - \kappa > 0$. Its stage-two target and coordination payoffs are independent of S_2 under following; the state-dependent continuation term is $\varepsilon\mu_\theta S_2 d_\gamma$.

Exactly as in Proposition 11, the first-stage full-deviation gap is $\gamma + (\kappa - \varepsilon\mu_\theta q d_\gamma)\gamma^2 \geq \gamma$. Against any opponent first-stage strategy, it is at least $\gamma - \kappa - \varepsilon\mu_\theta q d_\gamma > 0$. Backward induction gives unique diagnostic action outcomes.

We next establish a common oracle and a positive cost of information. For $k \in (0, \kappa]$, put

$$b(k) = \frac{1-\gamma+k}{1+\gamma-k}, \qquad c(k) = \frac{2k}{1+\gamma-k}.$$

Fix a prescribed-path public history before the terminal signal, with state s. The fresh target W_2 is conditionally fair. For either receiver let $q_i = \Pr(R_i = 1)$ and $d_i = \Pr(R_i = 1, R_j = 0)$, conditionally on that history. Summing the fine-signal terminal IC rows recommending one gives

$$\mathbb{E}[(2W_2 - 1)\mathbf{1}\{R_i = 1\}] + k_s(q_i - 2d_i) \geq \gamma q_i.$$

Conditional fairness gives $\mathbb{E}[(2W_2 - 1)\mathbf{1}\{R_i = 1\}] \leq 1 - q_i$. Summing over receivers and using $\mathbb{E}x(A_2) = q_1 + q_2 - 1$, then repeating for labels zero, yields

$$\pm\mathbb{E}[x(A_2) \mid \text{history}] + c(k_s)\Pr(D_2 = 1 \mid \text{history}) \leq b(k_s). \tag{PR2}$$

This aggregation allows arbitrary private histories and fine auxiliary messages. It bounds both architecture classes. Both $b(k)$ and $c(k)$ are increasing. Integrating (PR2) gives

$$\pm\mathbb{E}x(A_2) + c(k_{\min})\Pr(D_2 = 1) \leq b(\kappa) = b_\star. \tag{PR3}$$

The original positive and negative biased matched-label oracles attain $\pm\mathbb{E}x(A_2) = b_\star$. At stage one reveal W_1 and recommend W_1W_1, so $S_2 = 0$. Use the original second-stage oracle kernel independently of the preceding history. Its weak-label posterior advantage is $\mu_\gamma = \gamma - \kappa > \kappa$, hence following is strictly dominant at every terminal history, including the off-path state $S_2 = 1$. Following has terminal margin exactly γ on the prescribed state $S_2 = 0$. Even after a first-stage deviation, the unique terminal actions match, giving terminal bonus $\varepsilon/2$ and otherwise the same continuation payoff. First-stage following is therefore uniquely optimal and has margin $1 + \kappa$. The two oracle mechanisms are obedient under both models. They attain $\Delta b_\star$ with unique action outcomes, while sequential completion and (PR3) upper-bound the robust oracle by that value. The prescribed and unique-action-outcome values agree as well.

The common public selection and these two robust unique oracles give the exact public minimax regret $\Delta b_\star T$, by the same two-model sum-of-regrets argument as in Theorem 1. For private learning let

$$p_\pm = \tfrac{1}{2} + \beta \pm \Delta, \qquad h_\beta = q d_\gamma^2 \left[1 - \sqrt{p_+ p_-} - \sqrt{(1-p_+)(1-p_-)}\right] > 0. \tag{PR4}$$

The diagnostic's informative gate has probability $q d_\gamma^2$; conditional on it, S_3 has law $\mathrm{Bernoulli}(p_\pm)$. Off the gate its complete law is common. Hence h_β is its squared Hellinger separation. Repeat the diagnostic $N = \lceil \log(2T)/h_\beta \rceil \wedge T$ times, apply the exact likelihood test, and deploy the corresponding robust unique oracle. The sum of testing errors is at most $(1 - h_\beta)^N$. For $N < T$, regret is at most $\Delta b_\star[N + 2(T - N)e^{-Nh_\beta}] = O(\log T)$; the finitely many truncated horizons do not affect the asymptotic rate. All deployed mechanisms are deterministically obedient under both models.

For the private lower bound, learning which oracle to use requires evidence that can only be obtained at the informative gate. Inequality (PR3) charges each such visit a positive regret cost. We relate these two requirements by change of measure under the prescribed-path sequential selection. Set $I_\beta = \mathrm{KL}(\mathrm{Bernoulli}(p_+) \| \mathrm{Bernoulli}(p_-)) \in (0, \infty)$, the information per gate visit. Before the final state, the fixed-following laws of a type-blind mechanism are model-independent. The adaptive KL chain rule and (PR3) imply

$$\mathrm{KL}(\mathbb{P}_+^{H_T} \| \mathbb{P}_-^{H_T}) = I_\beta\, \mathbb{E}_+ \sum_{t=1}^{T} S_{2,t} D_{2,t} \leq \frac{I_\beta}{\Delta c(k_{\min})}\, R, \tag{PR5}$$

where $R = \max_{\theta = \pm\Delta} \mathbb{E}_\theta R_T$. Let v_t be the predictable fixed-following expectation of $x(A_2)$; it is the same function of learner history under both models. For $E = \{\sum_t v_t \geq 0\}$, regret accounting gives $\Pr_+(E^c) + \Pr_-(E) \leq 2R/(\Delta b_\star T)$. Bretagnolle–Huber and (PR5) therefore yield

$$\frac{2R}{\Delta b_\star T} \geq \frac{1}{2} \exp\left\{-\frac{I_\beta R}{\Delta c(k_{\min})}\right\}.$$

This forces $R = \Omega(\log T)$. No unobserved sender payoff or receiver utility is included in the learner's feedback. The public and private comparisons change only receiver-side signal visibility within this fixed physical system. □

A.3 Unknown Dynamics in Receiver Continuation Values

Identification can remain separated even when oracle equality does not. Here we add a terminal-state reward without the one-sided offset. The calculation isolates how unknown dynamics enter both stages' deviation values and establishes identification, rather than a matched-oracle regret comparison.

Proposition 11 (Persistence with payoff-relevant continuation)**.** *In the same two-stage system, add terminal utility εS_3 to each receiver, where $0 < \varepsilon\Delta < \kappa < \gamma/2 < 1/2$. For every public dynamically γ-obedient mechanism, its prescribed law is sequentially selectable and has $D_2 = 0$. These laws are identical under the two models. The private binary-symmetric diagnostic remains dynamically γ-obedient with unique action outcomes and the same positive Hellinger separation as in Theorem 1. Its exact follow-minus-deviate gaps are*

$$\Gamma_{i,2}(s) = \gamma + (\kappa - \varepsilon\theta s)\gamma^2,$$
$$\Gamma_{i,1} = \gamma + (\kappa - \varepsilon\theta q d_\gamma)\gamma^2.$$

Consequently, uniform consistent identification is impossible publicly under worst-case sequential selection and possible privately.

Proof of Proposition 11. Write $a = \varepsilon\Delta < \kappa$, $p = (1+\gamma)/2$, and $d = 2p(1-p) = (1-\gamma^2)/2$. After observing $S_2 = s$, expected terminal utility is $\varepsilon/2 + \varepsilon\theta s D_2$. Up to an action-independent term, the terminal game therefore has coordination coefficient $\kappa_s^\theta = \kappa - \varepsilon\theta s \in [\kappa - a, \kappa + a]$. It is strictly positive. At a common public terminal information set with mismatched labels, the state-matching advantages cancel across receivers and the coordination gaps sum to $-2\kappa_s^\theta < 0$. Both receivers cannot have margin $\gamma > 0$. Hence every prescribed obedient public path has $D_2 = 0$.

The sequential-completion argument in Lemma 8 still applies. At an off-path terminal information set, replacing the other receiver's following rule by an equilibrium continuation leaves the deviator's state-matching payoff unchanged. For any fixed deviating continuation strategy, its expected coordination plus terminal payoff changes by at most $\kappa + a$. This bound is uniform over the deviator's strategies, so it also bounds the increase in the best deviation payoff. Because $\gamma > 2\kappa > \kappa + a$, on-path first-stage following remains strictly optimal. The same finite-game tremble construction yields a sequential equilibrium inducing the prescribed law. For a fixed type-blind public mechanism, that law has identical complete learner feedback under both models: all signals, payoff states and S_2 precede the only model-dependent transition, and $D_2 = 0$ makes S_3 a fair coin. The selector can choose this law after each adaptive history, preserving history indistinguishability.

For the private diagnostic, let each receiver independently receive a binary-symmetric label of W_h with accuracy p. Use independent channels across stages, with the second-stage kernel independent of the earlier history and S_2. Conditional on an own label, the state-matching advantage is γ; under the other receiver's following strategy, the match-minus-mismatch probability is γ^2. Thus the terminal prescribed gap is $\gamma + \kappa_s^\theta\gamma^2 \geq \gamma$. Against any opponent strategy the gap is at least $\gamma - \kappa - a > 0$. The second-stage action outcome is consequently unique, including after any first-stage deviation.

Under this unique continuation, expected second-stage target and coordination payoffs are independent of S_2. The only state-dependent continuation term is $\varepsilon\theta S_2 d$. Conditional on its own label at stage one, following rather than switching changes the probability of $S_2 = 1$ by $-q\gamma^2$. This gives the exact first-stage full-deviation gap

$$\gamma + \kappa\gamma^2 - \varepsilon\theta q d\gamma^2 = \gamma + (\kappa - \varepsilon\theta q d)\gamma^2 \geq \gamma.$$

Here subsequent deviations cannot improve payoffs, by terminal dominance. To establish uniqueness also at stage one, allow the opponent an arbitrary first-stage strategy. The immediate state advantage remains γ, the coordination contribution is at least $-\kappa$, and the continuation effect has magnitude at most aqd. Hence following has gap at least $\gamma - \kappa - aqd > 0$. Backward induction yields unique action outcomes.

The resulting action and transition law is unchanged from the diagnostic in Theorem 1. Its informative event still has probability qd^2, and the conditional observation is $\mathrm{Bernoulli}(1/2 + \theta)$. Therefore its

squared Hellinger distance is $qd^2(1-\sqrt{1-4\Delta^2})>0$. Independent repetitions give a consistent likelihood test under every selector. Under the common public selector, any binary estimator instead has maximum error probability at least $1/2$ for every sample size. This proves the identification separation. No claim about the perturbed sender-oracle equality is used. □

A.4 An Exact Boundary for Equal Oracle Values

The sign of the terminal coupling determines whether visiting the informative state weakens or strengthens coordination. In the parameter region below, this gives an exact boundary for oracle equality: weakening preserves the public benchmark, whereas strengthening admits a strictly better private mechanism.

Proposition 12 (Sign boundary for oracle preservation)**.** *Use the transition and receiver utility of Theorem 2, allowing either sign of β. Suppose $q>0$, $\varepsilon>0$, $0<\Delta$, $|\beta|+\Delta<1/2$, and, with $a_{\max}=\varepsilon(|\beta|+\Delta)$,*

$$0<a_{\max}<\kappa, \qquad \kappa+a_{\max}<\gamma/2<1/2, \qquad \gamma>1/2+a_{\max}.$$

For each $\theta\in\{-\Delta,\Delta\}$, all three public oracle values (prescribed, robust, and unique-action-outcome) equal $\Delta b(\kappa)$. The corresponding private values equal the public value if and only if $\beta+\theta\geq 0$. If $a=-\varepsilon(\beta+\theta)>0$, a private mechanism with unique action outcomes attains a gain over the public oracle of

$$G_\theta=\Delta q d_\gamma[b(\kappa+a)-b(\kappa)]=\frac{\Delta q(1-\gamma^2)a}{(1+\gamma-\kappa-a)(1+\gamma-\kappa)}>0.$$

Consequently, equal architecture-specific oracle values at both models hold exactly when $\beta\geq\Delta$ in this parameter region.

Proof. We first show that public obedience prevents reaching the altered coordination coefficient. Fix a model and put $v=\varepsilon(\beta+\theta)$. The terminal coefficient $k_s=\kappa-vs$ belongs to $[\kappa-a_{\max},\kappa+a_{\max}]\subset(0,\gamma/2)$. Public terminal mismatch gaps sum to $-2k_s<0$, so every public prescribed terminal path has $D_2=0$.

Suppose a positive-probability public first-stage signal recommended a mismatch. The shared posterior over W_1 makes the immediate gaps sum to -2κ. Following gives each receiver continuation payoff at most $1+\kappa+\varepsilon/2$, since prescribed terminal actions match. After switching its first action, the receiver can ignore its terminal signal and choose a fixed binary action. This gives expected target payoff $1/2$, nonnegative coordination payoff, and expected terminal bonus at least $\varepsilon/2+\min\{v,0\}$. The two full-deviation gaps therefore sum to at most $1+2\max\{-v,0\}\leq 1+2a_{\max}<2\gamma$, contradicting obedience. Hence $D_1=0$ and $S_2=0$ on every obedient public prescribed path.

Public play thus retains the original oracle bound. The terminal envelope (PR2) gives prescribed value at most Δb_0, where $b_0=b(\kappa)$. Sequential completion gives the same upper bound for the robust and unique-outcome oracles, since an opponent's terminal response changes continuation payoffs by at most $\kappa+a_{\max}<\gamma$. The original matched oracle attains it: its weak-label posterior advantage $\gamma-\kappa$ exceeds every k_s, so terminal actions match uniquely even after a deviation. Revealing W_1 and recommending W_1W_1 gives first-stage gap $1+\kappa$, with unchanged continuation payoff. This proves all three public equalities. If $v\geq 0$, the statewise bound $b(k_s)\leq b_0$ also holds privately, proving the private equalities, including $v=0$.

Now suppose $v<0$, and write $a=-v$. Use private independent binary-symmetric labels of accuracy $(1+\gamma)/2$ at stage one. At stage two use the matched-label oracle with coefficient $\kappa+as$ and sender bias $\operatorname{sgn}(\theta)b_s$, where $b_s=b(\kappa+as)$. The terminal weak-label state advantage is $\gamma-\kappa-as$, so the action is strictly dominant because $\gamma>2(\kappa+a_{\max})$. Both labels are matched; each receiver's expected terminal payoff is $1+\kappa+\varepsilon/2-b_s/2$. The exact first-stage full-deviation gap is consequently

$$\gamma+\kappa\gamma^2+\frac{q\gamma^2}{2}(b_1-b_0)>\gamma.$$

Against an arbitrary opponent first-stage strategy, the gap is at least $\gamma-\kappa-q(b_1-b_0)/2>0$. Indeed,

$$\frac{b_1-b_0}{2}=\frac{a}{(1+\gamma-\kappa-a)(1+\gamma-\kappa)}<a,$$

and $\gamma > 2(\kappa + a)$. This verifies full dynamic obedience and backward uniqueness, including continuations after deviations. The first-stage disagreement probability is $d_\gamma = (1 - \gamma^2)/2 > 0$, giving sender value $\Delta b_0 + G_\theta$. This mechanism is chosen for the known-model oracle; obedience at the other model is not required for this comparison. The strict gain proves the converse, and applying the sign condition to both models gives $\beta \geq \Delta$. □

For the signed example $\beta = 0$, $(\gamma, \kappa, \Delta, \varepsilon, q) = (4/5, 1/10, 1/5, 1/10, 7/10)$, the negative model has $b_0 = 3/17$, $b_1 = 4/21$, public value $3/85$, and private witness value $303/8500$. More generally, the gain is positive for every negative coupling, however small. The result identifies the boundary of oracle equality without requiring the exact private optimum on the unequal-oracle side.

B Obedient Geometry and Information Architecture

B.1 Set-Theoretic Preliminaries

Eliminating candidates removes incentive constraints and leaves fewer models to distinguish. The following lemma tracks these two effects together. Its greatest-fixed-point property will later rule out further obedient identification inside a terminal set.

Lemma 13 (Monotonicity and stabilization)**.** *If $C \subseteq D$ and both contain $\theta^\star$, then*

$$\Pi^\gamma_{\mathsf{A}}(C) \supseteq \Pi^\gamma_{\mathsf{A}}(D), \qquad \mathcal{T}_{\mathsf{A},\theta^\star}(C) \subseteq \mathcal{T}_{\mathsf{A},\theta^\star}(D).$$

Moreover, $\mathcal{T}_{\mathsf{A},\theta^\star}(C) \subseteq C$, and the closure iteration stabilizes after at most $|\Theta| - 1$ strict reductions. Its terminal set is the greatest fixed point of $\mathcal{T}_{\mathsf{A},\theta^\star}$ among subsets of Θ containing $\theta^\star$.

Proof. The first inclusion follows because the defining intersection for D contains every constraint in the intersection for C. The second follows because membership in the left-hand operator requires observational equivalence under the larger mechanism set. Contractivity is immediate from the definition. A strictly decreasing chain of subsets of a finite set containing $\theta^\star$ has at most $|\Theta| - 1$ strict steps. If $D = \mathcal{T}_{\mathsf{A},\theta^\star}(D)$, monotonicity and $D \subseteq \Theta = C_0$ imply inductively that $D \subseteq C_k$ for every k. Hence $D \subseteq C_\infty$, while stabilization makes C_∞ itself a fixed point. □

B.2 Obedient Experiment Geometry

To test whether the robust mechanism set distinguishes two models, compare their observable laws at one feasible mechanism and along all feasible changes from it. Affinity makes these checks sufficient; the base-point check retains differences that are constant across mechanisms.

Proposition 14 (Obedient experiment geometry)**.** *Under the finite normal-form realization above, fix C, $\theta, \theta' \in C$, and a nonempty robust polytope. For any $p^0 \in \mathsf{P}^\gamma_{\mathsf{A}}(C)$, the models agree under every currently obedient experiment if and only if*

$$(\mathsf{B}_{\mathsf{A},\theta} - \mathsf{B}_{\mathsf{A},\theta'})p^0 = 0, \qquad (\mathsf{B}_{\mathsf{A},\theta} - \mathsf{B}_{\mathsf{A},\theta'})\mathsf{L}_{\mathsf{A}}(C) = \{0\}. \tag{G}$$

If $C \subseteq D$, then $\mathsf{L}_{\mathsf{A}}(C) \supseteq \mathsf{L}_{\mathsf{A}}(D)$. Moreover, an observation-preserving replicated-public embedding E gives

$$E\mathsf{L}_{\mathrm{pub}}(C) \subseteq \mathsf{L}_{\mathrm{priv}}(C).$$

Thus candidate elimination and private disclosure can only add obedient experimental directions. They add pairwise information precisely when an added direction is not annihilated by the corresponding observation-law difference operator.

The operator image of $\mathsf{L}_{\mathsf{A}}(C)$ measures changes in observable laws attainable by varying mechanisms, not the dimension of the unknown parameter. We first justify the affine coordinates by sampling complete causal plans. Sampling only on-path actions would omit the continuation responses needed to evaluate deviations.

Lemma 15 (Finite normal-form realization)**.** *Suppose the mediator and receiver pre-action information histories, messages, and actions are finite. Let $\mathfrak{D}_{\mathsf{A}}$ contain every deterministic causal direct plan, including the mediator's continuation after unilateral deviations. If the mechanism's internal seed is independent of the physical chance moves and is not itself revealed to receivers, then every causal*

randomized direct mechanism is outcome-equivalent, under obedience and every unilateral causal deviation, to a distribution $p \in \Delta(\mathfrak{D}_{\mathsf{A}})$. For every model θ,

$$K_\theta^p = \sum_{d \in \mathfrak{D}_{\mathsf{A}}} p_d K_\theta^d. \tag{NR1}$$

Furthermore, dynamic γ-obedience is a finite family of affine inequalities in p.

Proof. Pre-sample the mechanism's complete internal seed, including its possibly correlated outputs at every mediator information history. Each realization determines a causal plan d whose distribution does not depend on θ. Executing it reproduces the original kernels, including after unilateral deviations. Conversely, sampling a hidden plan is a causal randomized mechanism. Conditioning on this draw proves (NR1) on any standard Borel learner observation space. Every column K_θ^d uses fixed following, even if the plan alone is not obedient. Receivers do not observe the draw, so this decomposition changes neither their information nor the response rule.

Conditional obedience divides by a recommendation-event probability that itself depends on the mixture. Multiplying through by that probability removes the denominator and handles null events in the same expression. Index a row j by a receiver, an on-path pre-action information event I_j, a first deviating action, and a pure causal continuation deviation. Let G_i be the receiver's discounted payoff from that event onward and define

$$c_{\theta j}(d) = \mathbb{E}_\theta^{d,\mathrm{obey}}[\mathbf{1}_{I_j} G_i] - \mathbb{E}_\theta^{d,j}[\mathbf{1}_{I_j} G_i] - \gamma \Pr_\theta^{d,\mathrm{obey}}(I_j). \tag{NR2}$$

The deviation starts at I_j, so its probability is common to the two counterfactual laws. Multiplying the conditional obedience inequality by this probability gives exactly

$$g_{\theta j}(p) := \sum_{d \in \mathfrak{D}_{\mathsf{A}}} p_d c_{\theta j}(d) \geq 0. \tag{NR3}$$

At a null event this is $0 \geq 0$. There are finitely many rows, and pure continuation deviations suffice because every randomized deviation is a mixture of pure ones. □

Allowing model-independent affine ex ante restrictions gives an ambient polytope $\mathcal{P}_{\mathsf{A}} \subseteq \Delta(\mathfrak{D}_{\mathsf{A}})$. The lemma yields

$$\mathsf{P}_{\mathsf{A}}^\gamma(C) = \{p \in \mathcal{P}_{\mathsf{A}} : g_{\theta j}(p) \geq 0 \ \forall \theta \in C, j\}. \tag{A.1}$$

The hidden-plan qualification is essential. If the realized plan is publicly announced, receivers condition on that branch and obedience must be checked branchwise; this is a different experiment class.

Proof of Proposition 14. Write $\mathsf{D} = \mathsf{B}_{\mathsf{A},\theta} - \mathsf{B}_{\mathsf{A},\theta'}$. If $\mathsf{D}p = 0$ for every robustly obedient p, then it vanishes at p^0 and on every difference of two feasible points, hence on $\mathsf{L}_{\mathsf{A}}(C)$. Conversely, for any feasible p, $p - p^0 \in \mathsf{L}_{\mathsf{A}}(C)$, so (G) gives

$$\mathsf{D}p = \mathsf{D}p^0 + \mathsf{D}(p - p^0) = 0.$$

If $C \subseteq D$, robustness over D adds affine rows. Therefore $\mathsf{P}_{\mathsf{A}}^\gamma(C) \supseteq \mathsf{P}_{\mathsf{A}}^\gamma(D)$, which proves inclusion of direction spaces. For the replicated-public embedding, public feasibility of p, q implies private feasibility of Ep, Eq, so $E(p-q) \in \mathsf{L}_{\mathrm{priv}}(C)$. Taking spans gives $E\mathsf{L}_{\mathrm{pub}}(C) \subseteq \mathsf{L}_{\mathrm{priv}}(C)$. Observation preservation yields

$$(\mathsf{B}_{\mathrm{priv},\theta} - \mathsf{B}_{\mathrm{priv},\theta'})E = \mathsf{B}_{\mathrm{pub},\theta} - \mathsf{B}_{\mathrm{pub},\theta'},$$

which proves the final claim. □

The same information test can be expressed through local feasible moves. Constraints binding at a base point determine which infinitesimal perturbations remain admissible. For a polytope $\mathsf{P} = \{p : Fp \geq 0, Ep = e\}$, collect these perturbations in the tangent cone

$$\mathsf{T}(\mathsf{P}; p^0) = \{d : \exists \epsilon_n \downarrow 0, d_n \to d,\ p^0 + \epsilon_n d_n \in \mathsf{P}\}.$$

Its exact active-set representation and span are

$$\begin{aligned} \mathsf{T}(\mathsf{P};p^0) &= \{d : Ed = 0,\ F_j d \geq 0 \text{ if } F_j p^0 = 0\}, \\ \operatorname{span} \mathsf{T}(\mathsf{P};p^0) &= \operatorname{span}\{p - q : p, q \in \mathsf{P}\}. \end{aligned} \tag{A.2}$$

The first identity follows by linearizing the active affine rows. Every feasible secant $p - p^0$ is tangent by convexity, while every tangent direction is a limit of scaled feasible secants, proving the second identity. Thus (G) is equivalently a base-point condition plus annihilation of every tangent direction. The base point cannot be dropped: if every atom produces δ_0 under θ and δ_1 under θ', the difference map annihilates all zero-sum tangent directions but is nonzero at every probability vector.

Proof of Proposition 4. We first find a common positive mixing weight that makes every ambient probe feasible. We then use affinity to show that the surviving models already agree on each probe itself. Index the robust obedience rows by a, write their base-point slacks as $s_a = g_a(p^0)$, and their worst ambient values as $m_a = \min_{q \in \mathcal{P}_\mathsf{A}} g_a(q)$. Compactness attains each minimum. Condition (RS) implies $s_a > 0$ whenever $m_a < 0$, so

$$\bar{\epsilon} = \min_{a : m_a < 0} \frac{s_a}{s_a - m_a} > 0, \tag{A.6}$$

with $\bar{\epsilon} = 1$ if no row can be violated. For every ambient q and $0 \leq \epsilon \leq \bar{\epsilon}$,

$$g_a((1-\epsilon)p^0 + \epsilon q) \geq (1-\epsilon)s_a + \epsilon m_a \geq 0.$$

Thus every ambient experiment can be diluted into the robust set.

Let $\theta' \in \mathcal{T}_{\mathsf{A},\theta}(C)$, fix q, and set $p_\epsilon = (1-\epsilon)p^0 + \epsilon q$. Observable laws agree at both p^0 and p_ϵ. Subtracting in the vector space of finite signed measures and using affinity gives

$$0 = (K_{\theta'}^{p_\epsilon} - K_\theta^{p_\epsilon}) - (1-\epsilon)(K_{\theta'}^{p^0} - K_\theta^{p^0}) = \epsilon(K_{\theta'}^{q} - K_\theta^{q}).$$

Since q was arbitrary, every surviving model agrees with the truth under the full ambient polytope. It therefore remains indistinguishable under all experiments unlocked after candidate shrinkage, proving one-step stabilization. □

Proof of Corollary 5. An informative probe must fail even at arbitrarily small mixture weights; otherwise it would already separate the models. Fix a current base point p^0. If every row binding at p^0 were nonnegative at q, those rows would remain feasible along the segment to q, while finitely many strict rows would remain positive for sufficiently small $\epsilon > 0$. Thus $(1-\epsilon)p^0 + \epsilon q$ would be currently robust. Equality of the two model laws at this mixture and at p^0, together with affinity, would imply $K_{\theta'}^q = K_\theta^q$, a contradiction. Some row must therefore bind at p^0 and be negative at q. □

The affine-law assumption is indispensable. Let the ambient experiment be $p \in [0,1]$, the safe set be $[0, 1/2]$, and compare a constant δ_0 observation with $\text{Bernoulli}(h(p))$, where $h(p) = 0$ on $[0, 1/2]$ and $h(p) = 2p - 1$ above $1/2$. Safe dilution is possible in every direction, but the models agree on all safe experiments and differ at $p = 1$. Likewise, if receivers observe which branch of a mixture was selected, obedience is checked conditional on that branch and the experiment polytope has additional signal-indexed rows.

Proposition 16 (Support-gated continuous unlocking)**.** *There is a three-model, one-receiver, binary-action problem over the full convex class of type-blind direct kernels for which, at either of two truths, the closure makes two strict reductions and the intermediate set is decision-incompatible. The construction has strict IC on every used recommendation event; unlocking is caused by a zero-support recommendation, not a payoff equality at the margin.*

Proof. We need one candidate to prohibit the probe entirely, while the safe action can rule out that candidate. Let $W \sim \text{Bernoulli}(1/4)$, let actions and recommendations be $\{S, P\}$, and set $\gamma = 1/4$. In models α, β, the terminal state determining receiver utility is

$$X = \mathbf{1}\{W = 1, A = P\} + \mathbf{1}\{W = 0, A = S\};$$

in model ζ, it is $X = \mathbf{1}\{A = S\}$. The utility map is fixed, $u_R = X$; only the terminal transition changes with the model.

Under ζ, action P is strictly dominated by S, with gap one. Therefore every γ-obedient direct kernel has $\mathbb{P}(R = P) = 0$, and the unique induced experiment always recommends S. Its follow-minus-deviate gap is $1/2 > \gamma$ in α, β and $1 > \gamma$ in ζ. To make this safe experiment remove the blocking model but leave the final decision unresolved, let the learner observe W, R, A and

$$Y = \mu_\theta(W, A) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

where

$$\begin{aligned} \mu_\alpha &= 0, \\ \mu_\beta &= \mathbf{1}\{W = 1, A = P\}, \\ \mu_\zeta &= \mathbf{1}\{A = S\}. \end{aligned} \tag{A.7}$$

The only initially robust experiment yields identical laws in α, β and a unit mean shift in ζ. Thus, at either of the first two truths, $C_1 = \{\alpha, \beta\}$.

After ζ is removed, recommend P iff $W = 1$ and S iff $W = 0$. Each recommendation then has obedience gap one in α, β, while the P recommendation is disobedient in ζ. The probe separates α and β, with per-episode

$$D_{\mathrm{KL}}(K_\alpha \| K_\beta) = \frac{1}{8\sigma^2},$$

so C_2 is the true singleton. Finally, take unobserved sender payoffs

$$\begin{aligned} G_\alpha &= \mathbf{1}\{A = S\}, \\ G_\beta &= \mathbf{1}\{W = 1, A = P\} + \mathbf{1}\{W = 0, A = S\}. \end{aligned}$$

The unique direct outcome oracle is always S in α and the probe in β, so the intermediate set has no common oracle. The initial robust set lies on the support face $\mathbb{P}(R = P) = 0$, and the vacuous zero-probability P-recommendation IC is violated in every direction that adds P mass; this is exactly the failure of (RS). □

A finite menu can also produce successive unlocking, but its restriction matters: hidden mixtures may make a nominally unavailable probe accessible immediately. The next example distinguishes this discrete-menu effect from the support boundary in Proposition 16.

Proposition 17 (Finite-library unlocking witness)**.** *Let $\gamma = 1/10$, $\vartheta = (\alpha, \beta)$, and*

$$\Theta = \{(0, 1/5), (0, 2/5), (4/5, 0)\}.$$

For two binary receivers, set $Y = \alpha\mathbf{1}\{A = 01\} + \beta\mathbf{1}\{A = 10\} + \varepsilon$ and $U_i = \mathbf{1}\{A_i = A_j\} + Y$. Restrict the private library to

$$\begin{aligned} b_0 = (1, 0, 0, 0), \quad b_1 = (0, 0, 0, 1), \\ p^\alpha = (5/12, 1/6, 0, 5/12), \quad p^\beta = (2/5, 0, 1/5, 2/5). \end{aligned}$$

For either of the first two truths, the closure path is $\Theta \to \{(0, 1/5), (0, 2/5)\} \to \{\theta^\star\}$. For the unobserved objective $J_{(\alpha,\beta)}(p) = (\beta - 3/10)(p_{11} - p_{00}) - c(p_{01} + p_{10})$ with $c \geq 0$, the intermediate set is decision-incompatible.

Proof of Proposition 17. Write $w_\alpha = 1 - \alpha$ and $w_\beta = 1 - \beta$. The four private obedience gaps for a law p, omitting zero-probability events, are

$$\begin{aligned} \Gamma_{1,0} &= \frac{w_\beta p_{00} - w_\alpha p_{01}}{p_{00} + p_{01}}, \quad \Gamma_{1,1} = \frac{w_\alpha p_{11} - w_\beta p_{10}}{p_{10} + p_{11}}, \\ \Gamma_{2,0} &= \frac{w_\alpha p_{00} - w_\beta p_{10}}{p_{00} + p_{10}}, \quad \Gamma_{2,1} = \frac{w_\beta p_{11} - w_\alpha p_{01}}{p_{01} + p_{11}}. \end{aligned}$$

Direct substitution gives

	ϑ_1	ϑ_2	ϑ_3
$\min_{i,r} \Gamma_{i,r}(p^\alpha)$	$2/7$	$1/7$	$1/5$
$\min_{i,r} \Gamma_{i,r}(p^\beta)$	$2/5$	$7/15$	$-1/5$.

Both matched laws have minimum gap at least $1/5$ in every model. Hence the mechanisms robust at Θ are exactly $\{b_0, b_1, p^\alpha\}$, and p^β is added after ϑ_3 is removed.

Matched actions make the telemetry mean zero. The only mismatch used by p^α is 01, so its observable law depends on α and separates ϑ_3 from the other two models, which both have $\alpha = 0$. The law p^β assigns positive mass to 10, so after it is unlocked its observable law separates $\beta = 1/5$ from $\beta = 2/5$. This proves the two displayed closure reductions.

Finally, the matched laws have values $(1/10, -1/10)$ under ϑ_1 and $(-1/10, 1/10)$ under ϑ_2. The diagnostics have values $-c/6$ and $-c/5$ in both models, which are nonpositive because $c \geq 0$. Thus the unique oracles are respectively b_0 and b_1, and the intermediate set has no common oracle. □

This finite-library example is not robust to hidden convexification. Since all IC rows of p^α are initially strict, a sufficiently small hidden mixture $(1-\epsilon)p^\alpha + \epsilon p^\beta$ is already robust on Θ and probes both mismatch directions in the first closure step. If the branch label is announced to receivers, however, obedience is checked inside the p^β branch and the shortcut is unavailable. The example therefore describes a discrete or branch-revealing protocol, whereas Proposition 16 covers the full convex direct class.

B.3 Stability of a Thresholded Finite-Library Recursion

Small errors can change exact closure because equality of laws is not a margin condition. For a single strictly obedient experiment, take three Bernoulli feedback laws with means $1/2, 1/2, 3/4$. At the first model, closure retains the first two models. Changing only the second mean to $1/2 + \epsilon$, for any $0 < \epsilon < 1/4$, makes the terminal closure a singleton despite positive obedience slack and separation from the third model. The next result replaces exact equality by a resolution threshold. Separate margins for feasibility and law separation ensure that bounded errors change neither which experiments are admitted nor which candidates they remove.

Proposition 18 (Thresholded closure stability)**.** *Fix finite model and mechanism sets $\Theta, \mathcal{L}$. Let $b_\theta(\pi)$ be a feasibility score satisfying $\pi \in \Pi^\gamma_{\mathsf{A}}(\theta)$ if and only if $b_\theta(\pi) \geq 0$. Suppose $|b_\theta(\pi)| \geq a > 0$ for every pair. Assume every unequal pair $K^\pi_\theta, K^\pi_{\theta'}$ has total variation distance at least $d > 0$. Let approximate scores and laws satisfy uniformly*

$$|\widetilde{b}_\theta(\pi) - b_\theta(\pi)| \leq \epsilon_b < a/2, \qquad \mathrm{TV}(\widetilde{K}^\pi_\theta, K^\pi_\theta) \leq \epsilon_K < d/4.$$

Define

$$\widetilde{\Pi}(C) = \{\pi \in \mathcal{L} : \widetilde{b}_\vartheta(\pi) \geq \epsilon_b \ \forall \vartheta \in C\},$$
$$\widetilde{\mathcal{T}}_\theta(C) = \{\theta' \in C : \mathrm{TV}(\widetilde{K}^\pi_{\theta'}, \widetilde{K}^\pi_\theta) \leq d/2 \ \forall \pi \in \widetilde{\Pi}(C)\}.$$

Then $\widetilde{\Pi}(C) = \mathcal{L} \cap \Pi^\gamma_{\mathsf{A}}(C)$ and $\widetilde{\mathcal{T}}_\theta(C) = \mathcal{T}_{\mathsf{A},\theta}(C)$, where the latter exact operator is restricted to $\mathcal{L}$. Starting from the same candidate set, the approximate and exact recursions have identical iterates and terminal sets.

Proof. If $b_\vartheta(\pi) \geq a$, then $\widetilde{b}_\vartheta(\pi) \geq a - \epsilon_b > \epsilon_b$. If $b_\vartheta(\pi) \leq -a$, then $\widetilde{b}_\vartheta(\pi) \leq -a + \epsilon_b < \epsilon_b$. Intersecting these decisions over $\vartheta \in C$ preserves the robust library exactly. The triangle inequality gives

$$\left|\mathrm{TV}(\widetilde{K}^\pi_\theta, \widetilde{K}^\pi_{\theta'}) - \mathrm{TV}(K^\pi_\theta, K^\pi_{\theta'})\right| \leq 2\epsilon_K.$$

An originally equal pair has approximate distance at most $2\epsilon_K < d/2$, while an unequal pair has distance at least $d - 2\epsilon_K > d/2$. Thus the threshold test preserves every equality decision. The operators agree for every candidate set containing the truth, so induction proves equality of the entire recursion, including adaptively unlocked libraries. □

The margin a separates both feasible and infeasible library mechanisms from the obedience boundary at level γ. For finitely many inequalities, take the minimum residual after omitting structurally zero rows preserved by perturbation. Thus $\gamma > 0$ alone does not supply the required slack, and nontrivial binding rows remain outside this certificate. Conditional on the uniform errors, the result preserves a finite library's elimination path at resolution $d/2$, not the perturbed world's exact closure or full convex occupancy support. Joint error bounds holding with probability $1 - \delta$ give the same guarantee for the entire path. Estimating those errors and preserving oracle compatibility or regret require additional analysis.

B.4 Architecture Nesting

Architecture refinement enlarges access, but the additional experiments may reveal no new model differences. To decide whether the final closures differ, test the private experiments at the public terminal set $C^{\mathrm{pub}} = C^{\mathrm{pub}}_{\infty}(\theta^\star)$. The criterion below asks whether any private direction exposes a distinction that survives all public experiments.

Proposition 19 (Terminal kernel-escape criterion)**.** *Under the assumptions of Proposition 6,*

$$C^{\mathrm{priv}}_{\infty}(\theta^\star) = C^{\mathrm{pub}} \quad \Longleftrightarrow \quad \mathcal{T}_{\mathrm{priv},\theta^\star}(C^{\mathrm{pub}}) = C^{\mathrm{pub}}. \tag{A}$$

In the finite direct affine setting of Proposition 14, assume $\mathsf{P}^{\gamma}_{\mathrm{pub}}(C^{\mathrm{pub}}) \neq \varnothing$ *and identify embedded public laws in the private ambient space. Define the common private observation kernel*

$$\mathsf{N}_{\theta^\star}(C^{\mathrm{pub}}) = \bigcap_{\theta \in C^{\mathrm{pub}}} \ker(\mathsf{B}_{\mathrm{priv},\theta} - \mathsf{B}_{\mathrm{priv},\theta^\star}).$$

Then

$$C^{\mathrm{priv}}_{\infty}(\theta^\star) = C^{\mathrm{pub}} \quad \Longleftrightarrow \quad \mathsf{L}_{\mathrm{priv}}(C^{\mathrm{pub}}) \subseteq \mathsf{N}_{\theta^\star}(C^{\mathrm{pub}}). \tag{B}$$

Equivalently, the terminal inclusion is strict exactly when there exist $\theta \in C^{\mathrm{pub}}$ *and* $d \in \mathsf{L}_{\mathrm{priv}}(C^{\mathrm{pub}})$ *such that*

$$(\mathsf{B}_{\mathrm{priv},\theta} - \mathsf{B}_{\mathrm{priv},\theta^\star})d \neq 0. \tag{C}$$

Proof of Proposition 6. The two recursions need not visit the same candidate sets. We account separately for private simulation and the extra access released by a smaller set. Let C^{priv}_k and C^{pub}_k start from Θ. Inclusion holds at $k = 0$. Suppose $C^{\mathrm{priv}}_k \subseteq C^{\mathrm{pub}}_k$, and take $q \in C^{\mathrm{priv}}_{k+1}$. Then $q \in C^{\mathrm{pub}}_k$, and

$$\Pi^{\gamma}_{\mathrm{priv}}(C^{\mathrm{priv}}_k) \supseteq \Pi^{\gamma}_{\mathrm{priv}}(C^{\mathrm{pub}}_k) \supseteq E(\Pi^{\gamma}_{\mathrm{pub}}(C^{\mathrm{pub}}_k)).$$

The law of q agrees with the true law under every mechanism in the leftmost set, hence under every embedded public mechanism in the rightmost set. Since the embedding preserves observable laws, $q \in C^{\mathrm{pub}}_{k+1}$. Induction and finite stabilization prove the claim. □

Proof of Proposition 19. Let $F = \mathcal{T}_{\mathrm{priv},\theta^\star}$ and $C^{\mathrm{pub}} = C^{\mathrm{pub}}_{\infty}(\theta^\star)$. If $C^{\mathrm{priv}}_{\infty} = C^{\mathrm{pub}}$, stabilization gives $F(C^{\mathrm{pub}}) = C^{\mathrm{pub}}$. Conversely, if $F(C^{\mathrm{pub}}) = C^{\mathrm{pub}}$, then C^{pub} is a private fixed point. The greatest-fixed-point property gives $C^{\mathrm{pub}} \subseteq C^{\mathrm{priv}}_{\infty}$, while Proposition 6 gives the reverse inclusion. This proves (A).

For the affine statement, choose $q^0 \in \mathsf{P}^{\gamma}_{\mathrm{pub}}(C^{\mathrm{pub}})$ and set $p^0 = Eq^0$. Because C^{pub} is a public fixed point, for every $\theta \in C^{\mathrm{pub}}$,

$$(\mathsf{B}_{\mathrm{priv},\theta} - \mathsf{B}_{\mathrm{priv},\theta^\star})p^0 = (\mathsf{B}_{\mathrm{pub},\theta} - \mathsf{B}_{\mathrm{pub},\theta^\star})q^0 = 0.$$

Thus the base-point condition in Proposition 14 is already discharged. Applying that proposition to each $\theta \in C^{\mathrm{pub}}$ shows that every candidate survives all private experiments robust on C^{pub} if and only if

$$\mathsf{L}_{\mathrm{priv}}(C^{\mathrm{pub}}) \subseteq \bigcap_{\theta \in C^{\mathrm{pub}}} \ker(\mathsf{B}_{\mathrm{priv},\theta} - \mathsf{B}_{\mathrm{priv},\theta^\star}).$$

Combining this equivalence with (A) proves (B). Negating the finite intersection condition gives (C) and the strict case. □

With finite feedback, the kernel test reduces to linear algebra on feasible vertices. Differences from one vertex span all feasible directions, so their images certify whether private access adds information.

Corollary 20 (Finite vertex/rank certificate)**.** *Assume the setting of Proposition 19 and a finite learner observation alphabet. Let* $v^0, \dots, v^{N_V - 1}$ *be all* N_V *vertices of* $\mathsf{P}^{\gamma}_{\mathrm{priv}}(C^{\mathrm{pub}})$*, and let*

$$\mathsf{D}d = \big((\mathsf{B}_{\mathrm{priv},\theta} - \mathsf{B}_{\mathrm{priv},\theta^\star})d\big)_{\theta \in C^{\mathrm{pub}}}.$$

Then the terminal architecture-identification gain

$$g_{\theta^\star} = \operatorname{rank}\begin{bmatrix} \mathsf{D}(v^1 - v^0) & \cdots & \mathsf{D}(v^{N_V - 1} - v^0) \end{bmatrix} \tag{A.8}$$

is zero if and only if the public and private terminal closures agree. When $g_{\theta^\star} > 0$*, any nonzero displayed column is a kernel-escape witness.*

Proof. The differences $v^j - v^0$ span the affine direction space of the private polytope. Hence the displayed column space is exactly $\mathsf{DL}_{\mathrm{priv}}(C^{\mathrm{pub}})$. Apply Proposition 19. Completeness of the vertex list is essential; an incomplete feasible sample gives only a lower bound on the rank. □

The following family assigns one unknown transition coordinate to each mismatched action. Public obedience excludes both rows; private correlation visits both. It gives a concrete instance in which the kernel-escape certificate has rank two.

Proposition 21 (Two-coordinate Bernoulli transition family)**.** *Consider two receivers, binary stage-one actions, and coordination utilities $U_i = \mathbf{1}\{A_i = A_j\}$. Let $\eta = (\eta_1, \eta_2) \in \Theta_\Delta = \{-\Delta, +\Delta\}^2$, where $0 < \Delta < 1/2$, and let the observed stage-two state satisfy*

$$Z \mid A \sim \mathrm{Bernoulli}\big(\tfrac{1}{2} + \eta_1 \mathbf{1}\{A = 01\} + \eta_2 \mathbf{1}\{A = 10\}\big). \tag{A.9}$$

The learner observes (A, Z). For one-way public and receiver-private direct labels and $0 \le \gamma < 1$, every truth $\eta^\star \in \Theta_\Delta$ satisfies

$$C_\infty^{\mathrm{pub}}(\eta^\star) = \Theta_\Delta, \qquad C_\infty^{\mathrm{priv}}(\eta^\star) = \{\eta^\star\}, \qquad g_{\eta^\star} = 2. \tag{A.10}$$

Proof. Under a publicly observed mismatch label, following yields zero while switching yields one, so public obedience forces $p_{01} = p_{10} = 0$. Matched labels have deviation gap one and are obedient for $\gamma \le 1$. The transition in (A.9) is independent of η on matched profiles, proving the public closure identity.

Private direct obedience gives $(1-\gamma)p_{00} \ge (1+\gamma)p_{01}, (1+\gamma)p_{10}$ and the same two inequalities with p_{11} in place of p_{00}. Put $r = (1-\gamma)/(1+\gamma) > 0$. The distribution

$$p^{\mathrm{diag}} = \frac{(1, r, r, 1)}{2(1+r)}$$

is robustly γ-obedient and assigns positive probability to both mismatch profiles. Any $\eta \neq \eta^\star$ differs in at least one coordinate. Conditional on the corresponding observed mismatch action, its Bernoulli law differs from the true law. Thus one private closure step leaves only $\eta^\star$.

Finally, public directions are spanned by $e_{00} - e_{11}$, which every model-law difference annihilates. The private direction space contains $e_{01} - e_{00}$ and $e_{10} - e_{00}$. Across Θ_Δ, their images occupy disjoint observed-action blocks and are both nonzero, so the rank in (A.8) is two. At the excluded boundary $\gamma = 1$, $r = 0$, the private polytope collapses to the matched face and the gain becomes zero. □

C Obedient Learnability and Diagnostic Complexity

C.1 Closure Characterization

The exact closure describes available distinctions, but a learner must realize them from noisy samples without deleting the truth. We first work with finite mechanism and model classes. The testing interface makes each candidate's predictions evaluable without revealing which candidate is true; predictability requires choosing the experiment before seeing its feedback. Diagnostic blocks then implement reductions, while a stopped-history argument rules out escaping an incompatible terminal set under all-round obedience.

Assumption 22 (Finite-class testing interface)**.** The finite classes $\Theta, \mathcal{P}_{\mathsf{A}}$, dominated kernels K_θ^π, and bounded payoffs are known. Every model has an obedient mechanism; choices are predictable; and obedient and oracle sets are computable.

Theorem 23 (Finite-class obedient learnability)**.** *Under Assumption 22, there exists a sequence of uniformly all-round (δ_T, γ)-obedient algorithms with $\delta_T \to 0$ and $\mathbb{E}_\theta R_T(\theta) = o(T)$ for every $\theta \in \Theta$ if and only if*

$$C_\infty^{\mathsf{A}}(\theta) \quad \textit{is decision-compatible for every } \theta \in \Theta.$$

Proof of Theorem 23: sufficiency. The learner needs another diagnostic only while its retained models lack a common oracle. Let $m = |\Theta|$, and assume every terminal closure is compatible. For any

nonempty incompatible C and any $\theta \in C$, C cannot be a fixed point of $\mathcal{T}_{\mathsf{A},\theta}$: it would then lie in the greatest fixed point $C^{\mathsf{A}}_{\infty}(\theta)$, whose compatibility is hereditary. Hence the robust mechanism set is nonempty and $D_C(\theta) = C \setminus \mathcal{T}_{\mathsf{A},\theta}(C) \neq \varnothing$. Choose a full-support distribution ρ_C on the finite robust set. Its minimum squared Hellinger separation over pairs (θ, q) with $q \in D_C(\theta)$ is $h_C > 0$. Including the chosen mechanism in each observation makes the joint separation equal to the ρ_C-average of the individual distances.

Each block must retain the truth and remove all currently distinguishable candidates. Allocate failure budget $\beta = \delta/m$, set $a = \log(2m/\beta)$ and $N_C = \lceil 3a/(2h_C) \rceil$, then draw N_C independent experiments from ρ_C and retain

$$C^+ = \{q \in C : \Lambda_{N_C}(q) \geq e^{-a} \max_{r \in C} \Lambda_{N_C}(r)\},$$

where Λ is the fresh block likelihood. Conditionally on entry with the truth retained, all block experiments are obedient under that truth. The expectation of $\Lambda(r)/\Lambda(\theta)$ on the true support is at most one. The probability of deleting the truth is therefore at most $me^{-a} = \beta/2$. For a distinguishable q, Markov's inequality on the square-root likelihood ratio bounds its retention probability by

$$e^{a/2}(1 - h_C)^{N_C} \leq e^{a/2 - N_C h_C} \leq e^{-a}.$$

A union bound shows that truth retention and removal of all distinguishable candidates hold with probability at least $1 - \beta$. These arguments also cover singular laws and require no moments of log-likelihood ratios.

Stop at a compatible set and use its common oracle. Cap the number of diagnostic blocks at $m - 1$, with an arbitrary fixed fallback if the cap is reached while the set remains incompatible. A union bound over the conditional block guarantees gives a joint success event of probability at least $1 - \delta$; on that event the fallback is never used, every round is obedient, and exploitation has zero regret. There are finitely many candidate sets and each relevant h_C is positive. Hence the deterministic cap on diagnostic rounds is $O(\log(1/\delta))$. For a known horizon, truncate any unfinished block at T and use $\delta_T = T^{-2}$ for $T \geq 2$. Bounded payoffs give $\mathbb{E}_\theta R_T = O(\log T) + O(T\delta_T) = O(\log T)$. The all-round obedience failure probability is at most δ_T. □

The block construction allocates samples using the horizon. To retain the truth when observations are checked after every episode, compare cumulative likelihoods and remove a candidate only when its likelihood is sufficiently smaller than a retained competitor's. This gives the nested confidence sequence used in Corollary 26:

$$L_t(q) = \prod_{s=1}^{t} k_q^{\pi_s}(Z_s), \qquad C_t = C_{t-1} \cap \left\{ q : L_t(q) \geq \frac{\delta}{m} \max_{r \in C_{t-1}} L_t(r) \right\}.$$

Select only mechanisms robust over the current set. For each alternative, its likelihood ratio to the truth, stopped when either is removed, is a nonnegative supermartingale on the true support. Ville's inequality and a union bound imply $\Pr_\theta(\theta \in C_t \ \forall t) \geq 1 - \delta$. This is anytime coverage at fixed δ, not almost-sure obedience. The block construction above gives the vanishing-failure expected-regret claim; no fixed failure probability is silently discarded.

Proof of Theorem 23: necessity. Inside a terminal set, every robust experiment gives the same evidence across models. Obtaining different evidence therefore requires leaving that robust set and violating obedience in at least one candidate. Fix θ_0 with incompatible $C = C^{\mathsf{A}}_{\infty}(\theta_0)$; then $K^{\pi}_q = K^{\pi}_{\theta_0}$ for every $q \in C$ and $\pi \in \Pi^{\gamma}_{\mathsf{A}}(C)$. Let τ be the first episode that selects a mechanism outside this robust set. Before τ, the complete adaptive history law, including learner randomization, is common to all $q \in C$. Let $B_T(q)$ denote use of a mechanism nonobedient under q by time T. The common stopped law and all-round safety give

$$\mathbb{P}\{\tau \leq T\} \leq \sum_{q \in C} \mathbb{P}_q(B_T(q)) \leq |C|\delta_T.$$

If $\Pi^{\gamma}_{\mathsf{A}}(C) = \varnothing$, then $\tau = 1$, and the same union bound gives $1 \leq |C|\delta_T$, contradicting $\delta_T \to 0$. Hence it remains to consider a nonempty robust set.

For $\pi \in \Pi^{\gamma}_{\mathsf{A}}(C)$, set $\Delta_q(\pi) = J^\star_q - J_q(\pi)$. Finiteness and incompatibility imply

$$g_C = \min_{\pi \in \Pi^{\gamma}_{\mathsf{A}}(C)} \sum_{q \in C} \Delta_q(\pi) > 0.$$

On $\{\tau > T\}$, the sum of regrets across models in C is at least $g_C T$. Bounded payoffs limit the contribution of $\{\tau \le T\}$ to $O(|C|^2 \delta_T T)$, so

$$\sum_{q \in C} \mathbb{E}_q R_T(q) \ge g_C T(1 - |C|\delta_T) - O(|C|^2 \delta_T T).$$

Thus $\max_{q \in C} \mathbb{E}_q R_T(q) = \Omega(T)$ for all large T, contradicting pointwise $o(T)$ regret on the finite set C. □

The all-round safety requirement is essential. Take two models whose common obedient mechanisms are uninformative and whose oracle sets are disjoint. Add one diagnostic whose action reveals the model but is disobedient in one model. Using it once and then committing gives $O(1)$ regret and one violation, so the violation fraction tends to zero despite an incompatible closure. This falsifies necessity under asymptotic violation frequency.

Proof of Proposition 25. For a mixture ρ, include the sampled mechanism in the observation and write $\overline{K}^{\rho}_{\theta}(d\pi, dz) = \rho(\pi) K^{\pi}_{\theta}(dz)$. Since the mechanism marginal is common,

$$\mathsf{h}^2(\overline{K}^{\rho}_{\theta}, \overline{K}^{\rho}_{q}) = \sum_{\pi} \rho(\pi) \mathsf{h}^2(K^{\pi}_{\theta}, K^{\pi}_{q}).$$

If $q \in D_C(\theta)$, some robust mechanism induces different laws and therefore has positive squared Hellinger distance. A full-support mixture has positive distance for every $q \in D_C(\theta)$; finiteness makes their minimum positive. The converse is immediate. Combining this equivalence at $C = C^{\mathrm{pub}}_{\infty}(\theta^\star)$ with Proposition 19 proves the architecture statement.

Now suppose C is incompatible and the terminal-closure condition holds. If $D_C(\theta)$ were empty for some $\theta \in C$, then C would be a fixed point contained in the greatest fixed point $C^{\mathsf{A}}_{\infty}(\theta)$. Hereditary decision compatibility would make C compatible, a contradiction. Hence every indexed set is nonempty. A full-support mixture is positive on every distinguishable ordered pair, so finiteness also gives $\underline{\chi}_{\mathsf{A}}(C) > 0$.

Let $a = \log(2m/\beta)$, draw N_C independent joint experiments from a maximizing mixture in (RHE), and let $\Lambda_N(q)$ be the block likelihood. Update to

$$C^+ = \left\{ q \in C : \Lambda_N(q) \ge e^{-a} \max_{r \in C} \Lambda_N(r) \right\}.$$

If the truth is removed, some $r \in C$ has $\Lambda_N(r)/\Lambda_N(\theta) > e^a$. The expectation of this likelihood ratio under θ is at most one, so Markov's inequality and a union bound give probability at most $me^{-a} = \beta/2$.

For $q \in D_C(\theta)$, retention implies $\Lambda_N(q)/\Lambda_N(\theta) \ge e^{-a}$. Markov's inequality applied to its square root gives

$$\begin{aligned} \mathbb{P}_\theta\{q \in C^+\} &\le e^{a/2} \left(\int \sqrt{d\overline{K}^{\rho}_{\theta}\, d\overline{K}^{\rho}_{q}} \right)^{N_C} \\ &\le \exp\{a/2 - N_C \underline{\chi}_{\mathsf{A}}(C)\} \le e^{-a} = \frac{\beta}{2m}. \end{aligned}$$

The second inequality uses $1 - x \le e^{-x}$ and the mixture identity above. Union bounding over q proves that, with probability $1 - \beta$, the truth is retained and $C^+ \subseteq \mathcal{T}_{\mathsf{A},\theta}(C)$. There are at most $m - 1$ incompatible blocks. Taking $\beta = \delta/m$ proves (HBC) and uniform all-round obedience on the same event. □

Proof of Corollary 26. Use threshold $\delta/(2m)$, so Ville's inequality spends at most $\delta/2$ on simultaneous true-model coverage. At entry to a visited set C, retention of both θ and q implies

$$\left| \log \frac{L_{\tau_C}(\theta)}{L_{\tau_C}(q)} \right| \le A_\delta.$$

Conditional on this entry history, extend the fixed design ρ_C to an infinite i.i.d. stream. The actual experiment agrees with this stream until the candidate set changes. Condition (LT) and the sub-gamma Chernoff bound give, for each $q \in D_C(\theta)$, with

$$a_{n,Cq} = n\mu_{Cq} - \sqrt{2nv\mu_{Cq}x_\delta} - bx_\delta,$$

$$\mathbb{P}_\theta\left\{\sum_{s=1}^{n} X_{Cq,s} < a_{n,Cq} \,\middle|\, H_{\tau_C}\right\} \le e^{-x_\delta}.$$

The inequality $\sqrt{2nv\mu_{Cq}x_\delta} \le n\mu_{Cq}/2 + vx_\delta$ shows that the corollary's sample size makes the right-hand lower bound at least $3A_\delta$. This exceeds the worst $2A_\delta$ future evidence needed to move the total likelihood ratio beyond the deletion threshold. If the set remains unchanged for that many samples, every distinguishable candidate would therefore cross its deletion threshold. Hence the stage must end within the stated duration, except with probability e^{-x_δ}.

Each strict set change removes at least one model. There are therefore at most $m-1$ visited incompatible sets and fewer than m^2 conditional stage–candidate events. A union bound spends at most $\delta/2$ on their durations. Combining it with the coverage event proves the simultaneous $1-\delta$ statement.

Without (LT), finite termination still holds almost surely on the truth-retention event $E_\theta = \{\theta \in C_t \text{ for all } t\}$. Indeed, at any incompatible entry set $C \ni \theta$, choose $q \in D_C(\theta)$ separated by the full-support design, with joint squared Hellinger distance $h > 0$. If C remains unchanged for n rounds, retention and the entry offset imply that the future likelihood ratio of q to θ is at least e^{-2A_δ}. The square-root likelihood bound makes the probability of this event at most $e^{A_\delta}(1-h)^n$. Hence every such stage ends almost surely, and finitely many deletions yield a compatible set on E_θ. Writing τ for the diagnostic duration, dominated convergence gives, for fixed δ,

$$\mathbb{E}_\theta[(\tau \wedge T)\mathbf{1}_{E_\theta}] = o(T), \qquad \mathbb{E}_\theta[\tau \wedge T] \le o(T) + \frac{\delta T}{2}.$$

This does not assert unconditional almost-sure termination or expected $o(T)$ at fixed confidence. The horizon-dependent expected-regret guarantee uses the capped block construction in Theorem 23. □

C.2 Finite Witness Extraction and Compact-Class Learnability

An infinite mechanism class need not require infinitely many diagnostic choices. With finitely many models, it suffices to retain one experiment for each available pairwise distinction and one common oracle for each compatible set. This preserves both the reductions the learner can make and the decisions at which it can stop.

Lemma 24 (A finite library preserves closure and decisions)**.** *Let Θ have m elements, and suppose every individual oracle is attained in an arbitrary mechanism class $\mathcal{P}$. There is a library $\mathcal{F} \subseteq \mathcal{P}$ of cardinality at most $(m^2+1)2^m$ such that, simultaneously for every nonempty $C \subseteq \Theta$, its robust one-step closure equals the full-class closure for each truth in C, its individual oracle values equal those of the full class, and its common-oracle intersection is nonempty exactly when the full-class intersection is nonempty.*

Proof. For each nonempty C and each ordered pair $(\theta, q) \in C^2$ for which some $\pi \in \Pi^\gamma(C)$ satisfies $K_\theta^\pi \neq K_q^\pi$, select one such witness. For each decision-compatible C, also select one element of $\bigcap_{\theta \in C} \Pi^\star(\theta)$. Let $\mathcal{F}$ be the union of these choices. There are at most m^2+1 choices per subset, proving the size bound.

A robust library experiment is a robust full-class experiment. Conversely, every pair distinguishable by the full robust class has its selected witness in $\mathcal{F} \cap \Pi^\gamma(C)$. Thus the equality-of-laws relation, and hence every one-step closure, is preserved on every C. The singleton choices include a full-class oracle for each model, so maximizing within the library gives exactly the full-class oracle value. It follows that the library oracle set for each model is precisely the intersection of its full-class oracle set with $\mathcal{F}$. The common-oracle choice for each compatible C proves preservation in one direction; the other direction follows from inclusion in the full class. Iterating the equal one-step operators also preserves all terminal closures. □

Proof of Theorem 7. Compactness, closed feasibility, and continuity imply that each individual oracle is attained. Apply Lemma 24; all payoffs and kernels restricted to its finite library meet the finite testing interface. The library has the same closure and decision compatibility, so the block proof of Theorem 23 proves sufficiency and the fixed-instance $O(\log T)$ upper bound. Exact separation and oracle access suffice to supply the finitely many choices. The library can be exponential in m; the result is not a computational efficiency guarantee.

For necessity, let $C = C^{\mathsf{A}}_{\infty}(\theta_0)$ be incompatible. The stopped adaptive-history argument of Theorem 23 does not require finitely many mechanisms. Measurability permits integration over any randomized adaptive mechanism choice, and equality of every available kernel gives equality of the stopped laws. Thus the probability of choosing outside $\Pi^{\gamma}(C)$ by time T is at most $|C|\delta_T$. If that robust set is empty, this contradicts vanishing failure probability.

Otherwise $\Pi^{\gamma}(C)$ is compact. The continuous function

$$g(\pi) = \sum_{\theta \in C} [J^{\star}_{\theta} - J_{\theta}(\pi)]$$

is nonnegative there and has no zero: a zero would be a common oracle. It therefore attains a minimum $g_C > 0$. On the common event of no exit, total regret across C is at least $g_C T$. Uniform boundedness over the finite model set bounds the complement by $O(|C|^2 T\delta_T)$, giving the same linear lower bound as in the finite proof. □

Compactness cannot simply be omitted from necessity. Let $\Theta = \{+, -\}$, let every experiment return the same constant observation, and let every mechanism be obedient. Give the class distinct oracles a_+, a_- and a sequence b_n, $n \geq 2$. Set $J_+(a_+) = J_-(a_-) = 1$, $J_-(a_+) = J_+(a_-) = 0$, and $J_{\pm}(b_n) = 1 - 1/n$. The terminal set is incompatible, but playing b_T for $T \geq 2$ gives total regret one in either model. The missing limit is a common oracle. This example isolates exactly the positive-gap step supplied by compactness and continuity.

C.3 Finite-Sample Diagnostic Bounds

Closure determines whether a reduction is possible; diagnostic complexity also depends on how weakly the accessible experiments separate models. We quantify this through controlled testing (Nitinawarat et al., 2013; Deshmukh et al., 2021). For the compact-class application, use the witness library $\mathcal{L}$: it preserves closures and oracle intersections, but the quantitative separations depend on the chosen witnesses. Squared Hellinger distance controls the testing error of repeated experiments. For probability laws P, Q, write

$$\mathsf{h}^2(P, Q) = 1 - \int \sqrt{dP\, dQ}.$$

The design should separate every candidate removable at the current truth, rather than spend samples on pairs that no available experiment distinguishes. Let $m = |\Theta|$ and $D_C(\theta) = C \setminus \mathcal{T}_{\mathsf{A},\theta}(C)$. If this set is nonempty, maximize the weakest separation over those candidates to define the truth-specific obedient escape modulus

$$\chi_{\mathsf{A}}(C, \theta) = \max_{\rho \in \Delta(\Pi^{\gamma}_{\mathsf{A}}(C))} \min_{q \in D_C(\theta)} \sum_{\pi} \rho(\pi) \mathsf{h}^2(K^{\pi}_{\theta}, K^{\pi}_{q}), \tag{HE}$$

and set it to zero otherwise. This design depends on the truth and therefore describes available information, not an executable choice for the learner. To obtain one allocation valid before the truth is known, minimize over all currently distinguishable ordered pairs:

$$\underline{\chi}_{\mathsf{A}}(C) = \max_{\rho \in \Delta(\Pi^{\gamma}_{\mathsf{A}}(C))} \min_{\substack{\theta \in C \\ q \in D_C(\theta)}} \sum_{\pi} \rho(\pi) \mathsf{h}^2(K^{\pi}_{\theta}, K^{\pi}_{q}), \tag{RHE}$$

with value zero when the indexed pair set is empty. Both are finite max–min linear programs.

Proposition 25 (Quantitative obedient escape)**.** *Under Assumption 22,*

$$\chi_{\mathsf{A}}(C, \theta) > 0 \quad \Longleftrightarrow \quad \mathcal{T}_{\mathsf{A},\theta}(C) \subsetneq C.$$

Under the embedding assumptions of Proposition 6, consequently, for $C^{\mathrm{pub}} = C_\infty^{\mathrm{pub}}(\theta^\star)$,

$$C_\infty^{\mathrm{priv}}(\theta^\star) \subsetneq C^{\mathrm{pub}} \quad \Longleftrightarrow \quad \chi_{\mathrm{priv}}(C^{\mathrm{pub}}, \theta^\star) > 0.$$

Suppose the compatibility condition in Theorem 23 holds. At every incompatible C, $\underline{\chi}_{\mathsf{A}}(C) > 0$. A block design attaining (RHE)*, followed by the likelihood update with block failure budget β, retains the truth and removes every member of $D_C(\theta)$ with probability at least $1-\beta$ after*

$$N_C = \left\lceil \frac{3}{2\underline{\chi}_{\mathsf{A}}(C)} \log \frac{2m}{\beta} \right\rceil \tag{HB}$$

episodes. Allocating $\beta = \delta/m$ across at most $m-1$ incompatible blocks gives

$$N_{\mathrm{diag}} \le \sum_{C \in \mathcal{V}_\theta} \left\lceil \frac{3}{2\underline{\chi}_{\mathsf{A}}(C)} \log \frac{2m^2}{\delta} \right\rceil \tag{HBC}$$

with probability at least $1-\delta$.

The bound follows from finite-hypothesis testing over the accessible experiments. Enlarging that set can add weakly separated candidates to the minimum defining $\underline{\chi}$, so greater access need not increase this truth-independent modulus. The strictness equivalence instead asks whether private experiments reveal any additional distinction at the public terminal set and a fixed truth.

A bound in terms of average KL information requires an additional step: the realized likelihood evidence must concentrate around its mean. Let ρ_C be the diagnostic mixture at an incompatible set $C \ni \theta$. For each removable $q \in D_C(\theta)$, define its one-episode log-likelihood increment and expected value by

$$X_{Cq} = \log \frac{k_\theta^\pi(Z)}{k_q^\pi(Z)},$$
$$\mu_{Cq} = \sum_\pi \rho_C(\pi) D_{\mathrm{KL}}(K_\theta^\pi \| K_q^\pi),$$

where $\pi \sim \rho_C$, $Z \sim K_\theta^\pi$, and put $d_C(\theta) = \min_{q \in D_C(\theta)} \mu_{Cq}$.

Corollary 26 (Finite-sample diagnostic complexity)**.** *Suppose the terminal-closure condition in Theorem 23 holds and, uniformly over relevant (θ, C, q),*

$$\log \mathbb{E}_\theta e^{\lambda(\mu_{Cq} - X_{Cq})} \le \frac{\lambda^2 v \mu_{Cq}}{2(1-b\lambda)}, \qquad 0 \le \lambda < 1/b. \tag{LT}$$

Here $v, b \ge 0$, with $1/0 = \infty$. Run the nested-likelihood algorithm with coverage budget $\delta/2$. Let $\mathcal{V}_\theta$ be the incompatible candidate sets visited before the first compatible set. With

$$A_\delta = \log \frac{2m}{\delta}, \qquad x_\delta = \log \frac{2m^2}{\delta},$$

we have, with probability at least $1-\delta$, simultaneous true-model coverage and

$$N_{\mathrm{diag}} \le \sum_{C \in \mathcal{V}_\theta} \left\lceil \frac{6A_\delta + 2(v+b)x_\delta}{d_C(\theta)} \right\rceil, \qquad |\mathcal{V}_\theta| \le m-1.$$

Condition (LT) controls the lower tail of accumulated evidence on the same scale as its mean, giving $1/d_C$ sample complexity. A generic centered-increment bound instead yields a $1/d_C^2$ fluctuation term. Expected KL separation by itself therefore does not justify the faster bound.

D Controlled-Markov Support and Observability

D.1 Pure Nash versus Correlated-Obedient Support

The strategic source of transition-row coverage can be isolated at a terminal complete-information stage. Let the finite joint-action space be $\mathcal{A} = \prod_i \mathcal{A}_i$, and let receiver utilities be $u_i^\theta(a)$. Receivers

know θ, whereas a type-blind mechanism does not. A public signal reveals an auxiliary message and the full pure joint-action label R; a private signal reveals only (M_i, R_i) to receiver i. Under prescribed obedience, $A = R$, an outcome is drawn as $Z \sim P_\theta(\cdot \mid A)$, and the learner logs the messages, A, and Z. Messages affect utilities and the physical kernel only through the executed action.

With a public joint label, each receiver knows the others' prescribed actions. An implementable profile must therefore resist every unilateral switch under each candidate model. Collect these profiles in the robust positive-margin Nash set

$$\mathsf{N}_C^\gamma = \{a : u_i^\theta(a) - u_i^\theta(a_i', a_{-i}) \geq \gamma \\ \forall \theta \in C, i, a_i' \neq a_i\}$$

Private labels instead leave uncertainty about the other actions, so obedience compares payoffs averaged over that uncertainty. Multiplying by each own-label probability gives the robust correlated-obedience polytope

$$g^\theta_{i,r_i \to a_i'}(p) = \sum_{a_{-i}} p(r_i, a_{-i}) \big[u_i^\theta(r_i, a_{-i}) \\ - u_i^\theta(a_i', a_{-i}) - \gamma\big], \tag{E}$$

$$\mathsf{CE}_C^\gamma = \{p \in \Delta(\mathcal{A}) : g^\theta_{i,r_i \to a_i'}(p) \geq 0 \\ \forall \theta \in C, i, r_i, a_i' \neq r_i\}.$$

An inequality associated with a zero-probability recommendation is $0 \geq 0$. Write $\mathrm{asupp}(\mathsf{CE})$ for the union of the action supports of distributions in CE.

The equilibrium identities below are standard; the row characterization connects their support difference to identification of a fixed controlled kernel. The pure-Nash description requires the realized joint pure label to be public. An unrevealed mixed prescription need not have pure-Nash support, and a hidden payoff state requires a Bayes-correlated formulation.

A two-player congested-entry game gives a non-coordination witness. With rows indexed by a_1 and columns by a_2, let

$$u_1 = \begin{pmatrix} 2 & 1 \\ 3 & 0 \end{pmatrix}, \qquad u_2 = \begin{pmatrix} 2 & 3 \\ 1 & 0 \end{pmatrix}.$$

For $0 \leq \gamma \leq 1/3$, the only public profiles are 01 and 10, whereas the private distribution

$$(p_{00}, p_{01}, p_{10}, p_{11}) = (0, 2/5, 2/5, 1/5)$$

has minimum conditional obedience gap $1/3$. If $P_\theta(Z = 1 \mid a) = 1/2 + \theta \mathbf{1}\{a = 11\}$ for $\theta \in \{-\Delta, +\Delta\}$, $0 < \Delta < 1/2$, every public experiment has the same law under both models, while the displayed private experiment distinguishes them through its mass on 11.

Proposition 27 (Equilibrium support determines observable rows)**.** *In this terminal protocol, suppose a public signal reveals the realized pure joint-action label and a private signal reveals only each receiver's component. Finite type-blind auxiliary messages may accompany the labels but affect utilities and the physical kernel only through A. The learner logs these signals, A, and $Z \sim P_\theta(\cdot \mid A)$. Under prescribed obedience, the robustly implementable action laws are*

$$\mathsf{E}^\gamma_{\mathrm{pub}}(C) = \Delta(\mathsf{N}_C^\gamma), \qquad \mathsf{E}^\gamma_{\mathrm{priv}}(C) = \mathsf{CE}_C^\gamma.$$

Hence

$$\mathcal{T}_{\mathsf{A},\theta^\star}(C) = \{\theta \in C : P_\theta(\cdot \mid a) = P_{\theta^\star}(\cdot \mid a) \\ \forall a \in \mathsf{X}^\gamma_{\mathsf{A}}(C)\},$$

where $\mathsf{X}^\gamma_{\mathrm{pub}} = \mathsf{N}_C^\gamma$ and $\mathsf{X}^\gamma_{\mathrm{priv}} = \mathrm{asupp}(\mathsf{CE}_C^\gamma)$. If the private class contains replicated public signals, then at C^{pub} private refinement is strict exactly when some $a \in \mathsf{X}^\gamma_{\mathrm{priv}}(C^{\mathrm{pub}})$ has a transition row that differs between $\theta^\star$ and a model in C^{pub}.

Proof of Proposition 27. Fix a public realization $(\mathfrak{m}, a)$ with positive probability. Receiver i observes a_{-i}, so robust conditional obedience is exactly

$$u_i^\theta(a) - u_i^\theta(a_i', a_{-i}) \geq \gamma \quad \forall \theta \in C, i, a_i' \neq a_i.$$

Every action in the induced support therefore belongs to N_C^γ. Conversely, drawing any distribution on N_C^γ and revealing the realized label implements that distribution publicly.

For a private mechanism, multiply the obedience inequality conditional on $(M_i, R_i) = (m_i, r_i)$ by the probability of that signal and sum over m_i. For every θ, i, r_i, a_i', this gives

$$\sum_{a_{-i}} p(r_i, a_{-i})\big[u_i^\theta(r_i, a_{-i}) - u_i^\theta(a_i', a_{-i}) - \gamma\big] \geq 0.$$

The identity remains valid when $R_i = r_i$ has zero probability. Hence the induced action law lies in CE_C^γ. Conversely, draw $R \sim p \in \mathsf{CE}_C^\gamma$, use a constant auxiliary message, and send R_i privately. Dividing each positive-mass inequality by $\mathbb{P}(R_i = r_i)$ proves conditional γ-obedience. This establishes the two implementable-law identities.

For any allowed mechanism, type blindness and prescribed obedience give the logged-law factorization

$$K_\theta(dm, da, dz) = \mu(dm, da) P_\theta(dz \mid a),$$

where μ has no model argument and a is observed. Two logged laws are equal exactly when their controlled kernels agree at every action in the mechanism's positive support. Intersecting over all mechanisms and applying the implementable-law identities proves the closure formula.

Finally, Proposition 19 gives

$$C_\infty^{\mathrm{priv}}(\theta^\star) = C^{\mathrm{pub}} \iff \mathcal{T}_{\mathrm{priv},\theta^\star}(C^{\mathrm{pub}}) = C^{\mathrm{pub}}.$$

Substituting the closure formula shows that equality holds exactly when every candidate in C^{pub} agrees with the truth on every action in $\mathrm{asupp}(\mathsf{CE}_{C^{\mathrm{pub}}}^\gamma)$. Negating this condition proves the strict-refinement statement. □

D.2 Sequential Markov Support and Continuation Games

A current action can be obedient because of what follows a deviation, even if its immediate payoff is worse. Applying the terminal support identity stagewise would miss this incentive. Let $H = 2$, $\delta = 1$, and $\gamma = 1/4$. Suppose $u_{i,1}(a) = a_i/2$, while both receivers have terminal payoff matrix

$$u_{i,2} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}.$$

A public mechanism recommends 00 initially, then recommends 00 after obedience and 11 after a first-stage deviation. Following yields two; deviating and optimizing thereafter yields $1/2 + 1 = 3/2$. The mechanism is dynamically γ-obedient even though 00 violates both the Nash and correlated-obedience inequalities of the first-stage game. Recommending terminal 00 after every history preserves the prescribed trajectory but makes the initial deviation profitable. Dynamic accessibility therefore depends on counterfactual continuation design, not only the on-path law.

To recover a stagewise test, include future incentives in the action value rather than apply the terminal criterion to immediate utility. For this subsection, a finite public state contains all payoff-relevant history and mechanism memory. A type-blind mechanism draws $R_h \sim p_h^\pi(\cdot \mid S_h)$, carries no persistent private signal, and uses the same Markov kernels after every action history. Require obedience at every state and positive-probability recommendation, including after unilateral deviations. This lets backward induction evaluate a current deviation using obedient future values. With expected stage utility $\bar{u}_{i,h}^\theta$, define

$$\begin{aligned} v_{i,H+1}^{\pi,\theta}(s) &= 0, \\ q_{i,h}^{\pi,\theta}(s,a) &= \bar{u}_{i,h}^\theta(s,a) + \delta \sum_{s'} P_{\theta,h}(s' \mid s, a) v_{i,h+1}^{\pi,\theta}(s'), \\ v_{i,h}^{\pi,\theta}(s) &= \sum_a p_h^\pi(a \mid s) q_{i,h}^{\pi,\theta}(s,a). \end{aligned} \tag{D}$$

Proposition 28 (Sequential Markov support recursion)**.** *The Markov mechanism above is sequentially γ-obedient for every $\theta \in C$ if and only if the following conditions hold for every θ, i, h, s. Publicly, each a in the support of $p_h^\pi(\cdot \mid s)$ satisfies*

$$q_{i,h}^{\pi,\theta}(s,a) - q_{i,h}^{\pi,\theta}(s, a_i', a_{-i}) \geq \gamma \quad \forall a_i' \neq a_i. \tag{DN}$$

Privately, for every r_i and $a_i' \neq r_i$,

$$\sum_{a_{-i}} p_h^{\pi}(r_i, a_{-i} \mid s)\big[q_{i,h}^{\pi,\theta}(s, r_i, a_{-i}) - q_{i,h}^{\pi,\theta}(s, a_i', a_{-i}) - \gamma\big] \geq 0. \tag{DC}$$

Proof of Proposition 28. We use backward induction. At stage H, the continuation value is zero, so (DN) and (DC) are the public and private terminal obedience constraints.

Suppose the equivalence holds from stage $h+1$ onward. Because the local conditions apply after every unilateral history, the induction hypothesis implies that no causal future deviation improves on following the Markov plan. After a current unilateral deviation, the optimal future value is therefore the obedient continuation in (D).

Under a public label, the receiver knows a_{-i}. Comparing the prescribed action with each alternative gives (DN). Under a private label r_i, average the same comparison over $p_h^{\pi}(a_{-i} \mid r_i, s)$, multiply by $\mathbb{P}(R_{i,h} = r_i \mid S_h = s)$, and obtain (DC). A zero-mass recommendation yields $0 \geq 0$.

These arguments prove necessity. Conversely, (DN) or (DC) makes the first action of every current deviation lose at least γ, while the induction hypothesis rules out a gain from its causal continuation. Hence no multi-stage deviation is profitable. Backward induction proves sequential γ-obedience under every $\theta \in C$. □

At the terminal stage, (DN)–(DC) reduce to Proposition 27. Earlier continuation games depend jointly on the mechanism and model. Combining their recursively feasible occupancy supports with Proposition 31 gives the dynamic identifiable rows. This is a one-step-deviation decomposition; persistent private histories or hidden payoff states instead require an extensive-form or Bayes-correlated obedience system.

D.3 Dynamic Continuation-Punishment Separation

Private continuation recommendations can make an initially unattractive action obedient by reducing the payoff after deviation. To isolate this effect, let the informative first-stage action forgo an immediate gain, and let terminal utility reward coordination. Fix undiscounted utilities $\delta = 1$ and $0 < \gamma < 1/3$, let

$$g = \frac{1-3\gamma}{2}, \qquad r = \frac{1-\gamma}{1+\gamma},$$

and consider two stages with $u_{i,1}(a) = g a_i$ and $u_{i,2}(a) = \mathbf{1}\{a_i = a_j\}$. Joint actions are publicly monitored. After the first action, the learner observes

$$Y \mid A_1 \sim \text{Bernoulli}\left(\frac{1}{2} + \theta \mathbf{1}\{A_1 = 00\}\right), \qquad \theta \in \{-\Delta, +\Delta\}, \qquad 0 < \Delta < \frac{1}{2}. \tag{DP}$$

Proposition 29 (Dynamic continuation-punishment separation)**.** *Every dynamically γ-obedient public pure-label mechanism has $\mathbb{P}(A_1 = 00) = 0$. There is a dynamically γ-obedient private mechanism with $\mathbb{P}(A_1 = 00) = 1$. Hence all public observable laws in* (DP) *are model-independent, while the private mechanism distinguishes the two models.*

The private mechanism recommends 00 initially and rewards obedience with terminal 00. After a deviation it draws private terminal labels from

$$p^{\text{pun}} = \frac{(1, r, r, 1)}{2(1+r)}. \tag{PUN}$$

The result uses prescribed-obedience selection: private correlation lowers the payoff available after a deviation, whereas a public terminal label lets the deviator match the other receiver.

Proof of Proposition 29. Suppose a public mechanism uses the first-stage label 00 at a positive-probability on-path history. At the terminal stage, every obedient public pure label is matched, so following gives continuation payoff one. After a first-stage deviation, the deviator observes the other receiver's public terminal label and can match it, again obtaining one. Switching the first action from zero to one therefore gains $g > 0$, contradicting dynamic obedience. Thus every public obedient mechanism has $\mathbb{P}(A_1 = 00) = 0$.

For the private construction, recommend 00 initially and terminal 00 after obedience. After any first-stage deviation use (PUN). Conditional on either private terminal label, the other label matches with probability

$$\frac{1}{1+r} = \frac{1+\gamma}{2},$$

so following the terminal label beats switching by γ. The best terminal payoff after a first-stage deviation is therefore $(1+\gamma)/2$, and the first-stage obedience gap is

$$1 - \left(g + \frac{1+\gamma}{2}\right) = \gamma.$$

The on-path terminal label 00 has gap one, proving dynamic γ-obedience.

The public exclusion makes $Y \sim \mathrm{Bernoulli}(1/2)$ under both models. The private mechanism instead yields $Y \sim \mathrm{Bernoulli}(1/2 + \theta)$, completing the separation. □

This continuation argument uses prescribed following. The same utility family also admits an uninformative equilibrium that ignores messages, so the access gain does not extend to arbitrary equilibrium selection.

Corollary 30 (Selection barrier for (DP))**.** *For every type-blind public or private mechanism in* (DP)*, the receivers have a strict sequential equilibrium in which they ignore all messages, play* 11 *initially, and play* 00 *at the terminal stage after every history. The induced observable law is the same under both models. Consequently, no mechanism in this utility family guarantees identification under adversarial equilibrium selection.*

Proof of Corollary 30. Fix any allowed mechanism. At every terminal information set, if the other receiver ignores its message and plays zero, playing zero yields coordination payoff one and playing one yields zero. Thus terminal action zero is a strict best response after every history, independently of beliefs and messages.

Suppose both receivers use this terminal strategy and the other receiver plays one initially. Either initial action leads to terminal coordination payoff one, while initial action one adds the stage payoff $g > 0$. Hence playing one initially is also a strict best response at every information set. A consistent belief system exists by taking limits of completely mixed strategy profiles; sequential rationality holds for every such belief system because both comparisons are strict and belief-independent. The stated message-ignoring profile is therefore a strict sequential equilibrium.

Along this equilibrium, $A_1 = 11$, so Y is $\mathrm{Bernoulli}(1/2)$ under both models. Type blindness then implies by causal factorization that all mechanism messages have the same law under both models, and the selected actions are deterministic. The complete observable laws coincide. An adversarial selector may choose this equilibrium under both models for every mechanism, precluding guaranteed identification. □

D.4 Controlled-Markov Occupancy Coverage

When states and actions are logged, the learner can attribute each outcome to the transition row that produced it. This converts equality of trajectory laws into agreement on visited rows. Let state and joint-action spaces be finite, let the initial-state law be model-independent, and suppose the learner logs $(R_{1:H}, A_{1:H}, S_{1:H+1}, F_{2:H+1})$. Denote a row's next-state and feedback law by

$$Q_{\theta,h}(\cdot \mid s, a) = \mathcal{L}_\theta\big((S_{h+1}, F_{h+1}) \;\big|\; S_h = s, A_h = a\big).$$

Mechanisms are model-independent causal direct kernels conditioned only on the logged history, and receivers follow their recommendations. Thus the mechanism contributes no separate conditional likelihood ratio; model information enters through the visited physical rows. For $\pi \in \Pi^{\gamma}_{\mathsf{A}}(C)$, write a row's visit probability as

$$d^{\pi}_{\theta,h}(s,a) = \mathbb{P}^{\pi}_{\theta}(S_h = s, A_h = a)$$

and collect the rows that some currently robust mechanism visits with positive probability:

$$\mathsf{X}_{\mathsf{A}}(C,\theta) = \{(h,s,a) : \exists \pi \in \Pi^{\gamma}_{\mathsf{A}}(C), \\ d^{\pi}_{\theta,h}(s,a) > 0\}.$$

The same row coverage answers two different questions. For a finite model class, a visited row can rule out a candidate whose transition law differs. For a smooth parameter family, row Fisher information instead describes first-order sensitivity; its common nullspace collects directions with zero local information under every obedient experiment. The proposition states both conclusions under their respective assumptions.

Proposition 31 (Controlled-Markov occupancy closure and rank)**.** *For every $\theta, \theta' \in C$ and every robust mechanism π,*

$$D_{\mathrm{KL}}(K^{\pi}_{\theta} \| K^{\pi}_{\theta'}) = \sum_{h,s,a} d^{\pi}_{\theta,h}(s,a) \\ \times D_{\mathrm{KL}}\big(Q_{\theta,h}(\cdot \mid s,a) \\ \big\| \, Q_{\theta',h}(\cdot \mid s,a)\big). \tag{M}$$

Consequently,

$$\mathcal{T}_{\mathsf{A},\theta}(C) = \{\theta' \in C : \; Q_{\theta',h}(\cdot \mid s,a) \\ = Q_{\theta,h}(\cdot \mid s,a) \\ \forall (h,s,a) \in \mathsf{X}_{\mathsf{A}}(C,\theta)\}. \tag{MC}$$

For finite Θ, let $C^{\mathrm{pub}} = C^{\mathrm{pub}}_{\infty}(\theta^{\star})$. Under the embedding in Proposition 6,

$$C^{\mathrm{priv}}_{\infty}(\theta^{\star}) \subsetneq C^{\mathrm{pub}}$$

if and only if some $\theta \in C^{\mathrm{pub}}$ differs from $\theta^{\star}$ on a row in $\mathsf{X}_{\mathrm{priv}}(C^{\mathrm{pub}}, \theta^{\star})$.

Suppose additionally that $\theta = \eta \in \mathbb{R}^r$ and the rows form a regular dominated differentiable family. Let $\mathcal{I}_{\eta,h}(s,a)$ be the Fisher information of row (h,s,a). Then

$$\mathcal{I}^{\pi}_{\eta} = \sum_{h,s,a} d^{\pi}_{\eta,h}(s,a)\mathcal{I}_{\eta,h}(s,a), \\ \mathsf{N}_{\mathsf{A}}(C,\eta) := \bigcap_{\pi \in \Pi^{\gamma}_{\mathsf{A}}(C)} \ker \mathcal{I}^{\pi}_{\eta}, \\ \mathsf{N}_{\mathsf{A}}(C,\eta) = \bigcap_{(h,s,a) \in \mathsf{X}_{\mathsf{A}}(C,\eta)} \ker \mathcal{I}_{\eta,h}(s,a). \tag{MI}$$

Hence private local information rank is at least public rank. The inequality is strict exactly when a privately covered row has positive information in some direction in the common public nullspace. A finite mixture of at most r obedient mechanisms realizes the common nullspace whenever the rank is positive.

The decomposition into visitation-weighted information is standard in Markov experiment design (Mutny et al., 2023), and transition coverage is central to active model estimation (Tarbouriech et al., 2020; Gu et al., 2026). The architecture-specific content is which rows are obediently reachable. Occupancy support, not its affine direction span, is the correct object because a single fixed experiment can already have positive information. Under arbitrary partial state feedback, physical-row coverage is not exact and the general observable law K^{π}_{θ} must be retained.

Proof of Proposition 31. For a prescribed-obedience trajectory z, its density has the causal factorization

$$k^{\pi}_{\theta}(z) = \lambda(s_1) \prod_{h=1}^{H} \big[\mu^{\pi}_{h}(r_h \mid z_{<h}, s_h)\mathbf{1}\{a_h = r_h\} \\ \times q_{\theta,h}(s_{h+1}, f_{h+1} \mid s_h, a_h)\big],$$

where λ and the recommendation kernels μ_h^π have no model argument. Put

$$D_h(s,a) := D_{\mathrm{KL}}(Q_{\theta,h}(\cdot \mid s,a) \| Q_{\theta',h}(\cdot \mid s,a)) .$$

The conditional KL chain rule gives

$$D_{\mathrm{KL}}(K_\theta^\pi \| K_{\theta'}^\pi) = \sum_{h=1}^{H} \mathbb{E}_\theta^\pi [D_h(S_h, A_h)],$$

which is (M) after conditioning on (S_h, A_h). The identity holds in the extended nonnegative reals, so no reverse absolute continuity is needed.

The right side is zero exactly when the two row laws agree at every truth-positive occupancy. Thus, for a fixed π, $K_\theta^\pi = K_{\theta'}^\pi$ exactly under this row-agreement condition. Intersecting over all mechanisms in $\Pi_{\mathsf{A}}^\gamma(C)$ proves (MC).

The replicated-public embedding preserves the complete prescribed trajectory law and hence every occupancy, so

$$\mathsf{X}_{\mathrm{pub}}(C,\theta) \subseteq \mathsf{X}_{\mathrm{priv}}(C,\theta).$$

At C^{pub}, (MC) says that one private closure step removes a candidate exactly when it differs from the truth on a privately covered row. Proposition 19 identifies this one-step condition with strict terminal refinement, proving the global architecture claim.

For the local statement, let

$$g_{\eta,h} = \nabla_\eta \log q_{\eta,h}(S_{h+1}, F_{h+1} \mid S_h, A_h).$$

Differentiating the causal factorization shows that the trajectory score is $\sum_h g_{\eta,h}$. Regularity gives $\mathbb{E}_\eta[g_{\eta,h} \mid S_h, A_h] = 0$. The score increments are therefore martingale differences, and all cross terms vanish:

$$\begin{aligned} \mathcal{I}_\eta^\pi &= \mathbb{E}_\eta^\pi \left[\left(\sum_h g_{\eta,h} \right) \left(\sum_h g_{\eta,h} \right)^\top \right] \\ &= \sum_{h,s,a} d_{\eta,h}^\pi(s,a) \mathcal{I}_{\eta,h}(s,a). \end{aligned}$$

For positive-semidefinite matrices, the kernel of a positively weighted sum is the intersection of the kernels of its summands. Taking the intersection over mechanisms gives (MI). Public-row coverage is contained in private-row coverage, so the private nullspace is contained in the public one. Strict containment holds exactly when some v in the public nullspace has $v^\top \mathcal{I}_{\eta,h}(s,a) v > 0$ on a privately covered row.

Finally, start with $\mathbb{R}^r$ and add information matrices of obedient mechanisms whenever doing so strictly reduces the current intersection of kernels. Each addition reduces its dimension by at least one, so at most r mechanisms are needed when the final rank is positive. Any mixture assigning positive mass to the selected mechanisms has information matrix with the common nullspace. □

E Equilibrium Selection and Alternative Mechanisms

E.1 Adversarial Equilibrium Feedback

Under equilibrium selection, a mechanism supplies a set of possible feedback laws rather than one fixed law. A diagnostic must make progress for every law the selector can choose. For a given law, its *fiber* is the set of models that can produce it; successful learning requires every possible successor fiber to admit further progress or a common robust oracle. Feedback and payoff come from the same selected equilibrium, so we begin with their joint laws and then extract the observable marginal and guaranteed payoff. Convexification allows randomized selection. The recursion below characterizes worst-equilibrium guarantee regret by combining controlled composite testing (Nitinawarat et al., 2013; Deshmukh et al., 2021) with belief-support reachability (Chatterjee & Doyen, 2012).

Proposition 32 (Finite equilibrium-law selection game)**.** *Let the model, mechanism, and feedback alphabets be finite, and let the sender payoff $G \in [-B, B]$. Let $\mathcal{E}_\theta^\pi$ be the nonempty compact convex hull of joint laws of (Z, G) induced by all admissible equilibrium selections after π. For $\nu \in \mathcal{E}_\theta^\pi$, put $j_\theta(\nu) = \mathbb{E}_\nu G$, and define*

$$\begin{aligned} \mathcal{Q}_\theta^\pi &= \{\nu_Z : \nu \in \mathcal{E}_\theta^\pi\}, \\ \underline{J}_\theta(\pi) &= \inf_{\nu \in \mathcal{E}_\theta^\pi} j_\theta(\nu), \\ V_\theta^{\mathrm{rob}} &= \max_{\pi \in \Pi^\gamma(\theta)} \underline{J}_\theta(\pi), \quad \mathcal{O}_\theta^{\mathrm{rob}} = \arg\max_{\pi \in \Pi^\gamma(\theta)} \underline{J}_\theta(\pi). \end{aligned}$$

Define

$$F_C(\pi, q) = \{\theta \in C : q \in \mathcal{Q}_\theta^\pi\}, \qquad q \in \bigcup_{\theta \in C} \mathcal{Q}_\theta^\pi.$$

Let $\mathcal{F}_C(\pi)$ collect its nonempty fibers. Let $\mathcal{W}_0$ contain the nonempty sets C satisfying $\cap_{\theta \in C} \mathcal{O}_\theta^{\mathrm{rob}} \neq \varnothing$, and recursively put

$$\mathcal{W}_{k+1} = \mathcal{W}_k \cup \{C : \exists \pi \in \Pi^\gamma(C),\ \mathcal{F}_C(\pi) \subseteq \mathcal{W}_k\}. \tag{31}$$

For the selector-robust guarantee regret

$$R_T^{\mathrm{gar}}(\theta) = \sum_{t=1}^{T} \big[V_\theta^{\mathrm{rob}} - \underline{J}_\theta(\pi_t)\big],$$

sublinear expectation with obedience failure probability $o(1)$, uniformly over adaptive equilibrium selectors, is possible if and only if $\Theta \in \mathcal{W}_\infty$. When it holds, for every $\delta > 0$ there is an algorithm with failure probability at most δ and regret $O(\log(T/\delta)) + O(\delta T)$. Constants may depend on the finite instance. The same upper bound holds for selector-realized positive-part regret: on the true-model obedience event, $[V_\theta^{\mathrm{rob}} - j_\theta(\nu_t)]_+ \le V_\theta^{\mathrm{rob}} - \underline{J}_\theta(\pi_t)$, and the failure event contributes at most $2B\delta T$ in expectation. Deterministic feedback is the point-mass special case.

Proof. The proof first makes exact fiber reductions tolerant to sampling error, then tests them under adaptive selection. For necessity, a minimal nonwinning set supplies a common feedback law. First, every $\mathcal{W}_k$ is downward closed. The base case follows from hereditary robust decision compatibility. Inductively, if $D \subseteq C$, then $F_D(\pi, q) = D \cap F_C(\pi, q)$, so a witness for C also witnesses D.

Fix a rank-$(k+1)$ witness (C, π). Compactness and downward closure give an $\epsilon_{C,\pi} > 0$ such that, for every $q \in \cup_{\theta \in C} \mathcal{Q}_\theta^\pi$,

$$\{\theta \in C : d_1(q, \mathcal{Q}_\theta^\pi) \le 2\epsilon_{C,\pi}\} \in \mathcal{W}_k. \tag{32}$$

Otherwise take $\epsilon_n \downarrow 0$ and bad q_n. A convergent subsequence and then a constant bad fiber D satisfy $D \subseteq F_C(\pi, \lim_n q_n) \in \mathcal{W}_k$, a contradiction.

For sufficiency, repeat π for n episodes. Conditional on the past, the selector chooses $\nu_t \in \mathcal{E}_{\theta^\star}^\pi$, whose feedback marginal q_t belongs to $\mathcal{Q}_{\theta^\star}^\pi$, so $\bar{q}_n = n^{-1} \sum_t q_t$ remains in that convex set. If d_Z is the feedback-alphabet size, coordinatewise martingale Hoeffding gives

$$\mathbb{P}\{\|\widehat{q}_n - \bar{q}_n\|_1 > \epsilon\} \le 2m \exp(-2n\epsilon^2/d_Z^2). \tag{33}$$

Take $n = O(\epsilon_{C,\pi}^{-2} \log(d_Z |\Theta| T/\delta))$ and retain models within $\epsilon_{C,\pi}$ of $\widehat{q}_n$. On (33), the truth remains and the retained set lies in the winning robust fiber (32). Rank decreases, so at most $|\Theta| - 1$ blocks reach a common robust oracle. Guarantee loss is bounded during the diagnostic blocks and zero thereafter. A split failure budget proves the upper bound and obedience claim.

For necessity, take an inclusion-minimal nonwinning set C. For every $\pi \in \Pi^\gamma(C)$, failure of (31) and minimality give

$$q_\pi \in \bigcap_{\theta \in C} \mathcal{Q}_\theta^\pi.$$

For each model, the selector chooses a joint law in $\mathcal{E}_\theta^\pi$ whose feedback marginal is q_π, stopped if the learner leaves $\Pi^\gamma(C)$. Thus all models in C have one common adaptive feedback-history law

while the learner is uniformly obedient. If the robust set is nonempty, finiteness and the absence of a common robust oracle give

$$g_C = \min_{\pi \in \Pi^\gamma(C)} \sum_{\theta \in C} [V_\theta^{\mathrm{rob}} - \underline{J}_\theta(\pi)] > 0.$$

On the common-law event, summed guarantee regret is at least $g_C T$. If per-model obedience failure is $\delta_T = o(1)$, a union bound and bounded payoffs change this by only $o(T)$. Hence $\max_{\theta \in C} \mathbb{E}_{\theta,\sigma_\theta} R_T^{\mathrm{gar}}(\theta) \geq g_C T/|C| - o(T)$ for the constructed selectors σ_θ. Since C is finite, this contradicts $o(T)$ expected regret for every model. If the robust set is empty, uniform obedience is impossible directly. □

Remark 33 (A selector may be unable to hide and harm simultaneously). The equivalence in Proposition 32 is specific to guarantee regret. It remains sufficient, but is not necessary, for selector-realized positive-part regret. Consider two models with robust values $V_1^{\mathrm{rob}} = V_2^{\mathrm{rob}} = 1$. A diagnostic d has deterministic joint outcomes

$$\mathcal{E}_1^d = \{(z_0, 1), (z_1, 0)\}, \qquad \mathcal{E}_2^d = \{(z_0, 1), (z_2, 0)\},$$

where convexification is understood. Two further mechanisms a_1, a_2 always reveal z_0; mechanism a_i pays one in model i and zero in the other model. All three mechanisms may be taken as obedient. The robust oracle sets are disjoint, and every mechanism admits the common feedback law δ_{z_0}, so the pair is a selection trap under (31).

Nevertheless, repeat d until z_1 or z_2 occurs, then use the corresponding a_i. Every hidden outcome z_0 pays one, while the first zero-payoff outcome identifies the model. Against any adaptive randomized selector, expected positive-part regret is at most one. In contrast, $\underline{J}_i(d) = 0$, so a selector that always uses $(z_0, 1)$ still incurs one unit of guarantee regret per round. An exact characterization of realized regret must therefore retain the joint feedback–payoff law and allow zero-loss self-loops; the strict fiber recursion deliberately avoids this larger partial-monitoring game.

A common feedback law lets the selector prevent any reduction of the current candidate set. The next criterion identifies when such an obstruction prevents reaching a prescribed decision or identification target.

Corollary 34 (Selection-trap boundary)**.** *Fix an architecture and let $\mathcal{G}$ be the common downward-closed target family used as $\mathcal{W}_0$ in* (31)*. A nonterminal set $C \notin \mathcal{G}$ is a selection trap if*

$$\bigcap_{\theta \in C} \mathcal{Q}_\theta^\pi \neq \varnothing \qquad \textit{for every } \pi \in \Pi^\gamma(C). \tag{34}$$

Then $\Theta \in \mathcal{W}_\infty$ if and only if no nonempty nonterminal subset of Θ is a selection trap. In particular, for two models and the singleton target family, that target is winnable if and only if some robust mechanism has disjoint modelwise equilibrium-law sets.

Proof. A trap cannot acquire positive rank: the common law in (34) has fiber C, so every candidate witness has a successor that has not appeared at an earlier rank. Downward closure then implies that if any trap exists, $\Theta \notin \mathcal{W}_\infty$. Conversely, if Θ is nonwinning, choose an inclusion-minimal nonwinning subset C. For every robust π, failure of the recursion gives a nonwinning fiber contained in C. Minimality makes that fiber equal to C, which supplies a common law in (34). For two models, the only nonterminal set is the full pair, and (34) fails exactly when the two compact law sets are disjoint for some robust mechanism. □

The same common-law obstruction has an information-theoretic interpretation. For a prior with positive mass on every retained model, ask how much mutual information is guaranteed after minimizing over model-wise selector choices.

Proposition 35 (Zero selector-robust epistemic information)**.** *Let b have finite support Θ_b, with $b(\theta) > 0$ for every $\theta \in \Theta_b$, and let the feedback alphabet be finite. For a fixed mechanism π, suppose every $\mathcal{Q}_\theta^\pi \subseteq \Delta(\mathcal{Z})$ is nonempty and compact. Define*

$$\underline{\mathrm{EIG}}_b(\pi) = \inf_{q_\theta \in \mathcal{Q}_\theta^\pi,\ \theta \in \Theta_b} I_b(\Theta; Z), \qquad Z \mid \Theta = \theta \sim q_\theta. \tag{35}$$

Then

$$\underline{\mathrm{EIG}}_b(\pi) = 0 \iff \bigcap_{\theta \in \Theta_b} \mathcal{Q}_\theta^\pi \neq \varnothing. \tag{36}$$

Proof. A common law makes Θ and Z independent. Conversely, compactness and continuity of entropy on a finite simplex make the infimum attainable. At a minimizer,

$$I_b(\Theta; Z) = H\left(\sum_\theta b(\theta) q_\theta\right) - \sum_\theta b(\theta) H(q_\theta)$$

is zero only if every positive-prior conditional law q_θ equals their mixture. Those minimizing laws therefore share one common element. □

This proposition identifies the common-law obstruction with zero guaranteed one-step information, but positive EIG is not exact identification. For example, $\mathcal{Q}_1 = \mathcal{Q}_2 = \{\delta_0\}$ and $\mathcal{Q}_3 = \{\delta_1\}$ have positive EIG under a full-support prior, although models 1 and 2 remain indistinguishable. The fiber recursion in Proposition 32 retains this distinction. Compactness is also necessary for the exact zero test: the disjoint sets

$$\mathcal{Q}_0 = \{\mathrm{Bern}(1/2)\},$$
$$\mathcal{Q}_1 = \{\mathrm{Bern}(1/2 + 1/n) : n \geq 3\}$$

have EIG infimum zero.

To compare architectures under selection, simulation must preserve the entire equilibrium-law correspondence, not just one prescribed law. We also hold the target family fixed, so adding mechanisms cannot change what counts as successful learning.

Proposition 36 (Fixed-target monotonicity under equilibrium selection)**.** *Let $\mathcal{G}$ be a common downward-closed family of terminal model sets, and construct $\mathcal{W}_{\mathsf{A},k}(\mathcal{G})$ from* (31) *with $\mathcal{W}_{\mathsf{A},0} = \mathcal{G}$ and architecture-specific robust mechanisms. Suppose every public mechanism π has a private simulation $E\pi$ such that, for every C and θ,*

$$\pi \in \Pi^\gamma_{\mathrm{pub}}(C) \Longrightarrow E\pi \in \Pi^\gamma_{\mathrm{priv}}(C),$$
$$\mathcal{Q}^{E\pi}_{\mathrm{priv},\theta} = \mathcal{Q}^{\pi}_{\mathrm{pub},\theta}.$$

Then

$$\mathcal{W}_{\mathrm{pub},k}(\mathcal{G}) \subseteq \mathcal{W}_{\mathrm{priv},k}(\mathcal{G}) \quad \forall k,$$
$$\mathcal{W}_{\mathrm{pub},\infty}(\mathcal{G}) \subseteq \mathcal{W}_{\mathrm{priv},\infty}(\mathcal{G}).$$

In particular, the conclusion holds for the singleton terminal family used for exact identification.

Proof. The rank-zero families agree. Suppose the inclusion holds at rank k and a public mechanism π witnesses membership of C at rank $k+1$. Its private simulation is robust on the same set and has exactly the same modelwise law correspondences. The public and private fibers are therefore identical, and every successor belongs to $\mathcal{W}_{\mathrm{priv},k}(\mathcal{G})$ by induction. Thus $E\pi$ witnesses private membership at rank $k+1$. □

The common terminal family is essential. If the public class contains only a, while the private class adds b_0, b_1, let all feedback laws be model-independent and set

$$J_0(b_0) = J_1(b_1) = 1,$$
$$J_0(a) = J_1(a) = J_0(b_1) = J_1(b_0) = 0.$$

Public own-oracle regret is zero, but the private modelwise oracles are disjoint and every algorithm has worst-case regret at least $T/2$. Hence the proposition does not order regret against architecture-specific oracle benchmarks.

Pairwise overlap does not tell us which models can remain possible together. In particular, two models may share a law and a common oracle even though no oracle works for the full candidate set. To see why this matters for learning, let one diagnostic have convex law sets

$$\mathcal{Q}_1 = \mathrm{conv}\{\delta_a, \delta_b\}, \quad \mathcal{Q}_2 = \mathrm{conv}\{\delta_b, \delta_c\},$$
$$\mathcal{Q}_3 = \mathrm{conv}\{\delta_c, \delta_a\}.$$

while the oracle sets are

$$\mathcal{O}_1 = \{x, y\}, \quad \mathcal{O}_2 = \{y, z\}, \quad \mathcal{O}_3 = \{z, x\}.$$

Every pair of law sets intersects, and the triple oracle intersection is empty. Nevertheless, every law fiber is either a singleton or one of the pairs at $\delta_a, \delta_b, \delta_c$, which admit common oracles x, y, z, respectively. Thus $\{1, 2, 3\} \in \mathcal{W}_1$, contradicting the naive pairwise-overlap criterion.

E.2 State-Gated Dynamic Separation

Pure coordination utilities suffice for the two-gate separation when receivers follow obedient recommendations. This variant isolates the support restriction without the hidden payoff targets used for selection robustness. Its conclusions below concern prescribed-obedience laws and pseudo-regret.

Proposition 37 (Prescribed-obedience state-gated separation)**.** *In* (1)*, let receivers observe S_2, let $0 \le \gamma < \beta \le 1$, and give them utility*

$$U_i = \mathbf{1}\{A_{i,1} = A_{j,1}\} + \beta \mathbf{1}\{A_{i,2} = A_{j,2}\}.$$

For one-way, type-blind direct labels under prescribed obedience,

$$C_\infty^{\mathrm{pub}}(\theta) = \{-\Delta, +\Delta\}, \qquad C_\infty^{\mathrm{priv}}(\theta) = \{\theta\}.$$

A private diagnostic has squared Hellinger separation

$$\frac{q(1-\gamma)(\beta-\gamma)}{4\beta}\left(1 - \sqrt{1-4\Delta^2}\right).$$

For the sender criterion in Section 4, minimax pseudo-regret is ΔT publicly and $\Theta(\log T)$ privately for fixed parameters.

Proof of the structural claim in Proposition 37. First consider an arbitrary dynamically obedient public mechanism. At the second stage, a public mismatch reveals the other receiver's prescribed action; switching to it raises the weighted coordination payoff from zero to β. A public mismatch is therefore impossible. The same is true at the first stage. Switching to the other prescribed action raises current payoff by one. After that deviation, at every public second-stage history the receiver can match the other prescribed action and obtain the maximal continuation payoff β, whereas obedience could not have yielded more. Thus continuation incentives cannot offset the current loss. Consequently $D_1 = D_2 = 0$ under every obedient public mechanism. Equation (1) then gives $S_2 = 0$ and $S_3 \sim \mathrm{Bernoulli}(1/2)$, independently of θ, proving the public closure claim.

For the private witness, independently at the two stages use

$$p^{(h)} = \frac{(1, r_h, r_h, 1)}{2(1+r_h)}, \qquad r_1 = \frac{1-\gamma}{1+\gamma}, \qquad r_2 = \frac{\beta-\gamma}{\beta+\gamma}, \tag{4}$$

and make the second-stage law independent of S_2 and the preceding history. The first-stage conditional deviation gap is γ. At the second stage the unweighted coordination gap is γ/β, so its weighted gap is γ. A first-stage deviation can change S_2, but not the continuation experiment; terminal obedience rules out any gain from a later causal deviation. Thus (4) is dynamically γ-obedient.

Its mismatch probabilities are $d_1 = (1-\gamma)/2$ and $d_2 = (\beta-\gamma)/(2\beta)$. Hence the observed gate $G = \{D_1 = 1, S_2 = 1, D_2 = 1\}$ has probability qd_1d_2. Conditional on G, S_3 is Bernoulli with parameter $1/2 + \theta$; off G, its law is independent of θ. The two models are therefore separated in one private closure step. Because G is observed and has the same probability under both models, Hellinger affinity decomposes over G and its complement, which gives the stated distance. At any excluded boundary $q = 0$, $\gamma = \beta$, or $\Delta = 0$, this distance is zero and the strict private refinement disappears. □

Proof of the regret claim in Proposition 37. The two modelwise oracles use a matched first-stage action and recommend 11 at stage two under $+\Delta$, or 00 under $-\Delta$. Both have state-independent continuations and deviation gaps one and β, so they are dynamically γ-obedient. Both architectures therefore have oracle value Δ. Under public obedience, the structural claim above makes the entire observed history model-independent. If $X_t = x(A_{2,t})$, the two expected regrets sum to

$$\Delta \mathbb{E}_{+\Delta} \sum_{t=1}^{T} (1 - X_t) + \Delta \mathbb{E}_{-\Delta} \sum_{t=1}^{T} (1 + X_t) = 2\Delta T.$$

Thus the worst case is at least ΔT, and equal randomization between the two matched stage-two actions attains equality.

Privately, repeat the diagnostic in (4) for N episodes and then use the likelihood-ratio test followed by the corresponding oracle. Put $H = g_{q,\beta,\gamma}(1 - \sqrt{1 - 4\Delta^2})$. The product-law Hellinger affinity is at most $(1 - H)^N \le e^{-NH}$, which upper-bounds the sum of the two testing errors. Each diagnostic has expected value zero and regret Δ. Taking $N = \lceil H^{-1} \log(2T) \rceil \wedge T$ therefore gives worst-case regret at most $\Delta N + \Delta = O(\log T)$.

For the converse, let R be the larger expected regret of any private policy under the two models, and let E be the event that 11 is played at stage two in at least $T/2$ episodes. Pathwise regret accounting gives

$$\Pr_{+\Delta}(E^c) + \Pr_{-\Delta}(E) \le \frac{3R}{\Delta T}.$$

Only $S_2 = D_2 = 1$ distinguishes the models. With $I_\Delta = 2\Delta \log((1/2 + \Delta)/(1/2 - \Delta))$, the adaptive KL chain rule and the fact that every stage-two mismatch costs Δ under $+\Delta$ yield

$$D_{\mathrm{KL}}(\mathbb{P}^{H_T}_{+\Delta} \| \mathbb{P}^{H_T}_{-\Delta}) \le I_\Delta \mathbb{E}_{+\Delta} \sum_{t=1}^{T} D_{2,t} \le \frac{I_\Delta}{\Delta} R.$$

Bretagnolle–Huber applied to E and these two inequalities gives

$$\frac{3R}{\Delta T} \ge \frac{1}{2} \exp\left(-\frac{I_\Delta}{\Delta} R \right).$$

For fixed $\Delta > 0$, this implies $R = \Omega(\log T)$: if $R \ge \sqrt{T}$ the claim is immediate, while otherwise taking logarithms gives $R \ge (\Delta / I_\Delta)[\frac{1}{2} \log T + \log(\Delta/6)]$. □

In the two-coordinate family, each mismatch measures one unknown coordinate. The weakest-coordinate information is therefore limited by how much probability obedience permits on both mismatches. This yields an explicit escape modulus and matching-order identification bounds.

Corollary 38 (Exact Bernoulli escape modulus)**.** *In Proposition 21, let*

$$a_\gamma = \frac{1 - \gamma}{4}, \qquad h_\Delta = 1 - \sqrt{1 - 4\Delta^2},$$
$$I_\Delta = 2\Delta \log \frac{1/2 + \Delta}{1/2 - \Delta}.$$

Then

$$\underline{\chi}_{\mathrm{pub}}(\Theta_\Delta) = 0, \qquad \underline{\chi}_{\mathrm{priv}}(\Theta_\Delta) = a_\gamma h_\Delta.$$

One private block of

$$N \ge \left\lceil \frac{3}{2 a_\gamma h_\Delta} \log \frac{8}{\alpha} \right\rceil$$

episodes identifies the truth with probability at least $1 - \alpha$. Conversely, any uniformly α-correct estimator based on N adaptively selected obedient episodes, where $0 < \alpha < 1/2$, satisfies

$$N \ge \frac{\mathrm{kl}(1 - \alpha, \alpha)}{a_\gamma I_\Delta}.$$

Thus, as $\alpha \downarrow 0$ with Δ, γ fixed, the fixed-confidence complexity is $\Theta_\Delta(\log(1/\alpha)/(a_\gamma h_\Delta))$. Uniformly over $\alpha \le \alpha_0 < 1/2$, its small-$\Delta$ order is $\Theta(\log(1/\alpha)/((1 - \gamma)\Delta^2))$.

Proof. The Hellinger affinity between $\mathrm{Bernoulli}(1/2 - \Delta)$ and $\mathrm{Bernoulli}(1/2 + \Delta)$ is $\sqrt{1 - 4\Delta^2}$. Because the action profile is observed, a single-coordinate alternative has squared Hellinger distance $p_{01} h_\Delta$ or $p_{10} h_\Delta$. If $m_p = \min\{p_{01}, p_{10}\}$, private obedience with $r = (1 - \gamma)/(1 + \gamma)$ gives

$$p_{00}, p_{11} \ge \frac{m_p}{r}, \qquad 1 \ge 2m_p + \frac{2m_p}{r},$$

so $m_p \le r/(2(1+r)) = a_\gamma$. The symmetric diagnostic attains equality, proving the modulus formula. The upper bound is the block-affinity argument in Proposition 25 with $m = 4$, applied directly to the

displayed diagnostic. This does not invoke the finite-mechanism assumption: the private polytope is compact and its maximizer was computed explicitly above.

For the lower bound, fix a truth and the alternatives obtained by flipping only its first or second coordinate. Uniform correctness and data processing give history KL at least $\mathrm{kl}(1-\alpha,\alpha)$ against each alternative. The KL chain rule makes these two divergences

$$I_\Delta \mathbb{E}_{\eta^\star} \sum_{t=1}^{N} p_{01,t}, \qquad I_\Delta \mathbb{E}_{\eta^\star} \sum_{t=1}^{N} p_{10,t}.$$

Every obedient action law satisfies $p_{01,t}+p_{10,t} \le (1-\gamma)/2 = 2a_\gamma$: indeed, writing $s = p_{01,t}+p_{10,t}$, obedience gives $p_{00,t}+p_{11,t} \ge s/r$, hence $1 \ge s(1+1/r)$. Type-blind messages contribute zero conditional KL. Adding the two change-of-measure inequalities yields the lower bound. Finally, $h_\Delta = 2\Delta^2 + O(\Delta^4)$ and $I_\Delta = 8\Delta^2 + O(\Delta^4)$. □

F Computational Verification

The regret simulation uses informative-event counts, so we separately check that complete trajectories produce the predicted gate and conditional observation laws. An independent run of 10^6 trajectories gives gate frequency 0.145093, versus exact 0.1449175 (standard error 0.000352). In the payoff-relevant system, conditional terminal means are 0.799715 and 0.600325, versus 0.8 and 0.6.

Computational checks cover incentive feasibility, observable support, and learning rates. Private LP vertex enumeration agrees with the analytic oracle on parameter grids. Two- and three-model closure enumeration checks architecture refinement and the terminal criterion. The vertex certificate gives public/private ranks $0/2$ in the Bernoulli family and rank 0 at $\gamma = 1$, and recovers the Hellinger modulus (Corollaries 20 and 38).

F.1 Online Learning with Endogenous Experiment Access

This experiment tests whether evidence acquired online releases an initially unavailable probe. We use Proposition 16, whose restriction holds over the full convex class of direct kernels. Initially the mechanism always recommends S; after ζ is excluded, a probe recommends P exactly when $W = 1$. The learner observes (W, R, A, Y), but neither receiver nor sender payoffs. The adaptive policy recomputes access after each observation, probes while α, β remain unresolved, and commits after identification. A frozen-access ablation always uses the unique initially robust mechanism. Evaluating both against the same full-class model-wise oracle isolates the effect of updating access.

Deletion uses sampled feedback rather than the exact closure operator. Let $\ell_t(\vartheta)$ be the cumulative log likelihood of the logged history under candidate ϑ, including the observed actions. Starting from $C_0 = \Theta$, retain

$$C_t = \left\{ \vartheta \in C_{t-1} : \max_{v \in C_{t-1}} \ell_t(v) - \ell_t(\vartheta) \le \log(2/\varepsilon_{\mathrm{fail}}) \right\}.$$

The next mechanism uses C_t, so the observation that deletes ζ precedes the first probe. For each false candidate, its likelihood ratio against the truth is a nonnegative supermartingale under the adaptive experiment sequence. Ville's inequality and a union bound over the two alternatives bound the probability of ever deleting the truth by $\varepsilon_{\mathrm{fail}}$. On its complement, every selected mechanism is γ-obedient. To specify behavior after a mistaken deletion, a receiver in ζ always chooses S, including after a probe recommendation; a history incompatible with every retained model triggers the always-S fallback. We also run ζ as a safety control, completing its sender payoff as $G_\zeta = \mathbf{1}\{A = S\}$.

Each model, policy, and noise level uses 10^4 independent trials, with $\sigma \in \{0.5, 1, 2\}$, $\gamma = 1/4$, and $\varepsilon_{\mathrm{fail}} = 10^{-8}$. Checkpoints $T \in \{100, 300, 1000, 3000, 10000\}$ share each trial's online trajectory. Cumulative pseudo-regret sums conditional expected payoff gaps; these values are used for evaluation and never for action selection. Once a run reaches an absorbing policy and candidate set, subsequent expected gaps are accumulated exactly. The likelihood threshold is fixed across checkpoints, so these curves describe a fixed-confidence experiment, not a fitted asymptotic regret exponent.

Table 2 separates acquiring access from resolving the final decision. Doubling the noise standard deviation approximately quadruples both times, consistent with the Gaussian information scale. Under

Table 2: Online unlocking at $T = 10^4$. Times are median episode indices; all adaptive trials under α, β completed both reductions by this horizon. Brackets give pointwise normal-approximation 95% Monte Carlo intervals for mean pseudo-regret. Frozen access incurs exactly zero regret under α and 2500 under β, at every noise level.

σ	Truth	Probe unlocked	Decision certified	Adaptive regret [95% interval]
0.5	α	10	50	10.41 [10.32, 10.49]
0.5	β	10	50	2.62 [2.60, 2.63]
1.0	α	38	191	39.54 [39.27, 39.80]
1.0	β	38	191	9.91 [9.85, 9.98]
2.0	α	148	751	155.99 [154.99, 156.99]
2.0	β	148	748	38.95 [38.70, 39.20]

β, the unlocked probe is itself optimal; under α, its use incurs a diagnostic cost before returning to S. Frozen access never separates these two models, even after excluding ζ. No trial deleted the truth or selected a nonobedient mechanism in any configuration. For each configuration, zero events in 10^4 trials gives a binomial Chernoff 95% upper bound of 3.69×10^{-4}; it does not empirically establish the much smaller theoretical failure budget.

F.2 Uncertainty for Diagnostic-Then-Commit Regret

A fixed diagnostic budget incurs a deterministic exploration cost; an incorrect likelihood decision adds the remaining exploitation loss. Thus each independent trial has pseudo-regret $L = aN + 2a(T - N)E$, where $a = \Delta b_\star$ and $E \sim \mathrm{Bernoulli}(p)$ records that error. Estimating mean regret reduces to estimating this Bernoulli error probability, including when no error is observed. If k of n trials err, write $\widehat{p} = k/n$ and define

$$\mathcal{C}_\alpha(k, n) = \{p \in [0, 1] : n \ \mathrm{kl}(\widehat{p}, p) \le \log(2/\alpha)\},$$

using Bernoulli relative entropy and its boundary conventions. Binomial Chernoff bounds give each excluded tail probability at most $\alpha/2$, so this interval covers the true error probability with probability at least $1 - \alpha$. Transforming its endpoints by $p \mapsto aN + 2a(T - N)p$ gives the corresponding confidence interval for expected regret. These are pointwise intervals across horizons, not simultaneous confidence bands.

At $k = 0$, the error-probability upper endpoint is $1-(\alpha/2)^{1/n} > 0$. At $N = T$, regret is deterministic even though the error probability is uncertain. The binomial intervals are conservative finite-sample bounds; the reported sample standard error is descriptive and is not used to construct them.

For example, the payoff-relevant positive model at $T = 10^5$ has $N = 3461$ and zero errors in 10^5 trials. Its mean regret is 230.7333, and the transformed 95% interval is $[230.7333, 231.2082]$. Error counts and diagnostic budgets suffice to reproduce these intervals without resampling.

AI Use Statement

The main research ideas and conceptual framework originated with the authors, who also carried out most of the proofs and mathematical derivations. Generative artificial intelligence (AI) tools, including OpenAI Codex, assisted with selected derivations, proof checks, and counterexample searches.

The experimental ideas and study designs were developed by the authors. Following the authors' specifications, AI tools assisted with writing and editing experimental code, debugging implementations, and running simulations. All results reported in the experimental section were obtained by executing the experimental code.

AI tools also assisted with literature searches, bibliographic checks, and polishing the manuscript for clarity and consistency. The authors take full responsibility for the manuscript, including the proofs, code, reported results, and their interpretation.