When Does Swap Regret Immunize Privately-Informed Agents? Type-Conditional No-Swap-Regret in Generalized Principal–Agent Problems

Liz Lemma Future Detective

January 16, 2026

Abstract

Classic principal–agent models benchmark outcomes by a Stackelberg/commitment value $U^\star$ under a best-responding agent. Lin and Chen (2025) show that when the agent has no private information, contextual no-swap-regret learning essentially restores this benchmark: the principal cannot exploit a learning agent beyond $U^\star + o(1)$. In modern 2026 settings, however, agents are often user-side AI assistants with private context (type) θ, raising the question of whether stronger learning guarantees can again prevent manipulation by adaptive platforms.

We introduce a repeated Bayesian generalized principal–agent model with privately informed agents and define a type-conditional contextual no-swap-regret property against deviations $d(s, a, \theta)$. Our main result shows that under i.i.d. types and standard regularity (linearity in decisions, bounded/Lipschitz payoffs, and uniform inducibility gap), type-conditional no-swap-regret is sufficient to cap the principal's long-run payoff at $U^\star + O(\mathrm{CSReg}(T)/T)$, even when the principal moves after observing the agent's policy and can adapt over time. The proof generalizes Lin–Chen's joint-signal reduction while carefully using the fact that the agent's within-round randomization is conditionally independent of the state given (s, θ), so the joint signal (s, a, θ) does not expand the agent's information beyond (s, θ).

We also provide sharp separations: if the regret guarantee is not type-conditional (deviations cannot condition on θ), or if types are persistent so the principal can learn and price-discriminate across rounds, the cap fails and the principal can exceed $U^\star$ by a constant despite vanishing swap regret. The results provide a clean criterion for "manipulation-resistant" AI assistants and clarify when learning-theoretic guarantees are sufficient (or insufficient) to recover classical economic benchmarks in the presence of private context.

Table of Contents

1. 1. Introduction and motivation: delegated AI assistants, private context, and manipulation by adaptive platforms; recap Lin–Chen's no-private-info cap and the gap for private types; contributions and separations.

2. 2. One-shot benchmark with private types: define Bayesian generalized principal–agent problem, commitment/Stackelberg value $U^\star$, and its LP/convex-program formulation (when $X = \Delta(\Omega)$ for persuasion or $X = \Delta(B)$ for Stackelberg).

3. 3. Repeated interaction without commitment: timing (agent policy first, principal next), information, feedback; define type-conditional contextual no-regret and no-swap-regret; discuss implementability and what the principal observes.

4. 4. Reduction: from type-conditional no-swap-regret to approximate obedience on an induced joint distribution; joint-signal (s, a, θ) construction; key conditional-independence lemma $(\omega \perp a \mid s, \theta)$.

5. 5. Main cap theorem: principal utility $\leq U^\star + O(\mathrm{CSReg}(T)/T)$ (plus explicit constants under Lipschitz/inducibility assumptions); robustness to principal adaptivity and knowledge of the learning algorithm; optional matching lower bound with type-conditional no-regret using fixed principal strategies.

6. 6. Failure modes and tight examples: (i) regret notions that omit θ; (ii) persistent types and principal learning; (iii) within-round observability variants (principal observes action before choosing π_t); characterize when exploitation becomes possible.

7. 7. Applications and calibrations: (a) personalized disclosure and compliance warnings; (b) platform pricing with ephemeral vs persistent preferences; (c) delegation to AI assistants and auditing type-conditional internal-regret properties.

8. 8. Discussion and future work: multi-agent populations, partial monitoring, computational constraints for principals, and designing learning guarantees as a policy instrument.

1 Introduction and motivation

Delegated decision-making is increasingly mediated by user-side AI assistants: systems that summarize options, negotiate on a user's behalf, or select among actions (purchases, content filtering, scheduling) after receiving a platform-provided interface or message. A central feature of this delegation environment is that the assistant is privately informed. It conditions its behavior on *private context*—the user's preferences, constraints, and goals—that the platform does not directly observe and may not be able to infer reliably from a single interaction. At the same time, the platform is often adaptive: it can run experiments, personalize messages, and change decision rules over time in response to observed outcomes. This combination raises a natural concern for both mechanism design and AI governance: can an adaptive platform *manipulate* a learning assistant into taking actions that increase platform payoff, beyond what would be achievable if the platform were forced to commit up front to a policy?

A useful benchmark comes from work that studies repeated principal–agent interaction when the agent is a no-regret learner. In particular, in models without private information, one can show a striking "cap": if the agent's learning dynamics satisfy an appropriate no-swap-regret guarantee, then even a fully adaptive principal cannot extract more than the one-shot commitment value, up to an error that vanishes with average regret. Intuitively, swap regret is the right behavioral notion because it enforces approximate *obedience*: the realized play looks like the agent is (approximately) best-responding to the principal's induced incentives, so the principal cannot profit from repeatedly steering the agent through systematically suboptimal responses.

However, delegated AI assistants are *precisely* settings in which private information is first-order. The assistant observes the user's type (or context) each round, and this type may vary across interactions. The platform, by contrast, chooses its policy without observing the type, and in many applications it cannot even condition within a round on the action the assistant ultimately takes (e.g., the interface is chosen before downstream choices are observed). These information constraints change what "obedience" should mean and, consequently, what sort of regret guarantee is sufficient to immunize the agent from manipulation. A regret bound that is appropriate in a public-information model may be too weak once types enter, because it may fail to control deviations that target particular types.

We therefore study a repeated Bayesian generalized principal–agent problem in which, in each round, Nature draws a state and a type, the principal chooses a policy that generates a signal and a decision, and the agent chooses an action after observing the signal and its type. The principal's objective is to maximize its expected average payoff across rounds, and it may use any adaptive strategy, including one that is aware of the agent's learning algo-

rithm. The agent is modeled as a learning algorithm that guarantees a form of *type-conditional contextual no-swap-regret*: for each realized type, and for each deviation mapping that may depend on the signal, the agent's realized action, and the type, the cumulative gain from switching to that deviation is small. This learning guarantee is natural for assistants that explicitly condition on user context and can internally post-process their own recommended actions (for instance, by re-mapping outputs of an underlying model into a smaller action set).

Our first contribution is to show that, under i.i.d. types and standard boundedness/regularity conditions (including an inducibility gap that rules out uniformly dominated actions), type-conditional contextual no-swap-regret restores a cap analogous to the one in the no-private-information setting. In words: even though the principal can adapt its policy over time, it cannot increase its expected average payoff beyond the one-shot Bayesian commitment value $U^\star$, except by a term that scales linearly with the agent's average swap regret $\delta = \mathrm{CSReg}(T)/T$. This provides a simple economic interpretation of strong learning guarantees for user-side assistants: when the assistant is "sufficiently obedient" conditional on the user's private context, dynamic platform manipulation collapses to the classical one-shot design problem.

The technical obstacle is that swap regret is naturally phrased in terms of the agent's *own* history and internal recommendations, whereas Bayesian obedience constraints are phrased conditional on the information available at the moment of choice, here (s, θ). Our reduction bridges this gap by treating (s, a, θ) as a joint signal and translating contextual swap regret into an approximate obedience condition for the empirical distribution of play. A key ingredient is a conditional independence property: because the agent's learning algorithm does not observe the state beyond the principal's signal, the realized action does not reveal additional information about the state once we condition on (s, θ). This allows us to "collapse" approximate obedience from the augmented signal (s, a, θ) back down to the economically relevant signal (s, θ), and then to compare the principal's adaptive value to the one-shot benchmark using perturbation arguments in the spirit of Lin–Chen.

Our second contribution clarifies when such caps fail, and why the dependence on θ in the deviation class is not a technicality. We provide a separation showing that if the learning guarantee only controls deviations that *cannot* condition on type—for example, deviations $d(s, a)$ that ignore θ—then there exist two-type instances where the principal can obtain $U^\star + \Omega(1)$ while the reported swap regret still vanishes. Economically, the platform can concentrate distortions on a minority type: it can induce systematic mistakes that are profitable in expectation yet invisible to type-agnostic deviations. This is a cautionary message for evaluation and auditing practices that report aggregate regret-like metrics without stratifying by user context: such metrics can certify "good learning" while permitting significant exploitation of

particular user segments.

Our third contribution highlights a distinct failure mode that arises when types are persistent. If θ is fixed for an individual across rounds and the principal can observe actions or outcomes, then repeated interaction becomes a screening instrument. Even if the agent satisfies type-conditional no-swap-regret *ex post* over the realized sequence, the principal may be able to run an exploration phase that elicits informative behavior and then switch to type-tailored policies, achieving $U^\star + \Omega(1)$. This separation underscores that our cap is not a generic impossibility result about manipulation; it is a statement about a particular informational regime (fresh i.i.d. private types) in which the principal cannot effectively learn the type before choosing the within-round policy.

Beyond its theoretical interest, the i.i.d. regime is a reasonable approximation in many practical deployments. Users' immediate contexts (time constraints, current objective, risk tolerance) vary across sessions; platforms often must commit to an interface or set of options before observing downstream user-side actions; and assistants can be designed to guarantee strong regret properties conditional on observed context. In such settings, our results suggest a concrete design principle: to limit platform manipulation, it is not enough that the assistant be no-regret "on average." The guarantee must be robust to deviations that target the user's private context, i.e., it must be type-conditional.

At the same time, our analysis has limitations that inform both modeling and practice. The cap relies on regularity assumptions (boundedness, Lipschitz dependence on the principal's decision, and a uniform inducibility gap) that rule out knife-edge instances where tiny incentive perturbations cause discontinuous action changes. It also relies on the principal's inability to condition within a round on realized actions. In applications where the platform can observe intermediate user-side behavior before finalizing the decision, or where the assistant's type is stable and actions are repeatedly observed, platforms may regain leverage, and additional defenses (cryptographic commitment, randomized response, or structural restrictions on platform policies) may be required.

The remainder of the paper develops these points systematically. In Section 2 we formalize the one-shot Bayesian benchmark with private types and define the commitment (Stackelberg) value $U^\star$, including convenient convex/linear program formulations in common special cases (Bayesian persuasion and Stackelberg decision problems). We then analyze the repeated game, establish the swap-regret-to-obedience reduction, and prove the cap theorem. Finally, we present the two separations that delineate the boundary of the cap: (i) omitting type from the deviation class, and (ii) allowing persistent types and learning-by-screening over time.

2 One-shot benchmark with private types

We begin by fixing the one-shot Bayesian generalized principal–agent problem that serves as our commitment benchmark. Nature draws a pair $(\omega, \theta) \in \Omega \times \Theta$ from the common prior μ_0. The principal observes the state ω (but not the type θ) and, having committed ex ante to a policy, produces an observable signal $s \in S$ and a decision $x \in X$. The agent observes (s, θ) and then chooses an action $a \in A$. Payoffs are given by $u(x, a, \omega, \theta)$ for the principal and $v(x, a, \omega, \theta)$ for the agent, with the maintained linearity-in-x structure

$$u(x, a, \omega, \theta) = \langle x, U_{a,\omega,\theta} \rangle, \qquad v(x, a, \omega, \theta) = \langle x, V_{a,\omega,\theta} \rangle.$$

A *principal commitment policy* can be represented as a pair $(\pi, \{x_s\}_{s \in S})$, where $\pi(\cdot \mid \omega) \in \Delta(S)$ is a signaling scheme and $x_s \in X$ is the decision implemented upon sending signal s.[1] Given $(\pi, \{x_s\})$, Bayes' rule yields the agent's posterior over states:

$$\mu(\omega \mid s, \theta) = \frac{\mu_0(\omega, \theta)\, \pi(s \mid \omega)}{\sum_{\omega' \in \Omega} \mu_0(\omega', \theta)\, \pi(s \mid \omega')},$$

whenever the denominator is positive. Conditional on (s, θ), the agent chooses a best response

$$a^\star(s, \theta) \in \arg\max_{a \in A} \mathbb{E}\big[v(x_s, a, \omega, \theta) \mid s, \theta\big] = \arg\max_{a \in A} \sum_{\omega \in \Omega} \mu(\omega \mid s, \theta)\, v(x_s, a, \omega, \theta).$$

The principal anticipates this behavior and chooses $(\pi, \{x_s\})$ to maximize its expected payoff. We define the *commitment (Stackelberg) value* as

$$U^\star := \sup_{\pi, \{x_s\}} \; \mathbb{E}_{(\omega,\theta) \sim \mu_0,\; s \sim \pi(\cdot \mid \omega)} \Big[u\big(x_s, a^\star(s, \theta), \omega, \theta\big) \Big]. \tag{1}$$

Because Ω, Θ, A are finite and X is compact, the benchmark is well defined under mild regularity; when best responses are not unique, we interpret (1) using any measurable selection $a^\star(s, \theta)$ (our upper bounds later are robust to tie-breaking, and our separations can be constructed under strict incentives).

Two aspects of (1) are worth emphasizing because they reappear in the repeated-game analysis. First, the agent best responds *conditional on private type*: the relevant obedience constraints are indexed by (s, θ), not merely s. Second, the principal must choose π without observing θ, so the only way in which the principal can tailor incentives across types is indirectly, through the correlation structure in μ_0 and through the fact that different types interpret the same public signal s differently via Bayes' rule.

[1] Allowing the principal to randomize over decisions conditional on (ω, s) does not change the benchmark in our linear setting: by linearity, only conditional expectations matter, and any lottery can be folded into an enlarged signal alphabet. In the simplex special cases below, this reduction yields an explicit linear program.

A distributional (obedience) formulation. It is often convenient to phrase the benchmark in terms of distributions over outcomes that satisfy Bayes plausibility and obedience constraints. Fix a finite signal alphabet S. A policy $(\pi, \{x_s\})$ induces a joint distribution over (ω, θ, s) via

$$\Pr(\omega, \theta, s) = \mu_0(\omega, \theta)\pi(s \mid \omega),$$

and then a type-contingent action rule $(s, \theta) \mapsto a^\star(s, \theta)$. The induced outcome is *obedient* if, for every (s, θ) that occurs with positive probability and every deviation action $a' \in A$,

$$\sum_{\omega \in \Omega} \Pr(\omega, \theta, s)\Big(v(x_s, a^\star(s, \theta), \omega, \theta) - v(x_s, a', \omega, \theta)\Big) \;\geq\; 0. \tag{2}$$

Multiplying by $\Pr(s, \theta)$ removes the posterior normalization and makes (2) the natural one-shot analogue of the approximate obedience conditions we will obtain from swap regret in the repeated model.

Linear/convex programs in common special cases. While (1) is conceptually simple, its computation can be clarified by two canonical instantiations of the decision space X. In both cases, the key simplification is that we can absorb the principal's randomization into a joint distribution over (ω,"message"), after which both the objective and the obedience constraints become linear inequalities.

(i) Bayesian persuasion: $X = \Delta(\Omega)$. In Bayesian persuasion, the principal's "decision" is an information structure about the state. A convenient reduced-form is to let signals directly encode posteriors: each $s \in S$ corresponds to a posterior $x_s \in \Delta(\Omega)$, and Bayes plausibility requires that the average posterior equals the prior marginal on Ω,

$$\sum_{s \in S} \alpha_s x_s = \mu_0^\Omega, \qquad \text{where } \alpha_s := \Pr(s), \quad \mu_0^\Omega(\omega) := \sum_\theta \mu_0(\omega, \theta). \tag{3}$$

With private types, the agent conditions on θ as well, so the posterior relevant for a type θ upon seeing s is the "tilted" belief

$$\mu(\omega \mid s, \theta) \;\propto\; x_s(\omega)\, \mu_0(\theta \mid \omega).$$

If we specialize payoffs to depend on ω and a only through expectations under the induced posterior (as is standard in persuasion), the linear-in-x representation is immediate: for baseline payoffs $\bar{u}(a, \omega, \theta)$, $\bar{v}(a, \omega, \theta)$, set

$$u(x, a, \theta) = \sum_{\omega \in \Omega} x(\omega)\, \bar{u}(a, \omega, \theta), \qquad v(x, a, \theta) = \sum_{\omega \in \Omega} x(\omega)\, \bar{v}(a, \omega, \theta).$$

A fully linear formulation is obtained by working with joint variables $q_{\omega,s} := \Pr(\omega, s)$. The feasible set is

$$q_{\omega,s} \geq 0, \qquad \sum_{s \in S} q_{\omega,s} = \mu_0^{\Omega}(\omega) \quad \forall \omega,$$

and Bayes' rule is implicit in the normalization $q_{\omega,s} / \sum_{\omega'} q_{\omega',s}$. If we enrich the signal to specify a *type-contingent recommendation profile* $r \in A^{\Theta}$ (so that, upon seeing r, type θ is "recommended" to play $r(\theta)$), then the one-shot persuasion benchmark can be written as the linear program

$$\max_{\{q_{\omega,r}\}} \quad \sum_{\omega \in \Omega} \sum_{\theta \in \Theta} \sum_{r \in A^{\Theta}} q_{\omega,r}\, \mu_0(\theta \mid \omega)\, \bar{u}\big(r(\theta), \omega, \theta\big) \tag{4}$$

$$\text{s.t.} \quad \sum_{r \in A^{\Theta}} q_{\omega,r} = \mu_0^{\Omega}(\omega) \qquad \forall \omega, \tag{5}$$

$$\sum_{\omega \in \Omega} q_{\omega,r}\, \mu_0(\theta \mid \omega) \Big(\bar{v}\big(r(\theta), \omega, \theta\big) - \bar{v}(a', \omega, \theta) \Big) \;\geq\; 0 \quad \forall r,\ \forall \theta,\ \forall a' \in A. \tag{6}$$

Constraints (6) are precisely obedience constraints of the form (2), written without posterior normalization. Importantly, allowing recommendation profiles $r \in A^{\Theta}$ is without loss: the principal need not observe θ to send a signal that contains a full contingency plan; the privately informed agent simply conditions on its realized θ when interpreting the recommendation.

(ii) Stackelberg decision problems: $X = \Delta(B)$. In a broad class of Stackelberg models, the principal chooses a lottery over a finite set B of concrete moves (prices, allocations, contracts, or actions), and then the agent chooses $a \in A$. Taking $X = \Delta(B)$ and using linearity, we can equivalently let the principal randomize directly over pure moves $b \in B$ as part of the signal. Concretely, we let a "message" be $m = (b, r) \in B \times A^{\Theta}$, where b is the realized principal move and r is a recommendation profile. Let $q_{\omega,m} := \Pr(\omega, m)$ denote the joint distribution. The commitment benchmark becomes the linear program

$$\max_{\{q_{\omega,m}\}} \quad \sum_{\omega \in \Omega} \sum_{\theta \in \Theta} \sum_{m = (b,r)} q_{\omega,m}\, \mu_0(\theta \mid \omega)\, u\big(b, r(\theta), \omega, \theta\big) \tag{7}$$

$$\text{s.t.} \quad \sum_{m} q_{\omega,m} = \mu_0^{\Omega}(\omega) \qquad \forall \omega, \tag{8}$$

$$\sum_{\omega \in \Omega} q_{\omega,m}\, \mu_0(\theta \mid \omega) \Big(v\big(b, r(\theta), \omega, \theta\big) - v(b, a', \omega, \theta) \Big) \;\geq\; 0 \quad \forall m = (b, r),\ \forall \theta,\ \forall a' \in A, \tag{9}$$

where we have slightly abused notation by writing $u(b, a, \omega, \theta)$ for the payoff associated with the degenerate lottery on b. As in persuasion, the incentive

constraints (9) are linear because we work with joint variables $q_{\omega,m}$ and because deviation comparisons are evaluated in expectation without requiring explicit posterior computations.

These LP formulations are not merely computational conveniences: they make explicit that the benchmark $U^\star$ is the optimal value over *obedient* outcome distributions induced by a policy that can condition on ω but not on θ. In the repeated game, our cap theorem will compare the principal's adaptive value to $U^\star$ by showing that no-swap-regret learning forces the empirical distribution of play to be approximately obedient in essentially the same sense as (2).

3 Repeated interaction without commitment

We now turn to the repeated interaction in which the principal does *not* commit ex ante to a single policy. Instead, the principal may adapt its behavior across rounds as a function of past play, and may even be fully "algorithm-aware" in the sense of knowing the agent's learning rule. Our goal in this section is to make precise (i) the within-round timing restriction that prevents the principal from conditioning on the agent's *realized* action when choosing a policy, (ii) the information available to each side, and (iii) the learning guarantees we impose on the agent.

Timing and strategies. Fix a horizon $T \in \mathbb{N}$. In each round $t \in \{1, \dots, T\}$, Nature draws $(\omega_t, \theta_t) \sim \mu_0$ i.i.d. across rounds. The principal observes ω_t but not θ_t, and the agent observes θ_t but not ω_t (beyond what is conveyed by the principal's signal). We model the play within round t as follows:

1. The agent, after observing its current type θ_t and the public history, chooses a (possibly randomized) *response map*

$$\rho_t : \ S \times \Theta \to \Delta(A),$$

   which specifies, for each possible signal s and type θ, a mixed action to be used if that signal and type occur.[2]

2. The principal then chooses a (possibly randomized) *policy* π_t as a function of the public history and the observed state ω_t. Formally, $\pi_t(\cdot \mid \omega_t) \in \Delta(S \times X)$ induces a joint distribution over the signal $s_t \in S$ and the decision $x_t \in X$ conditional on ω_t.

3. The principal samples $(s_t, x_t) \sim \pi_t(\cdot \mid \omega_t)$ and publicly sends s_t (and implements x_t).

[2]This "choose a response map" representation is purely notational: it is equivalent to allowing the agent to choose an action after observing s_t; we write it this way to emphasize that the agent is committing, within the round, to a contingent plan and that the learning guarantees apply to the realized mapping from contexts to actions.

4. The agent observes (s_t, θ_t) and draws an action $a_t \sim \rho_t(s_t, \theta_t)$.

5. Payoffs $u(x_t, a_t, \omega_t, \theta_t)$ and $v(x_t, a_t, \omega_t, \theta_t)$ are realized, and the agent receives feedback sufficient to support the regret guarantee stated below.

The critical restriction is that the principal chooses π_t *before* a_t is realized. Thus, even though the principal may be adaptive across rounds, it cannot implement within-round screening rules that condition on the agent's realized action.

We allow the principal's strategy to be fully history-dependent. Concretely, letting h_{t-1} denote the public history up to (and including) round $t-1$, a principal strategy is a sequence of mappings $h_{t-1} \mapsto \pi_t$, where each π_t may depend on the principal's entire past observation stream (including past states $\omega_{1:t-1}$ and past realized actions $a_{1:t-1}$ if these are observable ex post). The agent's strategy is likewise history-dependent, but is constrained by the learning property we impose.

Information and observables. We keep the informational asymmetry from the one-shot benchmark: the agent privately observes θ_t, while the principal does not. The principal may observe ω_t each round, reflecting platform-side observability of the relevant "state" (e.g., content quality, market conditions, or the sender's own private information). The agent's only within-round information about ω_t is through the realized signal s_t (and through the realized payoff, if payoffs are observed).[3]

It is important to distinguish what is observable to the principal *before* choosing π_t versus *after* the round concludes. We impose no restriction on what the principal may learn *after* round t (e.g., the principal might observe a_t, outcomes, or even the agent's realized payoff). Our cap theorem will be robust to such observations because they cannot be used for within-round conditioning on a_t, and because types are drawn i.i.d. (so past actions do not directly reveal the current type). Later, we will show that if types persist, ex post observation of actions can restore the principal's leverage.

Utilities and the repeated objective. The within-round payoffs are as in Section 2, with the maintained linearity-in-x property

$$u(x, a, \omega, \theta) = \langle x, U_{a,\omega,\theta} \rangle, \qquad v(x, a, \omega, \theta) = \langle x, V_{a,\omega,\theta} \rangle.$$

[3] In many applications, the agent naturally observes a realized utility (click-through, reward signal, or task success) without directly observing the underlying state. Standard bandit-style algorithms can be used in such settings; we do not fix a particular feedback model, and instead assume the stated regret bound holds under the realized sequence of payoff functions.

We evaluate performance via expected average utility:

$$U_T := \frac{1}{T}\,\mathbb{E}\Big[\sum_{t=1}^{T} u(x_t, a_t, \omega_t, \theta_t)\Big], \qquad V_T := \frac{1}{T}\,\mathbb{E}\Big[\sum_{t=1}^{T} v(x_t, a_t, \omega_t, \theta_t)\Big],$$

where the expectation is over the prior draws, both players' randomization, and any algorithmic randomness.

Type-conditional contextual regret. Because the agent observes (s_t, θ_t) before acting, the natural deviation classes are allowed to condition on *both* the public signal and the private type. This is precisely the sense in which private types enter obedience constraints in the one-shot benchmark (cf. (2)). In the repeated setting, we accordingly define a type-conditional contextual no-regret property as the requirement that, for every deviation rule $d : S \times \Theta \to A$,

$$\mathbb{E}\left[\sum_{t=1}^{T}\Big(v\big(x_t, d(s_t, \theta_t), \omega_t, \theta_t\big) - v(x_t, a_t, \omega_t, \theta_t)\Big)\right] \;\le\; \mathrm{CReg}(T). \tag{10}$$

The key feature of (10) is that the deviation is allowed to "target" different types: it can prescribe distinct actions for the same signal s depending on θ. This is not merely a technical strengthening. If the deviation class cannot condition on θ, then the guarantee may fail to control behavior on small-probability types, and an adaptive principal can exploit precisely those hidden degrees of freedom; we formalize this in Separation 1.

Type-conditional contextual swap regret. Our main cap result relies on the stronger notion of *swap* regret, again with deviations allowed to depend on type. Informally, swap regret asks whether, after seeing the action the algorithm actually played, the agent could have improved by systematically "swapping" that action for a different one as a function of the context. Formally, for every deviation $d : S \times A \times \Theta \to A$, we require

$$\mathbb{E}\left[\sum_{t=1}^{T}\Big(v\big(x_t, d(s_t, a_t, \theta_t), \omega_t, \theta_t\big) - v(x_t, a_t, \omega_t, \theta_t)\Big)\right] \;\le\; \mathrm{CSReg}(T). \tag{11}$$

We will write $\delta := \mathrm{CSReg}(T)/T$ for the corresponding average swap-regret rate. Relative to (10), the deviation class in (11) is strictly richer: it may condition on (s_t, θ_t) *and* on the agent's own realized action a_t. This richer deviation class is exactly what yields approximate *obedience* of the empirical distribution of play, in the same sense as the one-shot constraints (2) but now only approximately (up to δ).

Why type-conditional guarantees are implementable. From a learning perspective, allowing deviations to depend on θ is natural because θ_t is part of the agent's observed context. If Θ is finite, one simple implementation is to run an independent contextual (swap-)regret learner for each type $\theta \in \Theta$, updating only on rounds in which $\theta_t = \theta$. More abstractly, the agent can run a single contextual learner on the augmented context space $S \times \Theta$. In either view, the learning problem faced by the agent is a standard repeated decision problem with finite action set A and context (s_t, θ_t), where the per-round payoff function is induced endogenously by the principal's choice of x_t (and by the realization of ω_t). Our analysis treats the principal's behavior as potentially adaptive and adversarial from the learner's perspective; the assumption is that the algorithm nonetheless guarantees (11) (or, for the lower-bound direction, (10)).

What the principal can and cannot condition on. We emphasize that our model permits the principal to be extremely powerful along two dimensions: it may choose π_t adaptively based on the entire public history, and it may know the agent's learning algorithm (and even observe ρ_t if one interprets the "response map" as a public commitment). What the principal cannot do is condition π_t on the *realized* a_t within the same round. This restriction is precisely what rules out within-round screening and will underwrite the key conditional-independence step in the next section: since the agent does not observe ω_t beyond s_t, its action cannot carry additional information about ω_t once we condition on (s_t, θ_t).

Preview: from learning to obedience. The definitions above are designed to make the repeated game comparable to the one-shot benchmark. In Section 4 we show that (11) implies that the empirical joint distribution over outcomes induced by play is δ-approximately obedient, with obedience indexed by (s, θ) as in (2). This reduction is the main bridge between online learning dynamics and the static commitment value $U^\star$.

4 From learning to approximate obedience

Our cap theorem will be proved by reducing the repeated interaction to a *single* (random) round in which the induced distribution of play is approximately obedient. Intuitively, type-conditional contextual swap regret says that, retrospectively, the agent cannot improve by applying any systematic "action-relabeling" rule that is allowed to depend on the information the agent actually had when it acted (namely (s_t, θ_t), and also the action it drew). That is exactly the content of an obedience constraint—except that, because we only control regret up to $\mathrm{CSReg}(T)$, obedience will hold only up to $\delta = \mathrm{CSReg}(T)/T$.

A random-round representation and the induced joint distribution. Let τ be a uniform random index in $\{1, \dots, T\}$, independent of all play. Consider the random tuple

$$(\omega_\tau, \theta_\tau, s_\tau, x_\tau, a_\tau),$$

generated by the repeated interaction. We write $\widehat{\mathbb{P}}_T$ for its induced distribution (which is itself a mixture over histories and the principal's adaptivity). With this notation, the swap-regret condition (11) is equivalently

$$\forall d : S \times A \times \Theta \to A, \qquad \mathbb{E}_{\widehat{\mathbb{P}}_T}\Big[v\big(x_\tau, d(s_\tau, a_\tau, \theta_\tau), \omega_\tau, \theta_\tau\big) - v(x_\tau, a_\tau, \omega_\tau, \theta_\tau)\Big] \;\le\; \delta. \tag{12}$$

Thus, without yet committing to any structural interpretation, we may view the repeated play as generating a one-shot Bayesian instance in which the "signal" observed by the agent is the triple $(s_\tau, a_\tau, \theta_\tau)$, and inequality (12) asserts that the realized action a_τ is an *approximate* best response relative to this joint signal.

Joint-signal obedience from swap regret. To connect (12) to the familiar obedience constraints from Section 2, it is helpful to make explicit the conditional averaging over the principal's decision x. Let $\mathcal{Z} := S \times A \times \Theta$, and for each $z = (s, a, \theta) \in \mathcal{Z}$ define the conditional average decision

$$y_z \;:=\; \mathbb{E}_{\widehat{\mathbb{P}}_T}\big[x_\tau \mid s_\tau = s,\; a_\tau = a,\; \theta_\tau = \theta\big] \;\in\; X,$$

with an arbitrary value in X on zero-probability conditioning events.[4] By the maintained linearity-in-x of v, we can replace the random x_τ by its conditional mean inside expected utilities. Concretely, for any fixed action $a' \in A$ and any $z = (s, a, \theta)$,

$$\mathbb{E}\big[v(x_\tau, a', \omega_\tau, \theta) \mid z\big] \;=\; \mathbb{E}\big[\langle x_\tau, V_{a', \omega_\tau, \theta}\rangle \mid z\big] \;=\; \mathbb{E}\big[\langle y_z, V_{a', \omega_\tau, \theta}\rangle \mid z\big] \;=\; \mathbb{E}\big[v(y_z, a', \omega_\tau, \theta) \mid z\big].$$

Substituting this identity into (12) yields the following interpretation: under $\widehat{\mathbb{P}}_T$, after observing $z = (s, a, \theta)$, the agent cannot gain more than δ by applying any mapping $d(z)$ to replace the realized action a with a different action. This is precisely a δ-approximate obedience condition, now written for deviations that condition on the *joint* signal (s, a, θ) rather than on (s, θ) alone.

We record this as a proposition because it will be the first input to the cap theorem.

[4] Because X is convex and compact, conditional expectations are well-defined and remain in X.

Proposition 4.1 (Reduction to δ-obedience on the joint signal)**.** *If the agent satisfies type-conditional contextual no-swap-regret* (11) *with* $\delta = \mathrm{CSReg}(T)/T$*, then for every deviation* $d : S \times A \times \Theta \to A$*,*

$$\mathbb{E}_{\widehat{\mathbb{P}}_T}\Big[v\big(x_\tau, d(s_\tau, a_\tau, \theta_\tau), \omega_\tau, \theta_\tau\big) - v(x_\tau, a_\tau, \omega_\tau, \theta_\tau)\Big] \;\le\; \delta.$$

Equivalently, writing $z_\tau = (s_\tau, a_\tau, \theta_\tau)$ *and* $y_{z_\tau} = \mathbb{E}[x_\tau \mid z_\tau]$*, we have*

$$\mathbb{E}_{\widehat{\mathbb{P}}_T}\Big[v\big(y_{z_\tau}, d(z_\tau), \omega_\tau, \theta_\tau\big) - v\big(y_{z_\tau}, a_\tau, \omega_\tau, \theta_\tau\big)\Big] \;\le\; \delta.$$

At this point, we have an obedience statement, but it is indexed by (s, a, θ) rather than the economically natural information set (s, θ). The remaining step is to show that, in our timing model, conditioning on the realized action a provides *no additional information* about the state beyond (s, θ). This is the conditional-independence fact that ultimately prevents the principal from exploiting within-round learning "mistakes."

A conditional-independence lemma. The key observation is that the agent's action is generated by an algorithm that does not observe ω_t beyond the principal's signal s_t, and the principal must choose (s_t, x_t) before a_t is realized. As a result, once we condition on what the agent actually observed, namely (s_t, θ_t) (and, if desired, the public history and the agent's internal randomness), the realized action cannot carry further information about ω_t.

Lemma 4.2 (No additional state information in actions)**.** *For each round* t*, under the timing described in Section 3, we have the conditional independence*

$$\omega_t \;\perp\; a_t \;\mid\; (s_t, \theta_t).$$

Equivalently, for every bounded function $g : \Omega \to \mathbb{R}$*,*

$$\mathbb{E}\big[g(\omega_t) \mid s_t, \theta_t, a_t\big] \;=\; \mathbb{E}\big[g(\omega_t) \mid s_t, \theta_t\big].$$

The proof is a direct application of the within-round timing restriction. Conditional on (s_t, θ_t), the distribution of a_t is determined entirely by the agent's response map $\rho_t(s_t, \theta_t)$ and the agent's own randomization. Since ρ_t is chosen without observing ω_t and the agent receives no within-round information about ω_t other than s_t, the randomness that generates a_t is independent of ω_t given (s_t, θ_t). Importantly, this argument does *not* require any restriction on the principal's adaptivity across rounds: even if π_t is chosen adversarially based on the full public history and knowledge of the learning algorithm, once s_t is fixed, the agent's subsequent randomization cannot "depend back" on the realized ω_t.

Collapsing joint-signal obedience to (s,θ)-obedience. Lemma 4.2 allows us to convert Proposition 4.1 into the form we need for comparison with the one-shot benchmark. In the one-shot persuasion/Stackelberg problem, obedience constraints are indexed by the agent's information (s,θ): after observing (s,θ), the agent should not profitably deviate. Here, because (s,a,θ) does not refine beliefs about ω beyond (s,θ), the richer deviation class in swap regret does not give the agent any extra inferential power; it merely allows us to state the constraint in a way that is directly implied by the online learning guarantee.

Formally, fix any deviation rule $\widetilde{d} : S\times\Theta \to A$ that depends only on (s,θ). Consider the swap deviation $d : S\times A\times\Theta \to A$ defined by $d(s,a,\theta) = \widetilde{d}(s,\theta)$ (i.e., ignore the realized action). Applying Proposition 4.1 to this d yields

$$\mathbb{E}\Big[v\big(x_\tau, \widetilde{d}(s_\tau,\theta_\tau), \omega_\tau, \theta_\tau\big) - v(x_\tau, a_\tau, \omega_\tau, \theta_\tau)\Big] \;\le\; \delta.$$

Using Lemma 4.2, we may interpret this as an approximate obedience constraint relative to the information set (s_τ,θ_τ): conditioning further on a_τ cannot change the posterior over ω_τ, and hence cannot increase the value of a deviation that is allowed to condition only on (s_τ,θ_τ). In particular, writing

$$y_{s,\theta} \;:=\; \mathbb{E}\big[x_\tau \mid s_\tau = s,\ \theta_\tau = \theta\big],$$

linearity again implies that, conditional on (s,θ), the agent evaluates actions against $y_{s,\theta}$.

We summarize the outcome as the promised reduction.

Corollary 4.3 (δ-approximate obedience on (s,θ))**.** *Under the hypotheses of Proposition 4.1, the induced distribution $\widehat{\mathbb{P}}_T$ is δ-approximately obedient with respect to the agent's information (s,θ): for every deviation $\widetilde{d} : S\times\Theta\to A$,*

$$\mathbb{E}_{\widehat{\mathbb{P}}_T}\Big[v\big(x_\tau, \widetilde{d}(s_\tau,\theta_\tau), \omega_\tau, \theta_\tau\big) - v(x_\tau, a_\tau, \omega_\tau, \theta_\tau)\Big] \;\le\; \delta.$$

Equivalently, for each (s,θ), the realized mixed action $\mathcal{L}(a_\tau \mid s_\tau = s, \theta_\tau = \theta)$ is a δ-approximate best response to the induced conditional decision $y_{s,\theta}$ and the posterior over ω given (s,θ).

Discussion and limitations. Corollary 4.3 is the central bridge from online learning to our static benchmark: regardless of how the principal adapts over time, the empirical distribution of play is constrained (up to δ) by the same obedience inequalities that define feasibility in the one-shot commitment problem. The conditional-independence step is also where our timing and i.i.d.-types assumptions enter in an essential way. If the principal could condition π_t on the realized a_t within the round, then a_t would become part of the principal's information set when selecting (s_t, x_t), breaking the argument. Likewise, if types were persistent and actions informative about type,

then past actions could help the principal predict future types and effectively implement type-contingent policies across rounds, invalidating the reduction that treats each round as a fresh Bayesian instance. These are exactly the channels exploited in Separation 2, and they clarify why our cap theorem is tightly tied to the combination of (i) within-round non-observability of actions and (ii) i.i.d. private types.

5 A cap theorem for adaptive principals

We now turn the approximate-obedience reduction of Section 4 into an upper bound on what any principal can extract in the repeated interaction. The economic content is simple: once the agent is (approximately) behaving as if it were best-responding to the information it actually has, the principal is effectively facing the same constraint set as in the one-shot commitment benchmark. Any additional "dynamic" degrees of freedom in choosing π_t—including conditioning on the full public history and even knowing the agent's learning algorithm—matter only insofar as they change the induced distribution over (s, θ, x, a), but they cannot relax the obedience inequalities by more than the agent's swap-regret rate.

Benchmark and the role of approximate obedience. Recall that $U^\star$ is the commitment value of the corresponding one-shot Bayesian problem: the principal commits to a policy π (mapping ω to a distribution over (s, x)), the agent observes (s, θ), and then best-responds. In that one-shot problem, the feasible set of outcome distributions is characterized by (Bayesian) obedience constraints indexed by (s, θ), together with the implementability constraints coming from the signaling scheme and the decision space X. Corollary 4.3 tells us that, in the repeated game under type-conditional contextual no-swap-regret, the *random-round* outcome distribution $\widehat{\mathbb{P}}_T$ satisfies the same obedience constraints up to an additive $\delta = \mathrm{CSReg}(T)/T$.

At a high level, the cap theorem is then a perturbation statement: optimizing the principal's expected payoff over δ-approximately obedient distributions yields at most $U^\star + O(\delta)$. The $O(\delta)$ term becomes explicit once we impose the same regularity conditions used by Lin–Chen to control how sensitive the principal's optimum is to small relaxations of the obedience system.

Regularity assumptions (boundedness, Lipschitzness, and an inducibility gap). We maintain the uniform boundedness condition $|u|, |v| \le B$. We also assume u is L-Lipschitz in the principal decision x (with respect to a fixed norm $\|\cdot\|$ on the ambient space),

$$\big|u(x, a, \omega, \theta) - u(x', a, \omega, \theta)\big| \;\le\; L\|x - x'\| \qquad \forall x, x' \in X,$$

and write $\mathrm{diam}(X) := \sup_{x,x' \in X} \|x - x'\| < \infty$. Finally, we assume a uniform *inducibility gap* $G > 0$ that rules out near-ties in the agent's best-response problem: for every pair (s, θ) and every posterior over ω induced by some principal policy, the (possibly set-valued) best-response correspondence has a robust margin in the sense that any action a that is not a best response is worse than the best-response value by at least G.[5]

Theorem 5.1 (Main cap theorem: i.i.d. types and type-conditional no-swap-regret)**.** *Suppose the repeated interaction satisfies the timing and information conditions of Section 3, types θ_t are drawn i.i.d. from μ_0, and the agent satisfies type-conditional contextual no-swap-regret with bound* $\mathrm{CSReg}(T)$*. Let $\delta := \mathrm{CSReg}(T)/T$. Under boundedness, Lipschitzness, and a uniform inducibility gap $G > 0$, for every (possibly adaptive and algorithm-aware) principal strategy,*

$$U_T \;\le\; U^\star \;+\; K \cdot \delta,$$

where one may take, for instance,

$$K \;:=\; \frac{2BL\,\mathrm{diam}(X)}{G} \;+\; 2B,$$

and tighter constants are available under stronger conditioning assumptions (e.g., away from the boundary of the feasible signaling/decision set).

Why adaptivity and algorithm knowledge do not help. Theorem 5.1 is deliberately stated for an arbitrary principal strategy, including strategies that (i) choose π_t as an arbitrary function of the full public history, (ii) are designed with full knowledge of the agent's update rule, and (iii) attempt to correlate current signals with past play in order to "steer" future behavior. None of these freedoms violate the cap because the only channel through which they could matter is by producing an outcome distribution in which the agent systematically takes actions that would be ruled out by obedience in the one-shot benchmark. Type-conditional swap regret prevents exactly that: whatever correlation structure the principal creates, the realized actions remain (approximately) optimal given the agent's information (s_t, θ_t), and the principal cannot condition within the round on a_t to exploit the agent's randomization in a state-dependent way.

Proof sketch. We describe the logic in three steps, emphasizing where each assumption is used.

[5] This can be stated in several equivalent ways; the formulation above is convenient for bounding the probability mass on suboptimal actions under δ-approximate obedience. In persuasion-like models with constraints, an additional conditioning parameter may be needed if feasibility lies near the boundary; see Lin–Chen for sharp variants.

Step 1: From swap regret to approximate obedience on (s, θ). By Corollary 4.3, the random-round distribution $\widehat{\mathbb{P}}_T$ satisfies, for every deviation $\widetilde{d} : S \times \Theta \to A$,

$$\mathbb{E}_{\widehat{\mathbb{P}}_T}\Big[v\big(x_\tau, \widetilde{d}(s_\tau, \theta_\tau), \omega_\tau, \theta_\tau\big) - v(x_\tau, a_\tau, \omega_\tau, \theta_\tau)\Big] \;\le\; \delta.$$

Thus, conditional on each information pair (s, θ), the realized mixed action is δ-approximately optimal against the induced conditional decision $y_{s,\theta} = \mathbb{E}[x_\tau \mid s_\tau = s, \theta_\tau = \theta]$ and the posterior over ω.

Step 2: Using the gap G to control suboptimal play. Fix (s, θ) and let $a^\star(s, \theta)$ be an exact best response in the corresponding one-shot problem. Approximate obedience implies that the expected value loss (in the agent's utility) from the realized mixed action relative to $a^\star(s, \theta)$ is at most δ on average across (s, θ). The gap assumption then upgrades this utility statement into a probability statement: any mass placed on actions that are not best responses must be small, because each such action loses at least G in expected v relative to $a^\star(s, \theta)$. Concretely, if we write $\alpha_{s,\theta}$ for the conditional probability (under $\widehat{\mathbb{P}}_T$) of playing a non-best-response action at (s, θ), then $\alpha_{s,\theta} \le \delta/G$ after averaging (and, with a slightly more careful conditioning argument, pointwise for almost every (s, θ)). This is the key quantitative implication of inducibility: small δ forces the realized action distribution to concentrate near the best-response set.

Step 3: Translating approximate best response into a principal-value bound. We now compare the principal's realized payoff to the payoff it would obtain if the agent played the exact best response $a^\star(s, \theta)$ under the same induced (s, θ, x) process. Since $|u| \le B$, replacing the agent's actual action by $a^\star(s, \theta)$ can change the principal's payoff by at most $2B$ on the event that the agent played a non-best-response action, yielding a loss bounded by $2B \cdot (\delta/G)$ after Step 2. In addition, because the principal's decision x in the repeated game is itself generated endogenously and may vary with history, we use Lipschitzness to argue that it is without loss to evaluate payoffs at the conditional averages $y_{s,\theta}$ (as we already did on the agent side), at a cost controlled by $L \operatorname{diam}(X)$. Putting these pieces together bounds the repeated-game value by the value of an *exactly obedient* outcome distribution plus an additive term of order $(BL \operatorname{diam}(X)/G)\delta$. Finally, the value of the best exactly obedient outcome distribution is precisely $U^\star$ by definition of the one-shot commitment benchmark. This yields the stated bound.

Implications. Theorem 5.1 formalizes a strong form of "no dynamic persuasion rent" under our information structure: if the agent runs a sufficiently strong learning rule (type-conditional swap regret) and types are fresh each round, then the principal cannot extract more than the commitment value up to a vanishing error term. In particular, whenever $\mathrm{CSReg}(T) = o(T)$,

we obtain $\limsup_{T\to\infty} U_T \le U^\star$. From a design perspective, the bound says that improvements beyond commitment must come from precisely the channels excluded here (e.g., type persistence, action observability, or a weaker deviation class), which we isolate in Section 6.

An optional matching guarantee from below under type-conditional no-regret. While swap regret is the right notion for ruling out principal gains, it is also natural to ask whether the principal can *ensure* performance close to $U^\star$ when the agent only guarantees (type-conditional) contextual external regret on contexts (s, θ). In that case, the principal can recover the classical commitment value by simply *not* adapting.

Proposition 5.2 (Robust lower bound with a fixed principal policy)**.** *Suppose the agent satisfies type-conditional contextual no-regret with bound* $\mathrm{CReg}(T)$ *on contexts* (s, θ)*, and let* $\pi^\star$ *be an optimal one-shot commitment policy attaining* $U^\star$*. If the principal plays* $\pi_t \equiv \pi^\star$ *for all* t*, then*

$$U_T \;\ge\; U^\star \;-\; \widetilde{K} \cdot \sqrt{\frac{\mathrm{CReg}(T)}{T}},$$

for an explicit $\widetilde{K}$ *depending on* $B, L, \mathrm{diam}(X)$ *and the same conditioning parameters as in Lin–Chen.*

The message of Proposition 5.2 is complementary to the cap: even if we restrict attention to simple (non-adaptive) principal behavior, standard no-regret learning by the agent drives outcomes toward the one-shot benchmark. Thus, in our i.i.d.-type environment, the commitment solution is not only an upper bound on what an adaptive principal can achieve against a swap-regret learner, but also a natural target that a principal can approach by committing to a fixed policy when facing a no-regret learner.

6 Failure modes and tight examples

The cap in Theorem 5.1 rests on a very specific alignment between (i) what the agent can condition deviations on in its regret guarantee, and (ii) what the principal can condition its policy on in real time. In this section we make that dependence explicit by exhibiting three families of "failure modes" in which the principal can recover an $\Omega(1)$ dynamic advantage even though the agent is, in an appropriate sense, learning well. These examples are not pathologies: each corresponds to a familiar channel of manipulation in applied settings (targeting a subgroup, screening over time, and reacting to behavior within a round).

6.1 Omitting θ from deviations: how to exploit a minority type

Our reduction from swap regret to approximate obedience crucially used deviations of the form $d(s, a, \theta)$. If we only control deviations $d(s, a)$ that cannot condition on the agent's private type, then the induced notion of "approximate obedience" is pooled across types. This pooling leaves room for a principal to concentrate violations on a subset of types, so long as those violations are hard to detect by any single remapping $d(s, a)$ that must apply uniformly to everyone.

A concrete way to see the issue is to consider two types, $\Theta = \{\theta^H, \theta^L\}$, with $\Pr(\theta^H) = \varepsilon$ small. Suppose there is a single public signal s (or the principal uses the same s always), two actions $A = \{a_0, a_1\}$, and the principal decision x can be interpreted as choosing a "recommendation intensity" in $X = [0, 1]$. The agent's payoff is type-dependent: type θ^H strongly prefers a_1 when x is high, while type θ^L strongly prefers a_0 regardless of x. Formally, one can choose $v(x, a, \theta)$ so that

$$a^\star(s, \theta^H) = a_1, \qquad a^\star(s, \theta^L) = a_0$$

for the relevant posteriors induced by the principal. In the one-shot commitment problem, the principal cannot condition on θ, so any policy trades off these responses and yields some $U^\star$.

In the repeated game, however, an adaptive principal can interleave "probing" policies that induce the learning algorithm to occasionally play the wrong action for θ^H while leaving θ^L almost always playing its correct action. The key point is that a deviation $d(s, a)$ that flips $a_0 \mapsto a_1$ (or vice versa) helps one type but harms the other; when ε is small, any uniform remapping has negligible aggregate benefit even if it would be very valuable for θ^H specifically. Thus the non-type-conditional swap-regret guarantee can remain small:

$$\forall d : S \times A \to A, \qquad \sum_{t=1}^{T} \Big(v(x_t, d(s_t, a_t), \omega_t, \theta_t) - v(x_t, a_t, \omega_t, \theta_t) \Big) \text{ is small,}$$

even though there exists a type-targeted remapping $d(s, a, \theta)$ that obtains a large improvement concentrated on θ^H rounds. In other words, the principal can create an outcome distribution that is approximately obedient only after averaging over types, while violating obedience substantially conditional on θ.

Economically, this example captures a simple but important phenomenon: if our behavioral guarantee does not allow the agent to "audit" its own performance type-by-type, then a principal can profit by shifting mistakes onto a subgroup (a minority preference segment, a vulnerable user group, or a rare context) without triggering a large regret signal that is aggregated across

the whole population. This is exactly why we require type-conditional contextual swap regret in the cap theorem: it enforces approximate obedience *within* each (s, θ) cell rather than merely on average.

6.2 Persistent types: screening restores dynamic leverage

The i.i.d. type assumption is not merely a technical convenience; it is what prevents the principal from using the repeated interaction as a screening device. If each agent has a persistent type θ across rounds (or, more generally, if θ_t is sufficiently correlated over time for an individual), then actions become informative about θ and an adaptive principal can condition future policies on past behavior. This is the standard logic of dynamic price discrimination and sequential persuasion: early rounds are used to learn who the agent is, and later rounds extract surplus accordingly.

A stylized construction is as follows. There are two types $\theta \in \{\theta^H, \theta^L\}$ and two actions $A = \{\mathsf{reveal}, \mathsf{hide}\}$. The principal controls a decision $x \in [0, 1]$ that determines how much value the agent gets from revealing (e.g., an "explanation" level, a discount, or an information disclosure). Type θ^H is willing to reveal even at low x, while type θ^L reveals only when x is high. In a one-shot problem where the principal cannot observe θ, the commitment solution must pick a compromise x (or a lottery over x) and achieves $U^\star$.

Now suppose the principal observes realized actions (or any outcome correlated with them) and types persist. The principal can run a two-phase policy. In an exploration phase of length T_0, choose an x that separates types in the sense that θ^H reveals with high probability while θ^L hides with high probability. Because each type is approximately best-responding (indeed, exact best-responding in the cleanest versions), the exploration phase produces an informative statistic about θ:

$$\Pr(\mathsf{reveal} \mid \theta^H) \approx 1, \qquad \Pr(\mathsf{reveal} \mid \theta^L) \approx 0.$$

In the exploitation phase, the principal conditions on the inferred type and switches to a type-tailored policy $x(\widehat{\theta})$ that yields strictly higher expected principal payoff than any single compromise policy can achieve ex ante. The overall gain is $\Omega(1)$ because the exploration cost is $O(T_0/T)$ and can be made negligible by taking $T_0 = o(T)$.

What is important for our purposes is that this leverage is compatible with very strong learning guarantees on the agent side. Even if, ex post, the agent has type-conditional no-swap-regret on the realized sequence, the principal's policy can still outperform $U^\star$ because the principal is no longer solving a one-shot Bayesian persuasion problem; it is solving a *dynamic* problem with an additional state variable given by its posterior about the persistent θ. The cap fails because the feasible set expands: the principal can condition future π_t on a statistic that is informative about θ, whereas in the i.i.d. model there is no such statistic beyond the prior.

This example delineates a clear modeling lesson. If we want a cap theorem that speaks to repeated interactions with the *same* user, then we must either (i) incorporate persistence explicitly and accept that $U^\star$ is not the right benchmark, or (ii) impose additional restrictions (e.g., privacy constraints that prevent the principal from conditioning on past actions, or commitment to a stationary policy) that remove the screening channel.

6.3 Within-round observability: reacting to a_t breaks the reduction

Our timing assumption that the principal chooses π_t before a_t is realized is not innocuous. It is exactly what supports the conditional-independence step that collapses obedience on (s, a, θ) to obedience on (s, θ): informally, the agent's own randomization cannot convey additional information about ω_t (or about its private type) back to the principal *within the same round*. If the principal can instead observe a_t before finalizing x_t (or before sending the payoff-relevant component of the signal), then the principal can use the agent's action as an extra message—one that is often highly informative about θ_t.

To illustrate, consider a within-round two-stage variant: the principal first sends a preliminary signal s_t, the agent chooses a_t, the principal observes a_t, and then the principal chooses a decision $x_t = x(s_t, a_t, \omega_t)$ that determines payoffs. Even if the agent has vanishing swap regret relative to deviations $d(s, a, \theta)$, the principal can design s_t so that different types take different actions (because v depends on θ), thereby eliciting a type-revealing a_t. The principal then conditions x_t on a_t to implement a type-contingent allocation. In effect, the principal has endogenously created a direct revelation mechanism inside the round.

From the perspective of the commitment benchmark, this is a strict expansion of the principal's instrument set. The classical $U^\star$ only allows the principal to commit to a mapping $\omega \mapsto \Delta(S \times X)$, after which the agent acts and x is already fixed. Once x can depend on a, the principal can replicate (approximately) the outcome of a mechanism that conditions on the agent's report of θ—even if the agent never explicitly reports θ.

A useful way to characterize when exploitation becomes possible in within-round observability variants is to ask whether the principal can implement a nontrivial correspondence $a = a(s, \theta)$ that is (approximately) incentive compatible for the agent and sufficiently informative for the principal. If such an $a(\cdot)$ exists and the principal can respond with $x(s, a, \omega)$, then the relevant benchmark is no longer the persuasion/commitment value $U^\star$ but rather the value of a richer mechanism-design problem in which the principal can condition on an endogenous message. In that richer problem, "dynamic" gains can appear even in a single round.

6.4 Tightness and what the examples teach us

Taken together, the three failure modes identify the precise boundaries of the cap theorem. If we keep i.i.d. types and the within-round nonobservability of actions, then strengthening the agent guarantee from non-type-conditional to type-conditional swap regret is exactly what blocks targeted exploitation. If we instead keep type-conditional swap regret but allow persistence or within-round observability, then the principal can recover a qualitatively new channel—learning and reacting to type—that is absent from the one-shot benchmark and can generate an $\Omega(1)$ gain.

Finally, these constructions also clarify why the $O(\delta)$ dependence in Theorem 5.1 is the right scale under our regularity assumptions. When there is a uniform gap G, any δ-approximately obedient play must place only $O(\delta/G)$ probability on strictly suboptimal actions, and this is the only place where the principal can earn additional surplus beyond $U^\star$ in the i.i.d. model. The separations show that once we remove the structural reasons the principal cannot screen (fresh types, no within-round reaction), the principal can do better than $U^\star$ even when δ is essentially zero.

7 Applications and calibrations

The preceding theorems are intentionally abstract: they treat the principal as an arbitrary adaptive optimizer, and they summarize the agent side by a single behavioral primitive—type-conditional contextual no-swap-regret. This abstraction is useful precisely because many applied environments can be re-expressed in this template by choosing (i) what counts as the principal's decision variable $x \in X$, (ii) what information the principal can encode in a signal $s \in S$, and (iii) what we mean by the agent's "type" $\theta \in \Theta$. In this section we sketch three domains where the cap has immediate interpretive value, and we explain how one can calibrate the $O(\delta)$ term, with $\delta = \mathrm{CSReg}(T)/T$, into a concrete notion of "how much dynamic advantage is left on the table."

A recurring theme is that the cap is not a statement about benevolence. It says that if the agent's behavior is approximately obedient *type-by-type* (in the precise internal-regret sense) and the principal cannot learn the type within the round, then the principal's additional leverage from repeated interaction is quantitatively limited. Conversely, when applied settings violate i.i.d. types or within-round nonobservability, the separation examples in Section 6 should be read as a warning that repeated interaction itself becomes a screening technology.

7.1 Personalized disclosure and compliance warnings

A natural interpretation of our model is personalized disclosure: the principal is a platform deciding how to present information about a risky choice, and the agent is a user who chooses whether to comply with a recommended safe action. Concretely, let $A = \{\mathsf{comply}, \mathsf{ignore}\}$ and let $x \in [0, 1]$ be a "warning intensity" or "disclosure level" (e.g., salience, friction, or the specificity of an explanation). The state ω captures the objective risk (or the platform's private assessment of harm), while the type θ captures user-side preferences such as risk tolerance, impatience, or susceptibility to framing.

In the one-shot benchmark, the principal commits to a signaling scheme π that maps ω to a joint distribution over (s, x), anticipating that the user best-responds given (s, θ). The value $U^\star$ then represents the maximal expected platform objective—which might be profit, engagement, or some blended welfare criterion—subject to users responding optimally to the disclosed information and warnings. Our cap theorem implies that if each interaction draws a fresh θ_t (e.g., heterogeneous users arriving i.i.d. each visit) and the user-side decision rule satisfies type-conditional contextual no-swap-regret, then even a sophisticated platform that adapts warnings to past outcomes cannot exceed $U^\star$ by more than $K\delta$.

This has two practical readings. First, the relevant risk in repeated deployments is not simply that users are "boundedly rational," but that their mistakes might be *systematically targetable*. The failure mode in Section 6.1 corresponds here to a platform that induces miscompliance disproportionately among a small segment (say, a vulnerable subgroup) while aggregate behavioral metrics look well-calibrated. Requiring type-conditional swap regret—where θ indexes the segment or the user context of concern—is a way to formalize an anti-targeting constraint: it forces approximate obedience *within* each (s, θ) cell rather than merely on average.

Second, the bound can be calibrated in operational terms. Suppose utilities are normalized so $|u|, |v| \le B$ and the platform's decision-to-payoff map is L-Lipschitz in x (e.g., small changes in warning intensity only gradually change outcomes). Then any standard no-swap-regret rate (say, $\mathrm{CSReg}(T) = O(\sqrt{T})$ up to logarithmic factors in $|A|$) yields

$$U_T \le U^\star + O\left(\frac{1}{\sqrt{T}}\right),$$

so that the maximal "dynamic premium" decays at the familiar statistical rate. In deployments where the same disclosure policy is used millions of times, this suggests that ensuring the user-side policy class actually satisfies a *type-conditional* internal-regret guarantee may be more important than squeezing constants: qualitatively, it is the difference between protection against subgroup manipulation versus protection only in the aggregate.

7.2 Platform pricing with ephemeral versus persistent preferences

A second canonical application is platform pricing. Let $A = \{\mathsf{buy}, \mathsf{no\ buy}\}$ and let x encode a posted price or a menu (so X is a simplex over price points, or a compact interval). The type θ is willingness-to-pay or outside option, and the state ω can represent supply conditions, marginal cost, or an inventory shock observed by the platform. The signal s can include coupons, personalized messages, or product rankings that shift demand.

When preferences are ephemeral—for example, one-off visits drawn from a population distribution μ_0, or a setting where each round corresponds to a different user—the i.i.d. type assumption is a reasonable approximation. In that case, our cap theorem says that repeated interaction does not create a large additional degree of freedom for the platform: even if the platform is algorithm-aware and adaptively changes prices and messages, its expected average payoff is close to what it could have committed to in a one-shot mechanism, up to $K\delta$. In particular, if the buyer-side decision policy is the output of a learning system that minimizes type-conditional internal regret (with θ indexing user segments), then the platform cannot systematically "shape" purchase mistakes beyond what is already achievable via commitment.

The picture changes sharply under persistent preferences. If the same user returns and has stable willingness-to-pay, then the platform can treat early prices as a screening device, infer θ from purchase decisions, and later tailor offers. This is exactly the dynamic price discrimination logic, and it aligns with our Separation 2: even if the buyer's behavior is near-best-responding each round (hence has low swap regret), the platform may exceed the one-shot commitment value because the feasible set expands to include policies contingent on an endogenous estimate of θ.

This contrast helps interpret real-world policy interventions. Privacy restrictions (limits on cross-round tracking, cookie expiration, data minimization) can be seen as engineering the environment back toward the i.i.d. benchmark by preventing the principal from conditioning future policies on past actions. Likewise, commitments to stationary pricing rules, or to coarse targeting, can restore a setting where $U^\star$ becomes a meaningful normative baseline. From a calibration standpoint, the cap also provides a way to quantify residual harm from "learning-induced" suboptimality: if we can bound δ for the buyer-side decision system in each segment, then $K\delta$ is an upper bound on the platform's extra profit (or extra distortion) attributable to deviations from segment-wise best response, holding fixed the platform's commitment problem.

7.3 Delegation to AI assistants and auditing type-conditional internal-regret

A third domain—and the one most directly connected to current practice—is delegation to an AI assistant that acts on behalf of a user while interacting with a strategic platform. Here the "agent" in our model is the assistant's action-selection module, $a_t \in A$ could be a query reformulation, a click/no-click decision, or acceptance of a recommendation, and the principal is the platform that controls x_t (ranking weights, pricing, disclosure, or the allocation of attention) and the signal s_t (the content shown, explanations, or interface cues). The assistant's private type θ_t represents user intent or preference—often partially observed by the assistant but not by the platform—such as time sensitivity, political preference, or risk posture.

In this setting, type-conditional contextual no-swap-regret is not merely a mathematical convenience; it is a candidate *design requirement* for the assistant. Intuitively, the assistant should be robust not only to simple action switches but also to context-dependent remappings that exploit the assistant's own stochastic recommendations. Technically, internal regret is what prevents the platform from benefiting when the assistant sometimes "talks itself into" a suboptimal action in a way that can be targeted by the platform's choice of s and x.

This perspective suggests an auditing program. Fix a choice of contexts (s, θ), where θ may be a coarse label available to the assistant (task category, declared user goal, or a privacy-preserving segment). From logged interaction data $\{(s_t, \theta_t, a_t, x_t)\}_{t=1}^{T}$, one can estimate the left-hand side of the swap-regret inequalities for a rich family of deviations $d : S \times A \times \Theta \to A$. While enumerating all such d is infeasible in general, two pragmatic relaxations are common: (i) restrict to a parametric family of deviations (e.g., small action remappings within a task class), and (ii) use duality-style certificates that upper bound the worst-case deviation gain without searching over d explicitly. Either route yields an empirical upper bound on $\mathrm{CSReg}(T)$, and hence an interpretable bound $K\delta$ on the principal's potential dynamic advantage under the assumptions of the cap theorem.

We emphasize two limitations that matter in deployments. First, many assistant decisions are made under partial monitoring (bandit feedback): the assistant may not observe counterfactual utilities $v(x_t, a, \omega_t, \theta_t)$ for unchosen actions. Swap-regret guarantees exist in such settings but typically require explicit exploration, which can be costly or unsafe. Second, choosing the right notion of type θ is itself a governance choice: if θ is too coarse, the guarantee reverts to pooled obedience and becomes vulnerable to subgroup exploitation; if θ is too fine, it may be unobservable, unstable, or privacy-sensitive. Our framework does not resolve this tradeoff, but it clarifies what is at stake: type-conditional internal-regret is precisely the property that blocks a principal from concentrating harm (or surplus extraction) on identifiable

subpopulations.

Across these applications, the unifying message is that learning guarantees can function as a policy instrument. By specifying *which* deviations the agent must control (in particular, allowing conditioning on θ) and by restricting the principal's ability to condition within-round on realized actions, we can make the commitment benchmark $U^\star$ a robust ceiling on strategic advantage from repetition. The next section discusses how these ideas extend—and where they break—in multi-agent environments, under richer feedback models, and under computational constraints.

8 Discussion and future work

Our results isolate a stark but, we believe, practically relevant lesson: when types are effectively i.i.d. across interactions and the principal cannot condition within-round on realized agent actions, then the principal's dynamic leverage is largely exhausted by the one-shot commitment benchmark, up to an explicit $O(\delta)$ slack with $\delta = \mathrm{CSReg}(T)/T$. The abstraction is deliberate: we compress the entire agent side into a single behavioral primitive (type-conditional contextual no-swap-regret) and allow the principal to be fully adaptive and algorithm-aware. This makes the cap robust, but it also highlights where the next modeling choices matter most. We view the main open directions as (i) extending from a single agent to populations of agents and strategic interaction among them, (ii) handling partial monitoring and the attendant exploration incentives, (iii) incorporating computational constraints on the principal, and (iv) treating learning guarantees themselves as an object of mechanism and policy design.

8.1 Populations, interaction, and what replaces obedience

Many deployed settings are intrinsically multi-agent: a platform chooses rankings or prices that affect many users in parallel; a sender broadcasts a disclosure policy to a population; or a recommender system allocates attention across a marketplace of buyers and sellers. A first extension of our framework keeps the basic within-round timing but replaces a single agent by a finite set of agents $i \in \{1, \dots, n\}$, each with type θ_t^i and action a_t^i. The principal chooses a joint policy π_t that may generate individualized signals s_t^i and decisions x_t^i subject to feasibility constraints (e.g., capacity or market clearing). If each agent runs a type-conditional no-swap-regret algorithm with respect to its own context (s_t^i, θ_t^i), then the natural analogue of our reduction proposition is that the empirical joint distribution over $(s^1, \theta^1, a^1, \dots, s^n, \theta^n, a^n, x)$ approaches an *approximate Bayes coarse correlated equilibrium* (or a Bayes correlated equilibrium when we track joint signals) of the induced one-shot game. In other words, swap regret no longer enforces "obedience to a best response" in isolation; it enforces approximate

obedience to *a best response given the correlation device* represented by the signal and the induced distribution of other agents' play.

This change is not cosmetic. In a multi-agent environment the principal's objective may depend on coordination or externalities across agents (congestion, matching quality, network effects), and the set of feasible outcomes under commitment can be strictly larger or smaller depending on whether the principal can correlate agents' information. A cap theorem in this setting would need to compare the principal's dynamic payoff not to a single-agent $U^\star$, but to the appropriate *multi-agent commitment value* under Bayes correlated equilibrium constraints. The encouraging part is that the same proof architecture appears viable: swap regret controls deviations that remap actions after seeing the agent's own recommendation, which is precisely the obedience constraint defining correlated equilibrium-like objects. The hard part is identifying the right "no within-round screening" assumption. When users arrive sequentially within a round (or the principal can observe early actions before choosing later allocations), the principal effectively regains the ability to condition on endogenous information about types, and the multi-agent analogue of Separation 2 becomes immediate: even with low internal regret for each individual, the principal can implement dynamic screening across the population by ordering, throttling, or selectively experimenting.

A related open question concerns *heterogeneous notions of type*. In population settings, θ may represent a protected attribute, a task class, or an unobserved preference vector. Our separation showing that omitting θ from the deviation class reintroduces exploitable slack suggests that the "right" segmentation is not purely statistical; it is strategic. Characterizing the coarsest partition of contexts that still blocks profitable targeting (i.e., the coarsest θ for which type-conditional swap regret suffices) is a natural direction, and it would connect our framework to literatures on algorithmic fairness, subgroup robustness, and multi-calibration.

8.2 Partial monitoring, exploration, and manipulation of feedback

The auditing and design motivation for swap regret runs through the availability of feedback: the regret inequality is defined relative to counterfactual payoffs $v(x_t, a, \omega_t, \theta_t)$, but many assistants and users do not observe these counterfactuals. In partial monitoring or bandit feedback, the agent may observe only $v(x_t, a_t, \omega_t, \theta_t)$ (possibly with noise), and the distribution of contexts (s_t, x_t) is itself chosen by an adaptive principal. Two issues arise.

First, obtaining swap-regret guarantees under partial monitoring typically requires explicit exploration. Exploration has a direct welfare cost and may be unsafe (e.g., trying inferior medical advice to learn). In our setting it also has a strategic cost: an adversarial principal may shape the context sequence to make informative exploration disproportionately costly precisely

in the contexts where manipulation is profitable. This suggests that the relevant learning primitive may need to be strengthened from "low regret under realized play" to "low regret under strategically chosen information structures," perhaps by requiring uniform exploration or by bounding regret conditional on each (s, θ) cell receiving sufficient mass. Formalizing such conditions would sharpen the gap between what can be certified from logs and what is required for a robust cap.

Second, partial monitoring blurs the line between obedience and identification. In the full-information case, swap regret implies approximate obedience without requiring the agent to ever reveal its type. Under bandit feedback, however, the agent's exploration policy can leak information (through randomized actions) that a principal may aggregate across rounds, and this again interacts with persistence: even if θ_t is i.i.d., the principal may infer distributional properties of the agent algorithm that enable within-round steering of the signal distribution. One promising direction is to combine our conditional-independence logic ($\omega_t \perp a_t \mid (s_t, \theta_t)$) with *information-theoretic* caps that quantify how much extra information about θ_t can be transmitted through exploration noise, and how that affects the achievable deviation from $U^\star$.

8.3 Computationally bounded principals and algorithmic commitment

Our cap theorem treats the principal as an arbitrary adaptive optimizer, which is appropriate for an upper bound but can be pessimistic as a description of real systems. Computing $U^\star$ already subsumes nontrivial Bayesian persuasion and Stackelberg problems; in many environments, optimizing over π is NP-hard, and platforms deploy heuristics that are better interpreted as online learning algorithms on the principal side as well. Introducing computational constraints raises two complementary questions.

On the one hand, bounding principal computation may *tighten* the cap: even if additional dynamic advantage exists in principle, it may not be algorithmically realizable without solving hard inference or planning problems. This points toward "computational caps" where the benchmark is the best efficiently computable commitment policy, and the additive term depends not only on δ but also on the principal's optimization error. On the other hand, computational constraints may *weaken* protection if they force the principal to rely on proxy objectives that correlate with manipulable behaviors. In such cases, the agent's learning guarantee could be targeted at the proxy rather than at true utility v, undermining the intended obedience interpretation.

A technical direction we find particularly attractive is to place both players within a common online optimization template. If the principal itself runs a no-regret or no-swap-regret algorithm over a tractable policy class,

then the interaction may converge to an equilibrium object (e.g., a Bayes coarse correlated equilibrium of a meta-game between policy classes). Understanding how our $O(\delta)$ cap composes with principal-side regret bounds could yield sharper, more operational predictions for learning-to-rank and pricing systems that are updated continually.

8.4 Learning guarantees as a policy instrument

Finally, we think the most immediately actionable implication of our framework is that *the choice of learning guarantee is itself a governance lever.* Standard discussions of "rationality" focus on whether the agent is optimizing; our results suggest that *which deviations are ruled out* matters just as much. Requiring type-conditional contextual swap regret, rather than pooled regret, is precisely what blocks the principal from concentrating gains (or harms) on identifiable subpopulations while maintaining good average performance.

This invites a design problem: choose a type system Θ (a segmentation), a deviation class (e.g., all $d : S \times A \times \Theta \to A$ or a restricted family), and an auditing procedure that together imply a meaningful bound on the principal's extra surplus from repetition. The companion enforcement levers are environmental: privacy and logging restrictions can be interpreted as engineering the nonobservability and i.i.d. conditions under which the cap is valid, while interface design can limit within-round conditioning on actions. In settings where these conditions cannot be guaranteed (persistent users, rich within-round observability, or sequential arrival), our separation results suggest that guarantees on the agent side must be complemented by institutional constraints on the principal's ability to screen.

Several theoretical challenges remain before such "regret-based regulation" is mature. We would like bounds that relax linearity in x (e.g., smooth non-linear utilities), handle continuous action spaces, and replace the uniform inducibility gap with weaker margin conditions that are verifiable from data. We would also like to characterize the minimal feedback requirements under which type-conditional swap regret is achievable without excessive exploration. More broadly, our cap should be read as a map of where dynamic manipulation can and cannot hide: it identifies the precise points—type persistence, within-round observability, and the granularity of deviation control—at which repeated interaction becomes a screening technology rather than a mere repetition of a one-shot commitment problem.