Robust Nudging in Sequential Contracting: Distributionally Robust Incentive Compatibility for Principal–Agent RL

Liz Lemma Future Detective

January 16, 2026

Abstract

We study sequential contract design for orchestrating autonomous agents when the principal faces three deployment-era frictions: (i) the outcome model linking hidden actions to observable outcomes is estimated and may shift; (ii) the agent's continuation values are misspecified due to approximation and strategic learning; and (iii) contracts are computed from learned Q-functions. Building on principal–agent reinforcement learning with subgame-perfect equilibrium (SPE), we propose a distributionally robust contract design method that converts estimated minimal-implementation contracts into deployment-safe contracts via 'robust nudges'—small additional payments that stabilize the agent's best response under model uncertainty. Our core technical contribution is a clean per-state robust incentive-compatibility (IC) program that is linear/convex for standard uncertainty sets and yields closed-form nudging rules in low-dimensional cases. We introduce an identifiable signal-separability parameter that generalizes the overlap measure $d_{\min}$ in prior analysis: when separability is low, any contract is inherently fragile and requires large subsidies. We prove additive welfare guarantees for the principal under validation against a best-responding oracle agent, decomposing losses into (a) principal learning error and (b) robustness cost that scales inversely with separability. Empirically, we validate on adversarially shifted variants of sequential social dilemmas (e.g., Coin Game) and show robust nudging prevents incentive flips and preserves welfare at modest budget increases. The results provide an implementable safety layer for contract-based governance of agent ecosystems in 2026 settings, where principals must operate under distribution shift, limited observability, and learning agents.

Table of Contents

tems; contributions and roadmap.

2. 2. Baseline principal–agent MDP and SPE: Restate hidden-action principal–agent MDP; SPE meta-algorithm; where LP-based minimal implementation appears; failure modes under misspecification.

3. 3. Robust modeling primitives: Uncertainty sets for outcome models $\mathcal{U}_{s,a}$; bounded agent-value misspecification ε_s; principal Q error δ_s; discussion of what is observable/estimable from data.

4. 4. Robust IC and robust nudging: Define robust IC constraints as worst-case expected utility inequalities; define robust separability $\bar{d}$; derive the nudging margin needed for compliance.

5. 5. Closed-form contracts in tractable cases: Two-actions/two-outcomes (or finite $|\mathcal{O}|$) with TV balls; show 'single-diagnostic-outcome' payments; extend to f-divergence balls via convex duality; flag when numerical convex programs are needed.

6. 6. Welfare guarantees in sequential settings: Backward-induction-style bound for finite horizon; extension sketch to infinite horizon with discounting and regularity; decompose regret into learning error and robustness cost.

7. 7. Minimax lower bounds and necessity of separability: Show any robust-IC contract must pay at least $\Omega(\varepsilon/\bar{d})$; interpret $\bar{d}$ as an inherent 'price of unobservability + shift.'

8. 8. Algorithms: Robust LP/convex program per state; plug-in estimators for $\mathcal{U}$; practical heuristics for large $|\mathcal{O}|$ and deep RL (e.g., outcome classifiers, robust margins).

9. 9. Experiments: Adversarially shifted tree MDPs; Coin Game with controlled outcome perturbations and non-stationary agents; compare nominal vs robust nudging; budget–welfare tradeoff curves; separability diagnostics.

10. 10. Discussion and extensions: Multi-agent robustness (dominant-strategy IC vs learning-stable IC); partial observability; budget constraints; policy implications for platform design and regulation.

1 1. Introduction: Why learned contracts are fragile under distribution shift; robustness as an essential governance layer in 2026 agent ecosystems; contributions and roadmap.

In contemporary digital economies, contracts are increasingly not handwritten documents but learned objects: platform subsidy schedules that adjust to user behavior, performance-pay schemes for gig workers, incentive rules for content moderation and ranking, and, more recently, "contract-like" payment and access controls governing tool-using AI agents. In these settings the principal is typically operating a complex dynamical system, and the incentive rule is computed from data, simulation, or a learned model of how outcomes respond to hidden actions. This evolution creates an uncomfortable tension. On the one hand, learned contracts are attractive precisely because they can exploit fine-grained predictive signals and adapt to high-dimensional environments. On the other hand, the same dependence on predictive models makes incentive schemes fragile: small distribution shifts can flip best responses, and the resulting deviations can propagate through time, amplifying losses in dynamic systems.

We study this fragility in the canonical place it arises: a finite-horizon hidden-action principal–agent Markov decision process (MDP) with limited-liability, outcome-contingent contracts. The principal posts a nonnegative payment schedule $b(\cdot)$ over observable outcomes, the agent privately chooses an action a, an outcome o is realized, transfers occur, and the system transitions. The principal chooses contracts using an estimated environment model and an estimated description of the agent's continuation incentives. In deployment, however, the true outcome distribution may differ from the estimate, and the principal's continuation-value proxy for the agent may be misspecified. These two sources of misspecification are not merely technical nuisances. They are precisely the forms of error that arise when we train incentive rules on historical data, validate them on offline simulators, and then ship them into environments that evolve—because of seasonality, adversarial adaptation, novel products, new cohorts of users, or strategic gaming.

Why does a small modeling error matter so much? The reason is that incentive compatibility is a knife-edge property. Contracts implement actions by making one action *strictly* better than another in expected utility terms. Under limited liability $b(o) \geq 0$, we cannot "fine-tune" incentives by punishing deviations; we can only reward certain outcomes. When two actions induce similar outcome distributions, the principal has only weak statistical leverage: expected payments differ little across actions unless the contract has large payouts concentrated on rare but diagnostic events. In that regime, even a modest shift in outcome probabilities or a modest error in how we value the agent's continuation payoffs can reverse the ranking of

actions. The principal then faces a dual risk: an *implementation risk* (the agent chooses a different action than intended) and a *dynamic compounding risk* (the induced state trajectory changes, so subsequent contracts are computed in the wrong region of the state space).

This compounding risk has become more salient in 2026-era agent ecosystems. Many principals now govern not only human agents but also automated agents that can search, plan, and exploit system loopholes at high speed. These agents respond sharply to incentives, and their behavior can shift the environment itself (e.g., changing congestion, liquidity, or the distribution of observable outcomes). In such systems, robust incentive design functions as a governance layer: not an after-the-fact audit, but an ex ante guarantee that the deployed contract remains incentive compatible across a specified set of plausible shifts. Robustness is therefore not simply a conservative preference; it is an operational requirement when contracts are computed by learning pipelines that are inevitably imperfect.

Our starting point is the observation that principals already implicitly reason about robustness, but often in an ad hoc way: adding "buffers" to performance bonuses, capping certain rewards, or requiring manual review when a model extrapolates. We aim to formalize a principled version of this buffer, one that is explicitly tied to two interpretable quantities: (i) how uncertain the principal is about the mapping from actions to outcomes, and (ii) how wrong the principal might be about the agent's continuation values under future contracts. In our framework, the principal possesses a nominal outcome model $\hat{O}(s, a)$ and an uncertainty set $\mathcal{U}_{s,a}$ that contains the true $O(s, a)$. Separately, the principal computes contracts using an estimated truncated continuation value $\widehat{Q}(s, a)$ for the agent, but acknowledges a bounded misspecification ε_s at each state. The key design question becomes: how should the principal translate $(\mathcal{U}, \varepsilon)$ into a simple, statewise modification of the learned contract that restores incentive compatibility in deployment, while incurring minimal additional expected payments?

A second observation motivates our approach: although the environment is dynamic, the most immediate failure mode is local. At a given state s, a learned contract is computed to implement an intended action a_p. If, under the true model, the agent instead prefers some deviation $a \neq a_p$, then the remainder of the dynamic plan is moot. This suggests a modular architecture. We can treat the principal's learning and planning pipeline as producing a *recommended* action $a_p(s)$ (and a nominal contract), and then apply a lightweight robustification step—a "nudge"—that enforces incentive compatibility at the current state against worst-case distributions in $\mathcal{U}$ and worst-case value misspecification ε_s. The result is an implementable governance layer: a per-state rule that is easy to compute, easy to audit, and easy to stress-test because it depends only on explicit uncertainty sets and explicit error budgets.

Technically, the central object governing whether nudging is cheap or

expensive is a robust notion of outcome informativeness. Informally, we ask whether there exists some observable outcome that remains *uniformly* more likely under the intended action than under any deviation, even after we allow Nature to choose the worst-case distribution in each uncertainty set. When such a diagnostic outcome exists, paying only on that outcome creates leverage: the expected payment advantage of the intended action is proportional to the robust gap in probabilities. When the robust gap is small, the same incentive margin requires larger payments. This logic mirrors classical results on identification and contracting under limited liability, but here it becomes a quantitative robustness calculus: a map from statistical distinguishability (under uncertainty) to subsidy costs.

Our contributions develop this calculus in a dynamic setting and connect it to learned contracting practice.

- *Robust incentive compatibility with misspecified continuation values.* We formulate a robust IC constraint that accounts simultaneously for distribution shift $O \in \mathcal{U}$ and for bounded error ε_s in the principal's proxy $\widehat{Q}(s, a)$ for the agent's continuation payoff. The resulting constraint has a transparent interpretation: the contract must create an expected-payment advantage large enough to dominate the worst-case value misspecification.

- *A constructive "single-diagnostic-outcome" nudge.* Under a mild separability condition—that the intended action is robustly distinguishable from every deviation by at least one outcome—we show that robust IC can be achieved by a nonnegative contract supported on a single outcome. This contract is not merely convenient; it is the simplest possible governance intervention: it adds mass to one payment entry $b(o^\star)$ and leaves the remainder unchanged.

- *Closed forms under standard uncertainty sets.* For commonly used ambiguity sets, including total-variation balls and f-divergence balls, we derive explicit expressions for the robust separability gap and hence explicit robust nudging payments. This matters for practice because uncertainty sets are often chosen to match statistical confidence regions, and closed forms enable fast deployment-time computation.

- *A sequential validation guarantee.* We show that if the principal's recommended-action policy is approximately optimal up to a learning error δ_s, then adding the robust nudge at each visited state yields a lower bound on the principal's realized value in validation against an oracle best-responding agent and any $O \in \mathcal{U}$. The bound decomposes into a term for planning/learning suboptimality (the δ's) and a term for robustness payments that scales as ε_s divided by robust separability.

- *A necessity result.* In a single-state environment, we provide a minimax lower bound showing that any limited-liability contract guaranteeing implementation under misspecification must incur expected payments on the order of $\varepsilon_s/\bar{d}$. This establishes that the basic tradeoff we highlight—robustness requires subsidies when outcomes are weakly informative—is not an artifact of our construction.

Beyond the formal results, we emphasize an interpretation relevant for governance: robustness here is not "free safety." It is a priced constraint that depends on two design levers. First, one can invest in better prediction and better uncertainty calibration, shrinking $\mathcal{U}$ and thereby increasing robust separability. Second, one can invest in better behavioral modeling of the agent's continuation values, shrinking ε_s. Our welfare guarantee makes these levers commensurable: it expresses, in the same units as the principal's value, how much each source of misspecification costs once we require incentive compatibility in deployment. This provides a practical decision tool: when should a principal pay for more data collection, better simulation, or richer agent modeling, versus when should it simply pay larger subsidies to stabilize behavior?

Several limitations are worth flagging at the outset. We assume limited liability and focus on nonnegative outcome-contingent payments, which is appropriate for many subsidy and bonus schemes but not for settings where fines or clawbacks are enforceable. We also posit that the principal can specify credible uncertainty sets $\mathcal{U}_{s,a}$ and bounds ε_s. In practice, these may themselves be learned or contested, and misspecification of uncertainty sets is a genuine risk. Finally, our separability condition can fail in environments where actions are observationally nearly equivalent; in such cases, robust implementation may be prohibitively expensive or impossible without richer observables or monitoring technology. We view this not as a drawback but as a diagnostic: the model tells the principal when a desired behavior cannot be robustly incentivized under the available signals.

Roadmap. In the next section we formalize the baseline principal–agent MDP and the relevant equilibrium notion, highlighting where standard minimal-implementation linear programs arise and why they can fail under model and value misspecification. We then introduce our robust separability measure and derive the nudging rule that restores incentive compatibility under distribution shift, before turning to closed-form expressions for standard ambiguity sets, sequential validation guarantees, and lower bounds that clarify the fundamental tradeoff between robustness and subsidy costs.

2 Baseline principal–agent MDP and SPE

We begin with the baseline dynamic contracting problem absent any robustness modifications. The environment is a finite-horizon hidden-action

principal–agent MDP. At each date $t \in \{0, \dots, T-1\}$ the system occupies a publicly observed state $s_t \in \mathcal{S}$. The principal posts a limited-liability, outcome-contingent contract $b_t \in \mathbb{R}^{|\mathcal{O}|}_{\geq 0}$, where $b_t(o)$ is the transfer paid to the agent if outcome $o \in \mathcal{O}$ is realized at t. The agent then privately selects an action $a_t \in \mathcal{A}$. Nature draws an observable outcome o_t according to $O(s_t, a_t) \in \Delta(\mathcal{O})$, transfers are executed, and the system transitions to the next state $s_{t+1} \sim T(s_t, o_t)$. The principal observes (s_t, o_t) but not a_t; the agent observes s_t and b_t when choosing a_t.

Per-period payoffs take the additive form

$$R_A(s, a, b, o) = r(s, a) + b(o), \qquad R_P(s, b, o) = r_p(s, o) - b(o),$$

with discount factor $\gamma \in (0, 1]$ (allowing $\gamma = 1$ for finite-horizon analyses). A principal policy ρ maps states to contracts, $\rho : \mathcal{S} \to \mathbb{R}^{|\mathcal{O}|}_{\geq 0}$, and an agent policy π maps (s, b) to actions, $\pi : \mathcal{S} \times \mathbb{R}^{|\mathcal{O}|}_{\geq 0} \to \mathcal{A}$. For any (ρ, π), the principal's value is

$$V_P^{\rho,\pi}(s) = \mathbb{E}\Big[\sum_{t=0}^{T-1} \gamma^t \big(r_p(s_t, o_t) - b_t(o_t)\big) \,\Big|\, s_0 = s \Big],$$

and analogously for the agent. The key difficulty is that the agent's action affects outcomes (hence both parties' payoffs) but is not contractible; the principal must steer actions using only payments contingent on observable outcomes.

Markovian SPE and the agent's continuation incentives. We focus on Markov strategies and subgame-perfect equilibrium (SPE). The agent's decision at state s under contract b trades off immediate intrinsic reward $r(s, a)$, expected transfer $\mathbb{E}_{o \sim O(s,a)}[b(o)]$, and the continuation value induced by future play. A convenient representation is the agent's (truncated) action-value function under principal policy ρ, written $Q_A^\star(s, a \mid \rho)$, which collects the intrinsic term plus the discounted continuation value from future contracts (but, by construction, excludes the *current* transfer so that current incentives enter linearly through b). Formally, one can write

$$Q_A^\star(s, a \mid \rho) = r(s, a) + \gamma\, \mathbb{E}_{o \sim O(s,a),\, s' \sim T(s,o)} \Big[\max_{a' \in \mathcal{A}} \Big(\mathbb{E}_{o' \sim O(s', a')} \rho(s')(o') + Q_A^\star(s', a' \mid \rho) \Big) \Big],$$

with terminal condition $Q_A^\star(\cdot, \cdot \mid \rho) = 0$ at $t = T$ (or, equivalently, indexing $Q_A^\star$ by time as in standard finite-horizon dynamic programming). Given b, the agent's best response at (s, b) is

$$\pi^\star(s, b) \in \arg\max_{a \in \mathcal{A}} \big\{ \mathbb{E}_{o \sim O(s,a)}[b(o)] + Q_A^\star(s, a \mid \rho) \big\}.$$

This expression highlights the core contractual lever: because $b \geq 0$ and depends only on o, the principal can alter incentives only through outcome likelihoods $O(s, a)$.

The principal's per-state contracting problem. Given an intended action a_p at state s, a classical benchmark is the *minimal implementation* contract: among all nonnegative payment schedules that make a_p optimal for the agent, choose one that minimizes the principal's expected transfer (equivalently, maximizes the principal's immediate payoff holding continuation fixed). In our notation, for a fixed continuation term $Q_A^\star(\cdot, \cdot \mid \rho)$ and true outcome model O, this problem takes the linear form

$$\begin{aligned} \min_{b \in \mathbb{R}_{\geq 0}^{|\mathcal{O}|}} \quad & \mathbb{E}_{o \sim O(s,a_p)}[b(o)] \\ \text{s.t.} \quad & \mathbb{E}_{o \sim O(s,a_p)}[b(o)] + Q_A^\star(s, a_p \mid \rho) \;\geq\; \mathbb{E}_{o \sim O(s,a)}[b(o)] + Q_A^\star(s, a \mid \rho), \qquad \forall a \neq a_p. \end{aligned}$$

The objective and constraints are linear in b. Limited liability enters only through $b(o) \geq 0$, and hidden action appears only through the incentive-compatibility (IC) inequalities.

This LP is a statewise object, but it is not "static": the constants $Q_A^\star(s, a \mid \rho)$ depend on the *entire* principal policy ρ through future contracts. In equilibrium, contracts must be consistent with the continuation incentives they generate.

SPE via backward induction: a meta-algorithm. In a finite-horizon Markov setting, an SPE can be constructed by backward induction on time. The conceptual meta-algorithm proceeds as follows. At the terminal date $t = T - 1$, continuation values are zero, and the principal's problem at each state reduces to a one-step contracting problem: for each candidate action a_p, solve the minimal-implementation LP to compute the cheapest contract b that induces a_p, then evaluate the principal's one-step value $\mathbb{E}_{o \sim O(s,a_p)}[r_p(s, o) - b(o)]$, and select the maximizing action/contract pair. Moving backward, suppose we have already computed the principal's continuation value $V_{P,t+1}(\cdot)$ (and, implicitly, the future contracts). At time t, for each state s and each candidate intended action a_p, the principal considers contracts that implement a_p given the agent's continuation terms and then selects the action/contract that maximizes

$$\mathbb{E}_{o \sim O(s,a_p),\, s' \sim T(s,o)} \big[r_p(s, o) - b(o) + \gamma V_{P,t+1}(s') \big].$$

The agent's best response at each state is computed from the corresponding IC condition using $Q_{A,t}^\star$, which itself is computed backward using the future contracts. In this sense, equilibrium computation resembles dynamic programming, but with an inner LP at each state to translate a desired action into an implementable contract.

Two observations are useful. First, because the minimal-implementation LP is linear and the feasible set is a polyhedron, optimal contracts are typically *extreme points*, concentrating payments on a small number of outcomes.

In particular, if one outcome is especially diagnostic of a_p relative to a deviation, paying only on that outcome can be optimal (a point we will exploit later when we design simple "nudges"). Second, because b enters the principal's objective negatively and enters the agent's IC constraints positively, the principal generically prefers contracts that make the agent *just indifferent* between a_p and the most tempting deviation. Minimal implementation is therefore often a knife-edge construction even when the underlying dynamic program is well-behaved.

Where LP-based implementation fails under misspecification. The preceding description is a benchmark: it presumes the principal knows the true outcome model O and the true agent continuation terms $Q_A^\star(\cdot, \cdot \mid \rho)$. In modern applications, neither is available. Contracts are computed from estimated transition/outcome models and from approximations of how the agent values future play. Even if these approximations are statistically consistent in large samples, in finite samples (or under distribution shift) the minimal-implementation LP can be brittle. The core issue is that minimal implementation typically selects a contract with *zero slack* in the binding IC constraints. Thus, small perturbations in either (i) outcome probabilities $O(s, a)$ or (ii) continuation differences $Q_A^\star(s, a_p \mid \rho) - Q_A^\star(s, a \mid \rho)$ can flip the agent's preference ordering.

To see the mechanism in its simplest form, consider a fixed s and two actions a_p and a. The IC constraint compares

$$\Delta(b) \equiv \mathbb{E}_{o \sim O(s,a_p)}[b(o)] - \mathbb{E}_{o \sim O(s,a)}[b(o)]$$

to the (negative) intrinsic/continuation advantage of deviating:

$$\Delta(b) \;\geq\; Q_A^\star(s, a \mid \rho) - Q_A^\star(s, a_p \mid \rho).$$

Under limited liability, the only way to increase $\Delta(b)$ is to shift payment weight toward outcomes that are more likely under a_p than under a. If the two outcome distributions are close, then $\Delta(b)$ is small unless b is large on a narrow set of outcomes. Minimal implementation pushes exactly to the boundary: it chooses b so that $\Delta(b)$ matches the required incentive gap with equality. A small error in $O(s, \cdot)$ that reduces $\Delta(b)$, or a small error in the estimated continuation gap that increases the right-hand side, can violate IC.

These failures are not merely local. In a dynamic system, an implementation error at state s_t changes the distribution over outcomes o_t and hence the next-state distribution over s_{t+1}. But the principal's future contracts were computed under the assumption that the intended action was taken and the nominal state trajectory would be followed. Thus, a one-step IC failure can push the system into regions of the state space where the learned policy is poorly calibrated, where outcome models are less accurate, or where

the principal has not even specified meaningful contracts. Put differently: minimal implementation is designed to be payment-efficient *conditional on correct prediction*, but it is not designed to be stable to the prediction errors that are inevitable in learned contracting.

A second, closely related failure mode concerns *rare-event leverage*. Because LP extreme points concentrate transfers, the computed contract may put a large payment on a low-probability outcome that is estimated to be slightly more likely under a_p than under deviations. This can be rational from a cost-minimization standpoint, but it creates an operational fragility: rare outcomes are precisely where empirical estimates are noisiest and where distribution shift is most likely. When the diagnostic event is misspecified (e.g., its probability advantage disappears in deployment), the contract loses its incentive bite while still exposing the principal to potentially high transfers when the event occurs.

Finally, the hidden-action dynamic setting introduces a subtle feedback: continuation incentives depend on the future contracts, and future contracts depend on the future states, which depend on today's action. If the principal computes contracts using an approximate continuation model, then even if the one-step outcome model $O(s, a)$ were correct, the induced IC inequalities can still be wrong because the agent is optimizing through time. This is the dynamic analogue of a familiar static lesson: when we design incentives using the wrong model of the agent, we can create "perverse" rewards that are locally sensible but globally misaligned with the agent's true objective.

These considerations motivate the central design goal of the paper: rather than relying on knife-edge minimal implementation computed from nominal estimates, we seek a lightweight modification of the per-state LP solution that creates an explicit buffer against (i) errors in predicted outcome probabilities and (ii) errors in predicted continuation incentives. The next section formalizes the modeling primitives that represent these errors and clarifies what can be estimated from data versus what must be treated as an explicit robustness budget.

3 Robust modeling primitives: what the principal can (and cannot) know

Our robustness modifications are driven by a simple empirical premise: in realistic deployments, the principal does not know the true outcome model $O(s, a)$, and—because the agent optimizes intertemporally—does not know the agent's relevant continuation incentives either. The purpose of this section is to make these informational gaps explicit, to separate what can be learned from what must be taken as a design budget, and to introduce three primitives that will parameterize our guarantees: (i) an uncertainty set $\mathcal{U}_{s,a}$ around the principal's nominal outcome model $\hat{O}(s, a)$; (ii) a bound

ε_s on misspecification of the agent's truncated continuation values; and (iii) a bound δ_s on the principal's own dynamic-programming approximation error.

Throughout, we view the principal's policy computation as occurring in a *training* phase, while performance is evaluated in a *validation/deployment* phase in which the agent best responds to posted contracts under the true environment. The analysis is "distributionally robust" in the sense that we evaluate the principal's policy uniformly over all true outcome models O consistent with the uncertainty sets.

3.1 Outcome-model uncertainty via state–action ambiguity sets $\mathcal{U}_{s,a}$

Fix a state–action pair (s, a). The principal has a nominal (estimated or simulated) outcome distribution $\hat{O}(s, a) \in \Delta(\mathcal{O})$, but the true distribution $O(s, a)$ may differ. We summarize this deviation by a set-valued restriction

$$O(s, a) \in \mathcal{U}_{s,a} \subseteq \Delta(\mathcal{O}),$$

where $\mathcal{U}_{s,a}$ is known to the principal.[1] We impose no special structure beyond the minimal requirements needed for tractability in the contract program: $\mathcal{U}_{s,a}$ is typically taken to be convex and closed, and to contain $\hat{O}(s, a)$. The key feature is that $\mathcal{U}_{s,a}$ is *action-specific*: different hidden actions may be more or less uncertain because they are observed at different frequencies in training or because they are generated by different operational regimes.

Two canonical constructions capture much of what we need.

Total-variation balls. A transparent choice is a total-variation (TV) neighborhood

$$\mathcal{U}_{s,a} = \Big\{ p \in \Delta(\mathcal{O}) : \mathrm{TV}\big(p, \hat{O}(s, a)\big) \le \eta_{s,a} \Big\},$$

where $\eta_{s,a} \ge 0$ is a radius. TV balls are appealing because they translate directly into worst-case bounds on expected payments, and because $\eta_{s,a}$ can be calibrated from finite-sample concentration for multinomial data. For example, if outcomes are observed under a *known* action a at state s for $n_{s,a}$ independent draws, then standard inequalities imply that, with probability at least $1 - \alpha$,

$$\mathrm{TV}\big(O(s, a), \hat{O}(s, a)\big) \;\lesssim\; \sqrt{\frac{\log(|\mathcal{O}|/\alpha)}{n_{s,a}}} \quad \text{(up to universal constants)}.$$

This calibration is deliberately schematic: the exact form depends on whether one uses Dvoretzky–Kiefer–Wolfowitz-type bounds, empirical Bernstein inequalities, or a bootstrap. What matters for our purposes is that $\eta_{s,a}$ can be made state–action dependent and can be tightened with additional data.

[1] In applications, $\mathcal{U}_{s,a}$ may itself be chosen conservatively from a family of candidate radii or confidence levels. We treat it as fixed to keep the contracting problem well-posed.

f-divergence balls. A second widely used family is an f-divergence neighborhood,

$$\mathcal{U}_{s,a} = \Big\{ p \in \Delta(\mathcal{O}) : D_f\big(p \,\|\, \hat{O}(s,a)\big) \le \eta_{s,a} \Big\},$$

including KL-divergence and χ^2-divergence as special cases. These sets can be motivated by likelihood-based confidence regions (e.g., Wilks' phenomenon) and often yield tractable dual reformulations when the principal optimizes expected payments subject to worst-case constraints.

What is (and is not) observable. A delicate point in hidden-action environments is that outcomes are not naturally labeled by the action that generated them. In many operational settings, however, the principal can create *action-revealing episodes* in training by temporarily using high-powered incentives (or rigid protocols) so that the agent's optimal response is effectively pinned down; under such "instrumented" regimes, outcomes can be attributed to intended actions with small residual error. Alternatively, the principal may rely on domain models or simulators that map proposed actions to outcome distributions. When neither is available, $\mathcal{U}_{s,a}$ should be interpreted as a *partial-identification* device: it collects all outcome distributions consistent with the data and the institutional assumptions the analyst is willing to defend. Our results do not require point identification; they require only that the true $O(s,a)$ lies in the specified set.

In short, $\hat{O}$ is a nominal object computed from data or simulation, while $\mathcal{U}$ is the principal's formal encoding of residual model uncertainty. Later, robust incentive constraints will take worst cases over $\mathcal{U}_{s,a}$, so enlarging $\mathcal{U}$ directly increases the payment buffer required for reliable implementation.

3.2 Bounded misspecification of the agent's truncated continuation values ε_s

Even if the principal knew $O(s,a)$ perfectly, robust contracting would still be necessary because the agent's best response depends on the entire future stream of contracts. In the benchmark formulation, the relevant object is the agent's truncated continuation value $Q_A^\star(s,a \mid \rho)$, which depends on the principal policy ρ through the continuation of play. In practice, the principal computes contracts using an approximation $\widehat{Q}(s,a)$ (e.g., from approximate dynamic programming, from a behavioral model fit, or from a structural estimate of $r(s,a)$ and future opportunities). We summarize the resulting discrepancy by a uniform statewise bound:

$$\sup_{a \in \mathcal{A}} \big| Q_A^\star(s,a \mid \rho) - \widehat{Q}(s,a) \big| \le \varepsilon_s.$$

We interpret ε_s as a *robustness budget* for continuation incentives at state s. Several features are worth emphasizing.

First, ε_s is intentionally state-dependent. Continuation values are typically harder to approximate in parts of the state space that are rarely visited in training, that are sensitive to tail events, or that depend on unobserved agent attributes; allowing ε_s to vary lets the designer recognize this heterogeneity rather than imposing a single global slack parameter.

Second, ε_s is a bound on an *action-value* approximation, not merely on the agent's one-step intrinsic payoff. This is important because dynamic incentives can dominate static ones: two actions that look similar in immediate outcomes may lead to different future states where the agent expects different future rents. Misspecifying those future rents can flip the agent's preference even when $\hat{O}$ is accurate.

Third, while we state the bound as uniform over a, what ultimately matters for incentive compatibility is the accuracy of *differences* $Q_A^\star(s, a_p \mid \rho) - Q_A^\star(s, a \mid \rho)$. A convenient (and conservative) implication of the uniform bound is that any such difference may be misspecified by at most $2\varepsilon_s$, which will directly determine the "margin" we must build into robust IC constraints in the next section.

How can ε_s be chosen? Unlike $\mathcal{U}_{s,a}$, ε_s is not solely a statistical confidence radius around an observable conditional distribution; it reflects misspecification of the agent model, approximation error in value-function fitting, and (in some applications) genuine behavioral departures from full rationality. Nonetheless, there are several disciplined ways to set it.

If $\widehat{Q}$ is computed from a parametric or nonparametric regression that predicts realized agent payoffs (or revealed choices) in a controlled training environment, then ε_s can be taken as a high-probability bound on prediction error, possibly inflated by a model-misspecification penalty. If $\widehat{Q}$ is produced by approximate dynamic programming with a known Bellman residual bound, then ε_s can be chosen to dominate the implied error propagation to truncated action values. If the principal has only coarse knowledge of the agent's future opportunities (e.g., bounds on outside options or on maximum attainable continuation rents), then ε_s is best interpreted as a conservative envelope that protects against those unknowns.

Our main point is conceptual: ε_s quantifies *how wrong the principal may be about the agent's dynamic incentives*, and robust contracting will trade off higher transfers for immunity to such errors.

3.3 Principal approximation error δ_s: robustness is not optimality

Robust implementation ensures that the agent takes the action the principal intends, but it does not guarantee that the intended action is itself close to equilibrium-optimal under the true environment. To track this distinction, we introduce a separate error term δ_s capturing the principal's approximation

error in computing its own recommended-action policy (or, more generally, in approximating the relevant contractual value function used in dynamic programming).

Formally, one can think of δ_s as bounding the difference between (i) the principal's true equilibrium continuation value from state s and (ii) the value implied by the approximate dynamic program used in training, evaluated on the recommended actions. Because different algorithmic pipelines lead to slightly different formalizations, we keep δ_s deliberately high level: it is a nonnegative quantity that enters the final welfare bound additively, discounted along the realized trajectory. Intuitively, δ_s is the cost of using an approximate planner, while ε_s is the cost of not knowing the agent's continuation incentives well enough to implement the planner's recommendations without slack.

Estimating δ_s. In many applied dynamic programming settings, δ_s can be related to a Bellman error or an off-policy evaluation error computed on a holdout sample. For example, if the principal uses fitted value iteration and can bound the sup-norm Bellman residual of its value estimate on the relevant state distribution, then δ_s can be chosen to dominate the implied performance loss. In model-based pipelines, δ_s may also include the effect of outcome-model error $\hat{O} \neq O$ on the principal's *planning* objective (as distinct from the effect on incentive constraints). Conceptually, δ_s is where we "charge" the principal for choosing a suboptimal policy due to limited data or function approximation; robustness will not remove this loss, but it will prevent additional loss from unintended agent deviations.

3.4 Putting the primitives together: a robust information structure

The three primitives $\left(\mathcal{U}_{s,a}, \varepsilon_s, \delta_s\right)$ jointly define the robust environment we will analyze. In deployment, we allow the true outcome model O to be any selection with $O(s,a) \in \mathcal{U}_{s,a}$ for all (s,a). The agent is treated as an oracle best responder given the true model and the posted contracts. The principal, by contrast, commits to a policy computed from nominal objects $(\hat{O}, \widehat{Q})$ and then protects itself by adding explicit slack to the per-state implementation constraints.

This separation clarifies a practical tradeoff. Enlarging $\mathcal{U}$ and ε makes the incentive constraints more conservative, typically requiring larger transfers to guarantee the intended action, which lowers the principal's value mechanically. Shrinking $\mathcal{U}$ and ε reduces transfers but increases the risk that the agent's best response changes under deployment conditions. The role of δ is orthogonal: it measures how good the intended policy is even if perfectly implemented.

With these robustness primitives in place, we can now formalize incentive compatibility in worst-case terms and show how a simple per-state "nudge"—a small additional payment concentrated on a diagnostic outcome—creates the strict buffer required to stabilize the agent's best response under both outcome-model ambiguity and continuation-value misspecification.

4 Robust incentive compatibility and per-state nudging

We now turn to the central design problem created by the primitives in Section 3: how can a principal *reliably* implement a desired hidden action when both the outcome model and the agent's continuation incentives are only known up to bounded error? The key step is to replace the usual (model-specific) incentive-compatibility (IC) inequalities with *worst-case* inequalities that must hold uniformly over the ambiguity sets and over continuation-value misspecification. This robustification has a direct economic interpretation: we are asking for a contract that keeps the agent on the intended action even in the least favorable environment consistent with what the principal believes it has learned.

4.1 Robust IC as worst-case utility inequalities

Fix a state s and suppose the principal wishes to implement (or recommends) an action $a_p \in \mathcal{A}$. Given a posted contract $b \in \mathbb{R}_{\geq 0}^{|\mathcal{O}|}$, the agent compares actions using (i) the expected payment under the relevant outcome distribution and (ii) the continuation value from choosing that action. Under the true environment, the agent's best response at s solves

$$a \in \arg\max_{a' \in \mathcal{A}} \Big\{ \mathbb{E}_{o \sim O(s,a')}\big[b(o)\big] + Q_A^\star(s, a' \mid \rho) \Big\}.$$

The principal does not observe a, and in deployment it cannot condition on $O(s,a)$ or on $Q_A^\star(s, a \mid \rho)$ directly. What it *can* do is build a uniform slack buffer into the inequality that makes a_p strictly preferred to every deviation $a \neq a_p$.

Operationally, the principal computes contracts using $\widehat{Q}(s,a)$, while the true continuation term may differ by up to ε_s in either direction. The most conservative implication is that the difference in continuation values between a_p and a deviation a may be misspecified by as much as $2\varepsilon_s$. Similarly, the mapping from actions to outcome distributions is only known through the ambiguity sets $\mathcal{U}_{s,a}$. We therefore define *robust IC for implementing a_p at state s* as the collection of inequalities

$$\inf_{p \in \mathcal{U}_{s,a_p}} \mathbb{E}_{o\sim p}\big[b(o)\big] + \widehat{Q}(s,a_p) - \varepsilon_s \;\geq\; \sup_{q \in \mathcal{U}_{s,a}} \mathbb{E}_{o\sim q}\big[b(o)\big] + \widehat{Q}(s,a) + \varepsilon_s, \qquad \forall a \neq a_p. \tag{1}$$

The left-hand side evaluates the agent's utility from the intended action under the *least favorable* outcome distribution in $\mathcal{U}_{s,a_p}$ and the *least favorable* continuation-value realization within the ε_s envelope. The right-hand side evaluates the utility of deviations under the *most favorable* outcome distribution in $\mathcal{U}_{s,a}$ and the *most favorable* continuation-value realization. If (1) holds, then a_p is a best response for an oracle agent for every true model $O \in \mathcal{U}$, despite the principal computing the contract using $\widehat{Q}$.

Two remarks clarify why we write robust IC in this particular form. First, the robustness is deliberately one-sided in a way that mirrors economic risk: we only care about deviations becoming attractive in deployment, so we take the worst case that reduces the appeal of a_p and increases the appeal of a. Second, the bound ε_s enters as a *margin requirement*. Rearranging (1) yields

$$\inf_{p \in \mathcal{U}_{s,a_p}} \mathbb{E}_p\big[b(o)\big] - \sup_{q \in \mathcal{U}_{s,a}} \mathbb{E}_q\big[b(o)\big] \;\geq\; \big(\widehat{Q}(s,a) - \widehat{Q}(s,a_p)\big) + 2\varepsilon_s. \tag{2}$$

Thus, any contract that would merely offset the principal's *estimated* continuation-value difference must be strengthened by an extra buffer of size $2\varepsilon_s$. This is the precise sense in which continuation misspecification is "paid for" by higher-powered incentives.

4.2 Robust separability: when limited liability can work

Robust IC is meaningful only if there exists some outcome-contingent contract that can discriminate between a_p and deviations. Under limited liability $b \geq 0$, the principal cannot punish the agent for "bad" outcomes; it can only create incentives by rewarding outcomes that are (robustly) more likely under the desired action. This motivates a statewise notion of outcome informativeness that is robust to ambiguity.

For $a \neq a_p$, define the *robust separability* between a_p and a at state s as

$$\bar{d}(s, a_p, a) \;=\; \max_{o \in \mathcal{O}} \Big\{ \inf_{p \in \mathcal{U}_{s,a_p}} p(o) - \sup_{q \in \mathcal{U}_{s,a}} q(o) \Big\}. \tag{3}$$

The expression inside braces is the worst-case advantage, at outcome o, of the intended action relative to the deviation. Maximizing over o selects the most diagnostic outcome in the worst case. When $\bar{d}(s, a_p, a) > 0$, there exists at least one outcome whose probability under a_p is uniformly larger than under a, even after allowing nature to pick adversarial distributions within each ambiguity set. This is precisely what limited liability needs: if such an outcome exists, then paying on it raises the expected payment from choosing a_p more than it raises the expected payment from choosing a.

The converse is also instructive. If $\bar{d}(s, a_p, a) \leq 0$, then for every outcome o the deviation can be made at least as likely as the intended action under some admissible models. In that case, no nonnegative contract can

guarantee that the expected payment advantage of a_p over a is strictly positive uniformly over $\mathcal{U}$; robust implementation may fail unless the principal (i) enlarges the observable outcome space $\mathcal{O}$ with more informative signals, (ii) tightens the ambiguity sets through additional data or instrumentation, or (iii) relaxes limited liability (e.g., by allowing deposits or penalties). Our results therefore make explicit a practical limitation: robustness is not "free," and in poorly identifiable environments it may be infeasible without redesigning measurement.

4.3 A per-state "nudge" that guarantees robust compliance

The separability parameter $\bar{d}$ leads to a simple constructive implementation rule that we will use repeatedly in the dynamic analysis: add a small extra payment concentrated on a single diagnostic outcome to create the margin required by (1). The underlying logic is linear: if we increase $b(o)$ by one unit for some o, then the worst-case expected payment under a_p increases by at least $\inf_{p\in\mathcal{U}_{s,a_p}} p(o)$, while the worst-case expected payment under a deviation a increases by at most $\sup_{q\in\mathcal{U}_{s,a}} q(o)$. The gap therefore increases by at least their difference, and $\bar{d}$ captures the best such difference obtainable from any outcome.

Formally, fix s and a_p. Let

$$o^\star(s,a_p) \in \arg\max_{o\in\mathcal{O}} \Big\{ \inf_{p\in\mathcal{U}_{s,a_p}} p(o) - \sup_{q\in\mathcal{U}_{s,a}} q(o) \Big\},$$

where the dependence on the competing a is suppressed for readability; in the multi-action case we will select an outcome that is simultaneously diagnostic against the "hardest" deviation. Define the worst-case separability against any deviation as

$$\bar{d}_{\min}(s,a_p) = \min_{a\neq a_p} \bar{d}(s,a_p,a).$$

When $\bar{d}_{\min}(s,a_p) > 0$, consider the *single-outcome nudge contract*

$$b_s(o) = \begin{cases} n_s, & o = o^\star(s,a_p), \\ 0, & \text{otherwise}, \end{cases} \qquad \text{with} \qquad n_s = \frac{M_s}{\bar{d}_{\min}(s,a_p)}. \tag{4}$$

Then for any deviation $a \neq a_p$,

$$\inf_{p\in\mathcal{U}_{s,a_p}} \mathbb{E}_p[b_s] - \sup_{q\in\mathcal{U}_{s,a}} \mathbb{E}_q[b_s] \;\geq\; n_s\, \bar{d}(s,a_p,a) \;\geq\; n_s\, \bar{d}_{\min}(s,a_p) \;=\; M_s.$$

Thus, choosing M_s to match the required slack in (2) guarantees robust IC. The most important special case for our dynamic setting is when the principal has already computed a *nominal* contract (or recommended action) using $\widehat{Q}$, and wishes only to immunize the recommendation against misspecification. In that case, we set the nudge margin to

$$M_s = 2\varepsilon_s,$$

so that the incremental contract adds precisely the buffer needed to prevent a best-response flip due to continuation-value error. Substituting into (4) yields the statewise *robust nudge magnitude*

$$n_s = \frac{2\varepsilon_s}{\bar{d}_{\min}(s, a_p)}. \tag{5}$$

This construction is deliberately transparent: it isolates a single diagnostic outcome and pays only when that outcome occurs. In applied terms, the nudge is a "bonus" tied to an outcome signal that is most indicative (in a worst-case sense) of compliance with the intended action. The required size of the bonus scales linearly in the robustness budget ε_s and inversely in the informativeness $\bar{d}_{\min}$. Put differently, continuation uncertainty makes us demand a larger strict preference gap, and weak outcome information forces us to pay more to create that gap under limited liability.

4.4 Payment cost and the role of "minimal implementation"

Robust IC only constrains *differences* in expected payments across actions; the principal would like to satisfy these constraints at minimal cost. A natural benchmark is the per-state *robust minimal-implementation* problem: minimize the worst-case expected payment under the implemented action subject to robust IC and limited liability,

$$\min_{b \geq 0} \sup_{p \in \mathcal{U}_{s,a_p}} \mathbb{E}_p[b(o)] \quad \text{s.t.} \quad \inf_{p \in \mathcal{U}_{s,a_p}} \mathbb{E}_p[b(o)] - \sup_{q \in \mathcal{U}_{s,a}} \mathbb{E}_q[b(o)] \geq M_s, \ \forall a \neq a_p.$$

The single-outcome nudge is not merely feasible; it yields an interpretable upper bound on payment cost. Indeed, under b_s in (4),

$$\sup_{p \in \mathcal{U}_{s,a_p}} \mathbb{E}_p[b_s(o)] = n_s \cdot \sup_{p \in \mathcal{U}_{s,a_p}} p\big(o^\star(s, a_p)\big) \leq n_s = \frac{M_s}{\bar{d}_{\min}(s, a_p)}.$$

Thus, to buy an additional robust IC margin of size M_s, the principal never needs to increase the worst-case expected payment by more than $M_s/\bar{d}_{\min}(s, a_p)$. This is the bound that will appear in the deployment welfare guarantees: robustness can be "priced" in closed form by the ratio of the continuation-value misspecification scale to the robust informativeness of outcomes.

Finally, it is worth noting what this per-state construction does *not* do. It does not claim that paying on a single outcome is always globally optimal once one accounts for the principal's intrinsic payoff $r_p(s, o)$, risk considerations, or regulatory constraints on contract variability. Rather, our point is that in the limited-liability, risk-neutral benchmark, robust IC has a particularly simple sufficient statistic: if $\bar{d}_{\min}$ is bounded away from zero, then a small additional payment mass concentrated on a diagnostic outcome

stabilizes incentives. When $\bar{d}_{\min}$ is close to zero, the same logic reveals a limitation: any robust contract must become expensive, reflecting the fundamental difficulty of incentivizing hidden actions with weakly informative outcome signals.

4.5 Closed-form contracts in tractable cases

The per-state nudge rule above is intentionally modular: all of the economics is pushed into the single statistic $\bar{d}(s, a_p, a)$, while the payment level is then a simple ratio $2\varepsilon_s/\bar{d}_{\min}(s, a_p)$. To make this operational, we now show how $\bar{d}$ and the associated "single-diagnostic-outcome" payments can be computed in closed form for common ambiguity sets. The goal is pragmatic as much as theoretical: in many applications the principal will want an explicit mapping from estimated outcome frequencies and confidence radii to a contract that is guaranteed to be incentive-compatible in deployment.

4.5.1 Total-variation balls: explicit separability and a one-outcome bonus

Suppose the ambiguity set at (s, a) is a total-variation (TV) ball around the nominal model $\hat{O}(s, a)$,

$$\mathcal{U}_{s,a} = \Big\{ p \in \Delta(\mathcal{O}) : \mathrm{TV}\big(p, \hat{O}(s, a)\big) \le \eta \Big\},$$

with radius $\eta \in [0, 1]$ interpreted as a distributional confidence parameter. TV balls are attractive because they provide a transparent worst-case calculus: nature can move probability mass by at most η in the direction that hurts the principal's IC constraints.

Fix an outcome $o \in \mathcal{O}$. Over a TV ball, the extremal value of $p(o)$ is obtained by shifting as much mass as possible away from (or into) that outcome. Concretely,

$$\inf_{p \in \mathcal{U}_{s,a}} p(o) = \max\{\hat{O}(s, a)(o) - \eta, 0\}, \qquad \sup_{p \in \mathcal{U}_{s,a}} p(o) = \min\{\hat{O}(s, a)(o) + \eta, 1\}. \tag{6}$$

Substituting (6) into the definition of robust separability (3) yields the closed form

$$\bar{d}(s, a_p, a) = \max_{o \in \mathcal{O}} \Big(\hat{O}(s, a_p)(o) - \hat{O}(s, a)(o) - 2\eta \Big)_+, \tag{7}$$

where $(x)_+ = \max\{x, 0\}$. Economically, (7) says that ambiguity acts like a uniform shrinkage of diagnosticity: even if the nominal model suggests that outcome o is k percentage points more likely under a_p than under a, worst-case model error can erase up to 2η of that gap.

Once $\bar{d}_{\min}(s, a_p) = \min_{a \neq a_p} \bar{d}(s, a_p, a)$ is computed, the robust nudge contract that immunizes the principal against continuation-value error is immediate: pick any outcome $o^\star(s, a_p)$ attaining the maximum in (7) against the

hardest deviation (or, equivalently, attaining $\bar{d}_{\min}$ after taking the minimum over deviations), and set

$$b_s(o) = \begin{cases} \dfrac{2\varepsilon_s}{\bar{d}_{\min}(s, a_p)}, & o = o^\star(s, a_p), \\ 0, & \text{otherwise.} \end{cases}$$

This "bonus on a diagnostic event" interpretation is often exactly how practitioners describe robust incentives: reward a verifiable signal that is hard to fake, and scale the reward by how informative that signal is relative to plausible model misspecification.

Two actions, two outcomes. The simplest case highlights why the one-outcome bonus is not merely a convenient sufficient condition, but can coincide with the optimal limited-liability implementation. Let $\mathcal{A} = \{a_p, a\}$ and $\mathcal{O} = \{o_1, o_2\}$. A contract is a pair $b = (b_1, b_2)$ with $b_i = b(o_i) \geq 0$. Under TV ambiguity, robust IC reduces to a single inequality comparing worst-case expected payments under the two actions. Using (6),

$$\inf_{p \in \mathcal{U}_{s,a_p}} \mathbb{E}_p[b] - \sup_{q \in \mathcal{U}_{s,a}} \mathbb{E}_q[b] = \min_{p \in \mathcal{U}_{s,a_p}} \{p(o_1)b_1 + p(o_2)b_2\} - \max_{q \in \mathcal{U}_{s,a}} \{q(o_1)b_1 + q(o_2)b_2\}.$$

Because there are only two outcomes, any increase in the probability of o_1 mechanically decreases the probability of o_2. As a result, the "best" direction to push incentives is to put all mass on whichever outcome has the larger robust probability advantage under a_p. More explicitly, define the robust probability gaps

$$\Delta_i = \inf_{p \in \mathcal{U}_{s,a_p}} p(o_i) - \sup_{q \in \mathcal{U}_{s,a}} q(o_i), \qquad i \in \{1, 2\}.$$

At most one of Δ_1, Δ_2 can be positive in a two-outcome model, and $\bar{d}(s, a_p, a) = \max\{\Delta_1, \Delta_2\}$. If $\bar{d} > 0$, then to achieve a margin requirement M_s it is optimal (in the robust minimal-implementation sense) to choose b supported on the outcome $i^\star \in \arg\max_i \Delta_i$, with $b_{i^\star} = M_s/\bar{d}$ and the other component equal to zero. Any attempt to "spread" payments across both outcomes weakens the worst-case payment gap per unit expected cost, because nature can exploit the ambiguity to concentrate probability on the more expensive outcome under the implemented action and on the more remunerative outcome under the deviation.

This two-by-two case is useful not because real environments have only two outcomes, but because it clarifies the economic role of $\bar{d}$: limited liability forces incentives to be built on *relative likelihood* of observable events, and robustness forces us to focus on the event whose likelihood advantage survives adversarial perturbations.

Finite $|\mathcal{O}|$ and multiple deviations. When $|\mathcal{O}| > 2$ and there are many deviations $a \neq a_p$, the one-outcome contract remains a sharp and interpretable implementation tool, but two practical subtleties arise.

First, the "most diagnostic" outcome can depend on which deviation we are guarding against. In principle, one could design a contract that pays on several outcomes to separate a_p from different deviations along different dimensions. Our approach instead selects a single outcome that is diagnostic against the *hardest* deviation, capturing the idea that robust IC is only as strong as the weakest link. Under TV balls, this is straightforward to compute: for each deviation $a \neq a_p$, compute the vector of nominal gaps $\hat{O}(s, a_p)(o) - \hat{O}(s, a)(o)$ across outcomes, subtract 2η, truncate at zero, and take the maximum over outcomes to obtain $\bar{d}(s, a_p, a)$. The robust bottleneck deviation is then the minimizer over $a \neq a_p$.

Second, while the single-outcome bonus yields an upper bound on the robust minimal-implementation cost, it need not be the unique optimizer in richer outcome spaces: paying on multiple outcomes can sometimes reduce the *worst-case* expected payment under a_p by better matching how ambiguity moves probability mass across outcomes. In such cases, the relevant computation is a small linear program (because TV constraints are linearizable), solved per state and intended action. From an applied perspective, this is still quite tractable: it is offline, separable across states, and typically low-dimensional because $|\mathcal{O}|$ is the number of reportable performance bins, audit flags, or verifiable events.

4.5.2 f-divergence balls: convex duality and when we need numerics

TV balls are deliberately conservative and yield clean formulas, but they are not always the best description of statistical uncertainty. A common alternative is an f-divergence ball,

$$\mathcal{U}_{s,a} = \Big\{ p \in \Delta(\mathcal{O}) : D_f\big(p \,\|\, \hat{O}(s,a)\big) \leq \eta \Big\},$$

where $D_f(p\|\hat{p}) = \sum_o \hat{p}(o)\, f\big(p(o)/\hat{p}(o)\big)$ for a convex function f with $f(1) = 0$. Examples include KL divergence ($f(u) = u \log u$) and χ^2 divergence ($f(u) = \frac{1}{2}(u-1)^2$). These ambiguity sets often arise naturally from likelihood-based confidence regions and yield less "corner" worst-case distributions than TV balls.

The key object we need, both for $\bar{d}$ and for robust IC, is the worst-case expectation of a linear functional, $p \mapsto \mathbb{E}_p[b] = \sum_o p(o)b(o)$, over an f-divergence ball. This is a convex optimization problem, and convex duality provides a convenient representation. In particular, using the Fenchel conjugate $f^\star(v) = \sup_{u \geq 0}\{uv - f(u)\}$, one obtains the robust expectation

bound

$$\sup_{p:D_f(p\|\hat{p})\le\eta} \mathbb{E}_p[b] = \inf_{\lambda\ge 0,\ \nu\in\mathbb{R}} \left\{ \lambda\eta + \nu + \lambda \sum_{o\in\mathcal{O}} \hat{p}(o)\, f^\star\Big(\frac{b(o)-\nu}{\lambda}\Big) \right\}, \qquad (8)$$

where $\hat{p} = \hat{O}(s,a)$. An analogous expression holds for the infimum (either by replacing b with $-b$ and negating, or by writing the corresponding dual directly). Representation (8) turns the inner worst-case problem into a low-dimensional convex minimization over (λ,ν), even when $|\mathcal{O}|$ is large.

Two implications are worth emphasizing.

Single-diagnostic-outcome structure often survives. If we restrict attention to one-outcome contracts $b(o) = n\mathbf{1}\{o = o^\star\}$, then (8) simplifies substantially because b takes only two values: n on $o^\star$ and 0 elsewhere. In that case, the worst-case expectation depends on $o^\star$ only through $\hat{p}(o^\star)$, and computing $\inf_{p\in\mathcal{U}_{s,a}} p(o^\star)$ or $\sup_{p\in\mathcal{U}_{s,a}} p(o^\star)$ reduces to a one-dimensional "binary" divergence problem. For KL balls, for example, the extremal distribution is an exponential tilt of $\hat{p}$; for χ^2 balls, the extremal distribution has a quadratic form and can be found by solving a scalar equation enforcing the divergence constraint. Practically, this means that even beyond TV, the principal can often implement robust nudges by (i) scanning outcomes to find a diagnostic $o^\star$ and (ii) choosing n_s via a small scalar computation.

General contracts require convex programs, but they remain separable and tractable. If the principal wishes to go beyond one-outcome bonuses—for instance, to satisfy additional constraints (budget caps, monotonicity in outcomes, or fairness restrictions), or to exploit richer outcome informativeness across deviations—then the robust minimal-implementation problem becomes a convex program with f-divergence constraints. Dual form (8) implies that each robust expectation term can be evaluated efficiently, and the overall per-state design can be handled with standard solvers. The limitation is interpretability rather than feasibility: the resulting contract may put positive weight on many outcomes, making it less transparent as a policy instrument.

We view this as an economically meaningful tradeoff. The "single diagnostic outcome" contract is attractive because it is easy to communicate and audit: it resembles a targeted bonus or a contingent subsidy keyed to an event that is ex ante agreed upon as evidence of compliance. More complex contracts can reduce worst-case payment costs, but at the cost of complexity, potential regulatory scrutiny, and the possibility that small modeling changes materially alter the contract. In settings where contracts must be stable and explainable, the closed-form nudge is a reasonable robustness primitive; where fine-tuned incentives are feasible, the convex program provides a systematic way to compute them.

In sum, tractable ambiguity sets let us translate robustness assumptions directly into explicit contracts. Under TV balls, separability and nudges are available in closed form via (7). Under f-divergence balls, convex duality provides a low-dimensional representation (8) that supports efficient computation, with closed forms or scalar root-finding in several important special cases. Where neither applies—for example, with nonstandard ambiguity sets, coupled constraints across states, or additional institutional restrictions—the per-state design remains a numerical robust optimization problem, but one that is typically small and solved offline, leaving the sequential analysis to focus on how these local nudges accumulate in welfare over time.

4.6 Welfare guarantees in sequential settings

Having reduced robust implementation at a state to a modular "nudge" rule, we now ask what this buys us in a genuinely sequential environment. The central point is that per-state robust incentive compatibility does more than prevent myopic deviations: it allows us to treat deployment as if the principal could directly choose the agent's action, up to two explicit and economically interpretable loss terms. One term reflects how well the principal learned (or approximated) the optimal *recommendation* policy; the other reflects the unavoidable *cost of robustness*—the extra payments needed to keep the agent on-path when both outcome models and agent continuation values may be misspecified.

Setup and benchmark. Fix a finite horizon T and discount factor $\gamma \in (0, 1]$. In training, the principal computes a recommended action $a_p(s)$ at each state and attaches a contract $b_s \in \mathbb{R}^{|\mathcal{O}|}_{\geq 0}$. In validation (deployment), an oracle agent best responds under the true model O, while the principal is committed to the trained policy. We benchmark performance against the principal's true subgame-perfect equilibrium (SPE) value $V_P^{\mathrm{SPE}}(s_0)$ under the true environment.

Two sources of misspecification matter. First, the principal may not know the agent's truncated continuation values: we assume a uniform bound

$$\sup_{a \in \mathcal{A}} \left| Q_A^\star(s, a \mid \rho) - \widehat{Q}(s, a) \right| \leq \varepsilon_s,$$

which, from an economic perspective, captures errors in forecasting the agent's outside opportunities, future subsidies, or downstream costs. Second, even if the principal knew the agent perfectly, it may still recommend the wrong action because it learned the environment imperfectly or used approximate dynamic programming; we summarize this by a per-state principal learning error δ_s.

Backward-induction logic: robust IC pins down the action path. The key technical step is a backward-induction argument that propagates *action implementation* forward through time. Suppose that at every state s that can be encountered under the principal's policy, the posted contract b_s satisfies robust incentive compatibility with a strict buffer,

$$\inf_{p\in\mathcal{U}_{s,a_p(s)}} \mathbb{E}_{o\sim p}\big[b_s(o)\big] - \sup_{q\in\mathcal{U}_{s,a}} \mathbb{E}_{o\sim q}\big[b_s(o)\big] \geq 2\varepsilon_s, \qquad \forall a \neq a_p(s). \tag{9}$$

Because the agent's continuation values may be misspecified by at most ε_s in either direction, the $2\varepsilon_s$ margin ensures that even the most adversarial realization of $(O, Q_A^\star)$ consistent with the bounds cannot flip the agent's best response away from $a_p(s)$. Put differently, (9) makes the intended action locally dominant among deviations *in the worst case*. Once this holds at every reached state, we can treat the deployed play path as if the principal had direct control over actions: the agent takes $a_t = a_p(s_t)$ for all t along the realized trajectory.

This observation is what makes a finite-horizon welfare bound essentially a dynamic-programming exercise. After robust IC removes strategic slippage, the only remaining differences between deployment and the SPE benchmark come from (i) choosing suboptimal recommended actions and (ii) paying extra subsidies to enforce robustness.

A finite-horizon guarantee and a clean regret decomposition. Let $V_P^{\text{deploy}}(s_0)$ denote the principal's realized expected value in validation under the true model $O \in \mathcal{U}$ when it posts the trained contracts and the agent best responds. Under the robust IC condition (9), the deployed action at s_t coincides with $a_p(s_t)$, hence the only welfare loss relative to the SPE can be charged to two state-by-state terms.

First, we charge *learning/approximation error*. Conceptually, δ_s measures how far the principal's training procedure is from choosing an SPE-optimal recommendation at state s (for instance, because its estimated contractual Q-function is off by at most δ_s and it chooses actions greedily with respect to that estimate). Standard approximate dynamic programming arguments yield a local comparison: the continuation value from taking $a_p(s)$ is within $2\delta_s$ of the continuation value from taking the truly optimal SPE action at s (the factor 2 is the familiar price of using an approximate argmax).

Second, we charge the *robustness cost* induced by limited liability and misspecification. Let

$$n_s := \max_{a\neq a_p(s)} \frac{2\varepsilon_s}{\bar{d}(s, a_p(s), a)},$$

where $\bar{d}$ is the robust separability statistic. By the constructive per-state design, there exists a contract that attains (9) while increasing the worst-case expected payment under the intended action by at most n_s above the

nominal minimal-implementation payment. Economically, n_s is the (worst-case) subsidy "overhead" required to stabilize incentives when the principal may be wrong about the agent and the outcome model may drift within $\mathcal{U}$.

Combining these two local comparisons with backward induction yields the following path-wise bound along the states induced by the deployed policy:

$$V_P^{\text{deploy}}(s_0) \geq V_P^{\text{SPE}}(s_0) - 2\sum_{t=0}^{T-1} \gamma^t \, \delta_{s_t} - \sum_{t=0}^{T-1} \gamma^t \, n_{s_t}. \tag{10}$$

Expression (10) is the sequential counterpart of a one-shot robust implementation guarantee. The inequality is informative precisely because it separates what is *statistical* from what is *incentive-theoretic*. If the principal can drive δ_s down through better learning (more data, richer features, better planning), the first loss term shrinks. If the principal can model the agent more accurately (smaller ε_s) or work in environments with more informative outcomes (larger $\bar{d}$), the second loss term shrinks.

Proof sketch (finite horizon). The proof is a dynamic-programming sandwich argument. Fix any time t and state s. Consider the principal's deployed one-step payoff plus continuation value when the agent takes $a_p(s)$. Under robust IC, this is exactly the deployed outcome, and the principal's expected continuation is well defined under the true kernel $O \circ T$. Compare this value to the principal's SPE continuation at (t, s). We decompose the gap into two pieces:

1. *Action selection error:* the difference between the SPE action and the recommended action $a_p(s)$, bounded by $2\delta_s$ by the definition of approximate optimality.

2. *Payment overhead:* the difference between the payment needed for nominal implementation and the payment needed for robust implementation, bounded by n_s by construction of the nudge.

Discounting and iterating from $t = T - 1$ backward to $t = 0$ yields (10). The induction relies on the fact that robust IC is imposed with respect to a conservative estimate of the agent's continuation values, so the agent's best response at t does not depend on the principal having correctly forecast the downstream contract sequence.

Extension sketch: infinite horizon with discounting and regularity. In many applications (compliance, maintenance, platform governance), the principal–agent interaction is effectively ongoing. Extending (10) to an infinite horizon requires two additional ingredients: (i) discounting, and (ii) a regularity condition ensuring that per-state errors do not accumulate pathologically.

If $\gamma < 1$ and we have uniform bounds $\delta_s \le \bar{\delta}$ and $n_s \le \bar{n}$ for all states that can be visited under the deployed policy, then the infinite-horizon analogue follows immediately from summing a geometric series:

$$V_P^{\text{deploy}}(s_0) \ge V_P^{\text{SPE}}(s_0) - \frac{2\bar{\delta} + \bar{n}}{1-\gamma}. \tag{11}$$

This bound is coarse but transparent: it says that robustness and learning errors translate into a steady-state "tax" on value, scaled by the effective planning horizon $(1-\gamma)^{-1}$.

When errors are state dependent, the relevant object is the (discounted) occupancy measure induced by the deployed policy. Writing $d_\gamma^\rho(s)$ for the expected discounted number of visits to state s under the deployed policy, one obtains the refined decomposition

$$V_P^{\text{deploy}}(s_0) \ge V_P^{\text{SPE}}(s_0) - 2\sum_{s\in\mathcal{S}} d_\gamma^\rho(s)\,\delta_s - \sum_{s\in\mathcal{S}} d_\gamma^\rho(s)\,n_s,$$

provided robust IC holds wherever $d_\gamma^\rho(s) > 0$. Establishing that d_γ^ρ is well behaved typically requires a mild stability condition (e.g., geometric ergodicity under a stationary policy, or simply bounded visitation under discounting). Economically, this is a "no-explosions" assumption: the policy should not concentrate all its mass on rare states where the model is unlearnable and incentives are prohibitively expensive.

Interpretation and limitations. Bound (10) highlights a tradeoff we view as central for sequential contracting under partial observability. Robustness is not free: the nudge term n_s is the premium the principal pays to insure against two kinds of uncertainty at once—ambiguity about how actions map into observable outcomes and ambiguity about how the agent evaluates future continuation. Yet, precisely because this premium is state-by-state and separable, it can be engineered and audited locally, while the sequential welfare impact is obtained by discounting and summing.

At the same time, the guarantee is only as strong as the modeling primitives permit. If $\bar{d}(s, a_p, a)$ is near zero at states that matter, robust IC becomes expensive and the bound deteriorates, reflecting a real economic obstruction: outcomes simply do not reveal enough about actions to incentivize cheaply under limited liability. Moreover, our argument presumes that the principal can commit to the contract sequence generated by its policy; without commitment, the relevant equilibrium notion changes and the backward-induction implementation step must be revisited.

The next section formalizes the sense in which the robustness cost captured by $\varepsilon_s/\bar{d}(s, a_p, a)$ is not merely an artifact of our construction, but an inherent feature of hidden-action sequential environments with limited-liability contracts.

4.7 Minimax lower bounds and necessity of separability

Our per-state nudge construction delivers a simple *upper* bound on the subsidy overhead needed to stabilize incentives under outcome ambiguity and continuation-value misspecification. A natural question is whether the scaling in that bound is merely an artifact of the construction. In this section we show it is not: up to constant factors, the dependence on the ratio $\varepsilon_s/\bar{d}(s, a_p, a)$ is *minimax necessary* under limited liability. Economically, $\bar{d}$ is not just a convenient statistic for our design; it is an inherent measure of how much leverage observable outcomes provide for disciplining hidden actions once we allow for adversarial shifts within $\mathcal{U}$.

A one-state reduction. We isolate the economic obstruction in the simplest environment: a single state s (or, equivalently, a fixed period of a dynamic problem holding the continuation policy fixed), two actions a_p (the intended action) and a (a deviation), and a contract $b \in \mathbb{R}_{\geq 0}^{|\mathcal{O}|}$ paid after observing $o \in \mathcal{O}$. The agent compares actions by expected payments plus a continuation-value term. The principal does not know $Q_A^\star(s, \cdot \mid \rho)$ exactly and uses $\widehat{Q}(s, \cdot)$, with misspecification bounded by ε_s in either direction. As in the robust IC condition, to ensure that no combination of outcome ambiguity $O \in \mathcal{U}$ and continuation-value misspecification can flip the best response away from a_p, the contract must create a payment advantage that clears a strict buffer. Abstracting from known differences in $\widehat{Q}$ (which can be folded into the required margin), we can write the required robust payment advantage as a margin constraint

$$\inf_{p \in \mathcal{U}_{s,a_p}} \mathbb{E}_{o\sim p}[b(o)] - \sup_{q \in \mathcal{U}_{s,a}} \mathbb{E}_{o\sim q}[b(o)] \geq M, \qquad \text{with } M \asymp 2\varepsilon_s. \tag{12}$$

We emphasize that the left-hand side is evaluated in the *least favorable* model for implementing a_p: nature minimizes the expected payment under a_p and maximizes it under a.

The principal, however, bears the subsidy cost under the *true* $O \in \mathcal{U}$. In a robust (minimax) evaluation of payments, the relevant exposure is the worst-case expected payment under the intended action,

$$\sup_{p \in \mathcal{U}_{s,a_p}} \mathbb{E}_{o\sim p}[b(o)].$$

This is the sense in which robustness can be expensive: the same uncertainty set that forces us to be conservative about incentives can also contain models in which the subsidized outcome is frequent.

The separability parameter as a bound on implementable margins. Recall the robust separability statistic

$$\bar{d}(s, a_p, a) = \max_{o \in \mathcal{O}} \Big\{ \inf_{p \in \mathcal{U}_{s,a_p}} p(o) - \sup_{q \in \mathcal{U}_{s,a}} q(o) \Big\}.$$

It captures the best *worst-case* gap in the probability of any single outcome under a_p relative to a. When $\bar{d}$ is small, even the most diagnostic outcome can be made nearly equally likely under the deviation once we allow for ambiguity; when $\bar{d} = 0$, there is no outcome that remains robustly more likely under a_p than under a, so limited-liability incentives lose their bite.

The next proposition formalizes the converse: if $\bar{d}$ is small, then *any* robustly incentive-compatible contract must entail a large worst-case subsidy exposure. The statement is minimax: we exhibit an ambiguity structure consistent with a given $\bar{d}$ that forces the lower bound.

Proposition 4.1 (Minimax lower bound: $\varepsilon/\bar{d}$ is unavoidable)**.** *Fix a state s and two actions a_p and a. Let $M > 0$ be the required robust margin in* (12)*. For any $d \in (0, 1]$, there exist an outcome set $\mathcal{O}$ and uncertainty sets $\mathcal{U}_{s,a_p}, \mathcal{U}_{s,a}$ such that $\bar{d}(s, a_p, a) = d$ and the following holds: every limited-liability contract $b \geq 0$ satisfying* (12) *must have*

$$\sup_{p \in \mathcal{U}_{s,a_p}} \mathbb{E}_{o \sim p}[b(o)] \;\geq\; \frac{M}{d}. \tag{13}$$

In particular, taking $M = 2\varepsilon_s$ yields the scaling $\Omega(\varepsilon_s/\bar{d}(s, a_p, a))$.

Proof sketch and economic interpretation. The construction is deliberately stark and highlights the role of limited liability. Take $\mathcal{O} = \{o^\star, o^0\}$ with two candidate models for the intended action and one for the deviation:

$$\mathcal{U}_{s,a_p} = \{p^\ell, p^h\}, \qquad \mathcal{U}_{s,a} = \{q\},$$

where

$$p^\ell(o^\star) = d, \quad p^h(o^\star) = 1, \quad q(o^\star) = 0, \qquad \text{and probabilities on } o^0 \text{ fill the remainder.}$$

Then

$$\inf_{p \in \mathcal{U}_{s,a_p}} p(o^\star) = d, \qquad \sup_{q \in \mathcal{U}_{s,a}} q(o^\star) = 0,$$

and for the other outcome o^0 the corresponding difference is nonpositive, so indeed $\bar{d}(s, a_p, a) = d$.

Now consider any $b \geq 0$. Under this uncertainty structure, the worst case for incentive provision is p^ℓ under a_p (which makes the "diagnostic" outcome rare) and q under a (which makes it impossible). Thus the robust margin becomes

$$\begin{aligned}
\inf_{p \in \mathcal{U}_{s,a_p}} \mathbb{E}_p[b] - \sup_{q \in \mathcal{U}_{s,a}} \mathbb{E}_q[b] &= \mathbb{E}_{p^\ell}[b] - \mathbb{E}_q[b] \\
&= \big(d\, b(o^\star) + (1-d)\, b(o^0)\big) - b(o^0) \\
&= d\big(b(o^\star) - b(o^0)\big) \;\leq\; d\, b(o^\star),
\end{aligned}$$

where the inequality uses $b(o^0) \geq 0$. To satisfy (12), we therefore must have $b(o^\star) \geq M/d$. But $p^h \in \mathcal{U}_{s,a_p}$ puts probability one on $o^\star$, implying the worst-case expected payment under the intended action is at least

$$\sup_{p \in \mathcal{U}_{s,a_p}} \mathbb{E}_p[b] \geq \mathbb{E}_{p^h}[b] = b(o^\star) \geq \frac{M}{d},$$

which is (13).

The economics are immediate. Limited liability prevents the principal from using negative transfers to "undo" the payment in benign models. Once the contract must be large enough to create a margin when $o^\star$ is *rare* under a_p (the p^ℓ case), the principal is exposed to paying that large amount in any model within $\mathcal{U}_{s,a_p}$ where $o^\star$ is *common* (the p^h case). Robustness thus couples two forces: (i) the need to amplify a small diagnostic probability gap (captured by d) to overcome continuation-value misspecification (captured by $M \asymp \varepsilon_s$), and (ii) the inability to hedge the resulting large transfer across models because payments must remain nonnegative.

Why $\bar{d} > 0$ is not just sufficient but essentially necessary. Proposition 4.1 also clarifies why the strict separability condition $\bar{d}(s, a_p, a) > 0$ is qualitatively indispensable. When $\bar{d} = 0$, the bound becomes vacuous only because the correct statement is stronger: for some ambiguity structures, *no finite contract* can satisfy a strictly positive margin $M > 0$ under limited liability. Intuitively, if every outcome that could trigger a subsidy under a_p can be made at least as likely under the deviation once we take worst cases in $\mathcal{U}$, then no nonnegative payment vector can create a robust advantage for a_p.

This conclusion aligns with a familiar identification logic: under hidden actions, contracting power comes from statistical distinguishability of action-induced outcome distributions. Our contribution is to show that with ambiguity and agent misspecification, the relevant notion of distinguishability is not a likelihood ratio under a known model but the *worst-case* probability gap summarized by $\bar{d}$. In this sense, $\bar{d}$ is a "price of unobservability plus shift": unobservability because actions are not directly contractible, and shift because ambiguity allows outcome distributions to drift in the direction that erodes incentives.

Relation to common uncertainty sets. The lower bound above is minimax and therefore uses an extreme ambiguity structure to make the point sharply. For more regular sets, such as total-variation or f-divergence balls, the same comparative statics persist. As the radius of the uncertainty set grows, the infimum probability under a_p can fall and the supremum probability under a can rise for the same outcome, shrinking $\bar{d}$; simultaneously, the set typically also contains models that make high-payment outcomes

more frequent, increasing worst-case expected subsidies. Thus even when (13) does not hold with constant 1 uniformly over all such sets, the qualitative implication survives: robustness requires paying in inverse proportion to the residual diagnostic power that remains after allowing for adversarial probability shifts.

Implications for sequential contracting. Although we have presented the argument in a one-state reduction, its force is sequential. At any state where the principal needs to stabilize incentives against continuation-value errors of order ε_s, the local problem contains the same tension: if $\bar{d}(s, a_p, a)$ is small for some plausible deviation a, then the principal must either (i) accept large worst-case subsidy exposure in that state, or (ii) change the recommendation to an action whose deviations are more separable, or (iii) invest in measurement/monitoring that enlarges the outcome space and increases separability. In practice, this reframes "better incentives" as often being "better observables": richer signals (audits, sensors, verification, or informative intermediate outcomes) increase $\bar{d}$ and therefore reduce the minimax price of robustness.

Limitations and what can break the lower bound. Finally, it is important to be clear about what drives the necessity result. The lower bound hinges on (a) limited liability ($b \geq 0$), which prevents ex post clawbacks, and (b) robustness with respect to a nontrivial uncertainty set, which can couple "hard-to-incentivize" models (small gaps) with "expensive" models (high incidence of subsidized outcomes). If either ingredient is relaxed, the scaling can change. For example, allowing negative transfers (or allowing the principal to escrow funds and impose penalties) can hedge subsidy exposure across models; allowing additional instruments such as costly state verification can effectively create new outcomes and increase $\bar{d}$; and restricting $\mathcal{U}$ to rule out large swings in the frequency of the paid outcome can weaken the worst-case payment implication. We view these as economically meaningful design levers rather than technical loopholes: they correspond to changing the informational and legal constraints of the contracting environment.

The central takeaway is that the ratio $\varepsilon_s/\bar{d}(s, a_p, a)$ is not an artifact of a particular contract construction. It is a structural measure of how expensive it is to robustly implement hidden actions when (i) outcome data are only partially informative about actions and (ii) the principal cannot perfectly forecast how the agent values the future. The next section builds on this characterization to discuss computational methods: once we accept that robustness costs are pinned down by such local statistics, we can design algorithms that compute robust contracts efficiently and diagnose when (and where) incentives will be prohibitively expensive.

4.8 Algorithms: per-state robust programs, plug-in uncertainty, and scalable heuristics

Our theory is deliberately "local": at each visited state s, we stabilize a desired recommendation $a_p(s)$ by adding a robust nudge that clears a margin of order ε_s against outcome ambiguity. This locality is not only conceptually clean—it is computationally useful. In finite-horizon problems (and, more generally, in episodic training), contract computation can be decomposed into a collection of per-state subproblems that can be solved independently once we have (i) a recommended action $a_p(s)$ from the principal's learning procedure and (ii) state-dependent uncertainty and error summaries $(\mathcal{U}_{s,a})_{a\in\mathcal{A}}$ and ε_s.

A per-state robust optimization template. Fix a state s and an intended action a_p. The principal chooses a limited-liability contract $b \in \mathbb{R}_{\geq 0}^{|\mathcal{O}|}$. To guarantee that a_p remains a best response for any $O \in \mathcal{U}$ and any continuation-value misspecification consistent with ε_s, it suffices to enforce, for each $a \neq a_p$,

$$\inf_{p\in\mathcal{U}_{s,a_p}} \mathbb{E}_p[b] - \sup_{q\in\mathcal{U}_{s,a}} \mathbb{E}_q[b] \;\geq\; \Delta_s(a), \qquad \Delta_s(a) \equiv \widehat{Q}(s,a) - \widehat{Q}(s,a_p) + 2\varepsilon_s. \tag{14}$$

When $\Delta_s(a) \leq 0$, the corresponding constraint is slack and can be dropped; economically, the principal already believes (with buffer) that the agent prefers a_p absent additional subsidies.

Given (14), we can formalize a robust minimal-implementation objective as minimizing worst-case subsidy exposure under the implemented action,

$$\min_{b\geq 0} \sup_{p\in\mathcal{U}_{s,a_p}} \mathbb{E}_p[b] \quad \text{s.t.} \quad (14) \;\; \forall a \neq a_p. \tag{15}$$

Problem (15) is a distributionally robust linear program once $\mathcal{U}_{s,a}$ is chosen from standard convex families (total-variation balls, f-divergence balls, moment sets, or polytopes defined by confidence intervals). In small outcome spaces, solving (15) directly provides a useful benchmark: it returns the cheapest robust contract given the chosen uncertainty model. In large outcome spaces, (15) also clarifies what must be approximated: the computational burden is concentrated in evaluating the robust expectation operators $\sup_{p\in\mathcal{U}} \mathbb{E}_p[b]$ and $\inf_{p\in\mathcal{U}} \mathbb{E}_p[b]$.

Computing robust expectations under common $\mathcal{U}$. Two cases recur in practice.

(i) Total-variation balls. Let $\mathcal{U} = \{p \in \Delta(\mathcal{O}) : \mathrm{TV}(p,\hat{p}) \leq \eta\}$ where $\hat{p} = \hat{O}(s,a)$. For any fixed b, computing $\sup_{p\in\mathcal{U}} \mathbb{E}_p[b]$ is equivalent to moving at most η mass (in TV units) from low-payment outcomes to high-payment

outcomes. Operationally, one can compute these quantities by sorting outcomes by $b(o)$ and greedily reallocating probability mass subject to the ℓ_1 budget $\|p-\hat{p}\|_1 \le 2\eta$. This yields an $O(|\mathcal{O}|\log|\mathcal{O}|)$ routine per evaluation of sup (and similarly for inf by moving mass in the opposite direction). When b is supported on a small set of outcomes (in particular, a single diagnostic outcome), the computation collapses to constant time.

(ii) f-divergence balls. Let $\mathcal{U} = \{p : D_f(p\|\hat{p}) \le \eta\}$ for a convex f with $f(1) = 0$. Then $\sup_{p\in\mathcal{U}} \mathbb{E}_p[b]$ admits a standard convex dual representation: it can be computed as a one-dimensional convex minimization over a multiplier $\lambda \ge 0$ involving the convex conjugate f^*. For example, for KL divergence D_{KL}, one obtains an entropic risk form,

$$\sup_{p:D_{\mathrm{KL}}(p\|\hat{p})\le\eta} \mathbb{E}_p[b] = \inf_{\lambda>0} \left\{ \lambda\eta + \lambda \log\left(\mathbb{E}_{o\sim\hat{p}}\left[e^{b(o)/\lambda}\right]\right)\right\},$$

and $\inf_{p\in\mathcal{U}} \mathbb{E}_p[b]$ is computed by applying the same formula to $-b$ (with a sign flip). The practical advantage is that robust expectations can be evaluated accurately with a handful of Newton or bisection steps in λ, even when $|\mathcal{O}|$ is large, provided we can compute (or estimate) $\mathbb{E}_{\hat{p}}[\exp(b/\lambda)]$.

These evaluation routines can be embedded in an outer solver for (15). In the TV/polyhedral case, the overall program remains an LP (after standard epigraph transformations). In the f-divergence case, the program is convex but not linear; nonetheless, its dimension is still $|\mathcal{O}|$ and typically small in applications where outcomes are coarse categories.

Closed-form robust nudges as a computational shortcut. While (15) is a natural baseline, our constructive results imply a far simpler implementation that is often near-optimal and, crucially, scales to very large $\mathcal{O}$. Namely, compute a diagnostic outcome

$$o^\star(s,a_p) \in \arg\max_{o\in\mathcal{O}} \left\{ \inf_{p\in\mathcal{U}_{s,a_p}} p(o) - \sup_{q\in\mathcal{U}_{s,a}} q(o) \right\},$$

and then place all payment mass on $o^\star$. Concretely, let

$$\bar{d}_{\min}(s,a_p) \equiv \min_{a\neq a_p} \bar{d}(s,a_p,a), \qquad B_s \equiv \max_{a\neq a_p:\Delta_s(a)>0} \frac{\Delta_s(a)}{\bar{d}(s,a_p,a)},$$

and set $b_s(o^\star) = B_s$ and $b_s(o) = 0$ for $o \neq o^\star$. This "single-diagnostic-outcome" heuristic replaces solving (15) by computing $\bar{d}(s,a_p,a)$ for each deviation and taking a maximum. Under TV balls, $\bar{d}$ has a closed form in terms of $\hat{O}$ and η, so the computation is $O(|\mathcal{A}||\mathcal{O}|)$ per state and requires no numerical optimization.

We emphasize the economic interpretation: the contract is not "complex" but rather *targeted*. The algorithm is therefore well aligned with practice: it

suggests that robust incentives can often be implemented by paying on one verifiable event (an audit trigger, a diagnostic sensor pattern, a milestone) rather than by attempting to price every possible outcome.

Plug-in construction of uncertainty sets. The per-state computations above require $\hat{O}(s,a)$ and $\mathcal{U}_{s,a}$. In many applications, we estimate $\hat{O}(s,a)$ from counts in a generative model or from logged trajectories. A simple plug-in approach is:

$$\hat{O}(s,a)(o) = \frac{N(s,a,o)}{N(s,a)}, \qquad N(s,a) = \sum_o N(s,a,o),$$

with an uncertainty radius $\eta_{s,a}$ chosen from finite-sample concentration. For TV balls, a conservative but transparent choice is

$$\eta_{s,a} \;=\; \sqrt{\frac{\log(2|\mathcal{O}|/\alpha)}{2N(s,a)}},$$

which ensures $\mathrm{TV}(O(s,a), \hat{O}(s,a)) \le \eta_{s,a}$ with high probability (up to constants) under i.i.d. sampling. In settings with heterogeneous outcome probabilities, empirical Bernstein radii can materially tighten $\eta_{s,a}$, reducing subsidy costs by shrinking $\mathcal{U}$ precisely where the model is well measured.

For f-divergence balls, one may instead calibrate $\eta_{s,a}$ via likelihood-ratio confidence regions; for KL, this corresponds to a classical multinomial confidence set. The practical message is the same across constructions: *the radius is a policy lever.* Larger radii improve robustness but raise the required nudge; smaller radii economize on payments but risk incentive failure under shift.

Estimating and operationalizing ε_s and continuation-value uncertainty. The robustness margin in (14) scales with ε_s, the principal's bound on continuation-value misspecification for the agent. In applications, we rarely observe $Q_A^\star$ directly, so ε_s must itself be estimated or upper bounded. Two pragmatic approaches are common.

First, if $\widehat{Q}$ is obtained from a supervised or fitted value procedure (e.g., regression on Monte Carlo rollouts), we can form confidence bands using standard generalization bounds or, more practically, bootstrap/ensemble dispersion. This yields a state-dependent ε_s that is larger in regions with limited data support.

Second, when we cannot credibly quantify ε_s pointwise, we can work with a small number of bins or state abstractions $\mathcal{C}(s)$ and use a conservative groupwise bound $\varepsilon_{\mathcal{C}}$. Algorithmically, this amounts to computing nudges using $\varepsilon_{\mathcal{C}(s)}$; economically, it is a commitment to worst-case robustness within an operationally meaningful class (e.g., "early-stage projects" versus "late-stage projects").

Large outcome spaces: learning diagnostic events. When $|\mathcal{O}|$ is very large or outcomes are high-dimensional (images, text, rich sensor traces), contracts that specify a payment for each outcome are infeasible. Our framework suggests an alternative: *learn a diagnostic event* $g(o) \in \{0,1\}$ and contract only on $g(o) = 1$. Doing so induces a binary outcome space with probabilities

$$P_{s,a}(g=1) = \sum_o O(s,a)(o)\, g(o), \qquad \hat{P}_{s,a}(g=1) = \sum_o \hat{O}(s,a)(o)\, g(o),$$

and therefore a one-dimensional separability statistic

$$\bar{d}_g(s, a_p, a) = \inf_{p \in \mathcal{U}_{s,a_p}} P_p(g=1) - \sup_{q \in \mathcal{U}_{s,a}} P_q(g=1).$$

Computationally, choosing g becomes a classification problem: we seek a function of the observed outcome that best distinguishes data generated under a_p from data generated under a, subject to robustness penalties. In practice, we can train g_θ (e.g., logistic regression on handcrafted features, or a neural classifier) to maximize an empirical proxy for $\min_{a \neq a_p} \bar{d}_{g_\theta}(s, a_p, a)$ while controlling overfitting through held-out validation. Once g is fixed, we revert to the simple single-event contract: pay B_s when $g(o) = 1$, where B_s is chosen using the estimated (and robustified) gap $\bar{d}_g$.

This heuristic has an appealing institutional interpretation: rather than contracting on raw high-dimensional data, the principal designs an *auditable signal* (a rule for when the data counts as "success") and then uses a simple transfer on that signal.

Integration with deep RL and policy optimization. In large state spaces, the principal typically learns $a_p(s)$ via approximate dynamic programming or deep RL. The robust nudge can be integrated into this pipeline in two complementary ways.

First, in a *plug-in* mode, we learn a nominal recommended-action policy using standard RL on the principal's intrinsic reward, and only at deployment compute the per-state nudge using current $(\hat{O}, \mathcal{U}, \widehat{Q}, \varepsilon)$. This isolates robustness in the contract layer and is attractive when we cannot easily differentiate through the nudge computation.

Second, in a *robustness-aware* mode, we incorporate an estimate of the subsidy overhead into the principal's training objective. A simple surrogate is to penalize actions with low separability by subtracting an estimated cost term $\max_{a \neq a_p} 2\varepsilon_s / \bar{d}(s, a_p, a)$ (or its smoothed version) from the principal's per-step reward. This encourages the learned policy to avoid regions of the state space where incentives are intrinsically expensive, thereby converting a post hoc robustness fix into an endogenous feature of the recommended action rule.

Implementation heuristics and diagnostics. Two additional heuristics matter in practice. First, we often impose a hard cap $b(o) \leq \bar{B}$ or a per-episode budget; when the computed B_s exceeds this cap, a reasonable fallback is to change the recommendation to an action with higher $\bar{d}$ (if available), or to trigger a monitoring intervention that effectively enriches $\mathcal{O}$ and increases separability. Second, we advocate reporting $\bar{d}_{\min}(s, a_p)$ as an online diagnostic. From an operational perspective, $\bar{d}_{\min}$ is a leading indicator of when contracting is likely to be fragile: small $\bar{d}_{\min}$ flags states where either more data (to shrink $\mathcal{U}$), a different recommendation, or richer observables are required.

Taken together, these algorithmic components translate the theoretical objects in our bounds into a practical workflow: estimate $(\hat{O}, \mathcal{U})$ from data, learn a recommendation policy, compute (or approximate) per-state robust nudges using separability statistics, and use separability as a diagnostic to understand when robustness will be inexpensive versus prohibitive. The next section evaluates this workflow empirically in controlled environments where we can adversarially shift outcomes and directly observe the resulting budget–welfare tradeoffs.

4.9 Experiments: adversarial shift, non-stationary agents, and budget–welfare tradeoffs

We use controlled experiments to stress-test the algorithmic workflow from Section 4.8 under exactly the failure modes that motivate our robust IC constraints: (i) *distribution shift* in outcome generation relative to the principal's nominal model $\hat{O}$, and (ii) *misspecification and drift* in the agent's continuation values relative to $\widehat{Q}$. The experiments are not intended as realistic calibration; rather, they provide an "engineering sanity check" of the comparative statics and decompositions implied by our bounds. In particular, we ask whether the per-state robust nudge behaves as predicted: it should (a) prevent incentive reversals under adversarially chosen $O \in \mathcal{U}$, (b) incur additional payments that scale as $\varepsilon_s/\bar{d}$, and (c) expose states with small separability as the operational bottleneck.

Protocol: two-phase training/validation with adversarial shift. Across all environments we adopt a two-phase protocol aligned with our theoretical setup. In a *training phase*, the principal learns (or is given) a recommended-action policy $a_p(s)$ and constructs nominal estimates $(\hat{O}, \widehat{Q})$ from rollouts under a reference outcome model. In a *validation phase*, we freeze the principal's policy and contracts, but allow Nature to choose a true outcome kernel O within the declared uncertainty sets $\mathcal{U}_{s,a}$. The agent best responds as in the equilibrium condition (Agent BR), using the true O and its true continuation values. This validation design is deliberately asymmetric: the principal is "committed" to its estimates and uncertainty model, while the

environment and the agent are treated as adversarial (within the declared robustness envelope). Economically, this corresponds to deployment in a new regime (market conditions, platform composition, measurement drift) where the principal cannot re-fit models on the fly, and must rely on ex ante robustness.

Within validation, we consider two shift models. First, an *oblivious adversary* chooses a fixed $O \in \mathcal{U}$ before the episode begins. Second, an *adaptive adversary* selects, at each visited (s, a), an extremal distribution in $\mathcal{U}_{s,a}$ to minimize the principal's realized payoff subject to respecting the uncertainty description.[2] In both cases, we report (i) principal value, (ii) realized payments, and (iii) a direct incentive diagnostic: the fraction of visited states in which the agent deviates from the principal's intended recommendation.

Contracts compared: nominal vs robust nudging. At each state, we compare three contract layers built on the same underlying recommendation $a_p(s)$. (1) *Nominal minimal implementation*: solve the non-robust counterpart of (15) with $\hat{O}$ and $\varepsilon_s = 0$, i.e., implement a_p under the estimated model. (2) *Robust nudge*: enforce robust IC with margin $2\varepsilon_s$ using the per-state construction from Section 4.8 (either by directly solving (15) or via the single-diagnostic-outcome heuristic when applicable). (3) *No contract*: set $b \equiv 0$, to quantify the extent to which incentives are needed at all. To isolate the role of ambiguity, we hold fixed the recommended actions and vary the uncertainty radius (e.g., η for TV balls) and the misspecification envelope (the ε_s schedule). This produces budget–welfare curves directly interpretable through our theoretical comparative statics.

4.9.1 Adversarially shifted tree MDPs

Environment design: where robustness should matter. The first family is a finite-horizon "tree MDP" designed to make incentive failures easy to observe and easy to attribute. Starting from s_0, each period t the agent chooses an action $a_t \in \{0, 1\}$ and an outcome $o_t \in \{0, 1\}$ is realized; the next state is the history $(o_0, \ldots, o_t)$, so the episode induces a depth-T outcome tree. The principal receives a terminal reward depending on the leaf (e.g., a high reward on a sparse set of "good" leaves), and the agent has an intrinsic cost for the principal-preferred action. Outcome distributions $O(s, a)$ are Bernoulli with state- and action-dependent biases, so that $\hat{O}$ can be estimated from counts, and TV uncertainty balls provide a transparent robustness model:

$$\mathcal{U}_{s,a} = \{p \in \Delta(\{0, 1\}) : \mathrm{TV}(p, \hat{O}(s, a)) \leq \eta_{s,a}\}.$$

[2] The adaptive adversary is stronger than our equilibrium model if interpreted literally; we use it as a worst-case diagnostic of how "tight" the robustness layer is.

This construction deliberately creates two types of states: (i) *high-separability states* where the two actions induce substantially different outcome probabilities, and (ii) *low-separability states* where actions are observationally similar and incentives are intrinsically expensive.

Shift construction and what it targets. In validation, Nature perturbs outcome probabilities within $\mathcal{U}_{s,a}$ to reduce the diagnostic power of the contract. For single-outcome support contracts (paying only when $o = 1$, say), the worst case pushes down $\Pr(o = 1 \mid s, a_p)$ and pushes up $\Pr(o = 1 \mid s, a)$ for deviations. This directly attacks the separability term

$$\bar{d}(s, a_p, a) = \max_{o \in \{0,1\}} \left(\inf_{p \in \mathcal{U}_{s,a_p}} p(o) - \sup_{q \in \mathcal{U}_{s,a}} q(o) \right),$$

and therefore targets the exact leverage through which transfers create incentive differences.

Main findings: robust nudging prevents incentive collapse under shift. The nominal minimal-implementation contracts perform well when the true kernel equals $\hat{O}$, but degrade sharply as we increase η or allow the adaptive adversary. In particular, deviation rates concentrate in the low-separability region of the tree, where small perturbations can flip the ranking of actions in expected transfer. By contrast, robust nudging yields near-zero deviation rates throughout the episode across the tested uncertainty radii, consistent with robust IC being enforced by construction.

Welfare decomposes in the predicted way. Relative to the no-contract baseline, both nominal and robust contracts raise value by inducing the desired actions at many states. Relative to the nominal contracts, robust nudging sacrifices value primarily through higher payments, rather than through changes in state visitation.[3] Empirically, the additional payment mass concentrates on a small fraction of states with small $\bar{d}_{\min}(s, a_p)$, providing a sharp diagnostic: robustness is not uniformly expensive, but rather "spikes" at informational bottlenecks.

Budget–welfare curves and the role of separability. Varying η traces out a smooth budget–welfare frontier. As predicted, increasing η reduces $\bar{d}$ and therefore increases the required nudge magnitude approximately as $1/\bar{d}$. Plotting realized expected payments against realized principal value yields a curve with a distinct elbow: initially, small robustness investments buy large improvements in incentive stability (and hence welfare under shift),

[3] In a finite-horizon tree, changes in early actions can change the distribution over later states. The fact that robust nudging stabilizes actions makes later-state distributions closer to training, which is exactly the point of the approach.

but beyond a point the marginal cost rises quickly because the remaining problematic states have very small $\bar{d}_{\min}$. This is the operational content of our comparative statics: robustness is cheap when outcomes are informative, and expensive precisely when actions are hard to distinguish.

4.9.2 Coin Game with controlled perturbations and non-stationary agents

Why a "Coin Game"? The second family strips away state complexity to focus on the interaction between outcome ambiguity and agent non-stationarity. At each period the agent chooses $a \in \{H, T\}$ and a coin outcome $o \in \{0, 1\}$ is realized with probability $O(a)(1)$. The principal prefers H (because $o = 1$ is more valuable downstream, or because transitions are better), but the agent's intrinsic payoff drifts over time: we model this as a time-varying cost c_t for choosing H, unknown to the principal at training time. This drift captures a realistic deployment concern: even if the principal's model of outcomes is stable, the agent's private continuation values can change with market conditions, learning, or fatigue.

Controlling ε via drift. We connect the drift to the misspecification envelope by calibrating ε_t so that the principal's estimate $\widehat{Q}$ is accurate up to ε_t, while the agent's true continuation values reflect the current c_t. Concretely, in training we fit $\widehat{Q}$ under a reference cost process, and in validation we perturb costs within a band that implies a known ε_t bound. This makes the robust IC margin $2\varepsilon_t$ economically meaningful: it is exactly the buffer needed to prevent the agent from switching actions when its private incentives shift moderately relative to what the principal anticipated.

Outcome perturbations and ambiguity sets. Simultaneously, we perturb the coin bias within uncertainty sets around $\hat{O}$. We consider both TV balls (transparent and worst-case sharp) and KL balls (smoother, capturing likelihood-based confidence regions). In all cases, the qualitative pattern is the same: as ambiguity increases, nominal contracts become fragile because they price the wrong diagnostic event, while robust nudges remain stable because they explicitly purchase a gap that survives worst-case perturbations.

Nominal contracts fail under joint drift; robust nudges trade money for stability. When either (i) outcome shift is present but the agent is stationary, or (ii) the agent drifts but outcomes are known, nominal contracts can often limp along. The failure mode is the interaction of the two: drift reduces the intrinsic advantage of a_p, while outcome shift reduces the transfer advantage of the contract, and the combination flips best responses. Robust

nudging addresses exactly this interaction by requiring

$$\inf_{p \in \mathcal{U}_{a_p}} \mathbb{E}_p[b] - \sup_{q \in \mathcal{U}_a} \mathbb{E}_q[b] \geq 2\varepsilon_t,$$

so the contract remains valid even when both terms move against the principal. Empirically, the robust policy exhibits a clean substitution pattern: as drift (and hence ε_t) grows, it increases payments linearly; as outcome ambiguity grows (shrinking $\bar{d}$), it increases payments superlinearly via the $1/\bar{d}$ channel. This is the exact tradeoff highlighted by the welfare bound.

4.9.3 Separability diagnostics as an operational tool

Predicting fragility ex ante. Across both environments, we log $\bar{d}_{\min}(s, a_p)$ along the realized trajectory and compare it to (i) whether the nominal contract experiences a deviation, and (ii) the magnitude of the robust nudge payment B_s. Two patterns are consistent and practically useful. First, nominal deviations occur almost exclusively when $\bar{d}_{\min}$ is small, even when $\widehat{Q}$ is accurate on average. Second, B_s is well explained by the simple proxy $2\varepsilon_s/\bar{d}_{\min}(s, a_p)$, validating the interpretation of separability as "incentive leverage."

Implications for system design. These diagnostics suggest a concrete workflow for practitioners: before deployment, compute (or estimate) $\bar{d}_{\min}$ under the uncertainty model implied by available data; if $\bar{d}_{\min}$ is small in frequently visited regions, then robust contracting will be expensive or infeasible under limited liability. In such cases, the right intervention is often not "more clever contracting" but *better observables*: richer measurement, audits, or redesigned outcome categories that increase separability. This is where the economics connects to policy: platforms and regulators that mandate standardized reporting, auditing protocols, or verifiable milestones can expand the feasible set of low-liability incentive schemes by increasing $\bar{d}$.

Limitations of the experimental evidence. Our experiments are intentionally stylized. The adversarial shift is worst-case within $\mathcal{U}$, whereas many applications face structured (and sometimes benign) drift. We also treat the agent as an oracle best responder in validation, which is conservative when agents are boundedly rational, but may be optimistic when agents strategically manipulate observables beyond our modeled outcome space. Finally, we focus on per-state robustness rather than end-to-end learning with contract-aware exploration. These choices reflect the purpose of this section: to isolate the mechanism predicted by the theory. The next section discusses how the framework extends when these simplifying assumptions are relaxed, and how robustness interacts with multi-agent settings, partial observability, and budget constraints.

5 Discussion and extensions

Our framework is intentionally modular: a learning layer produces nominal objects $(\hat{O}, \widehat{Q})$ together with uncertainty envelopes $(\mathcal{U}, \varepsilon)$, and a contracting layer transforms these objects into per-state limited-liability transfers that are *stable* to the corresponding misspecifications. This separation clarifies what robustness is buying. The contract does not "fix" a bad recommendation a_p (that loss is tracked by δ_s), but it can prevent a second, operationally distinct failure mode: incentive reversals when the world at deployment differs from the world in which the contract was computed. In this section we discuss extensions that preserve this modularity while relaxing modeling choices that are restrictive in applications.

5.1 Multi-agent robustness: dominant-strategy IC versus learning-stable IC

Many platform and regulatory settings involve multiple strategic agents—e.g., multiple suppliers on a marketplace, multiple departments within a firm, or a committee of auditors—whose actions jointly determine observable outcomes. A natural extension is to index agents by $i \in \{1, \dots, N\}$, let $a^i \in \mathcal{A}_i$ denote agent i's hidden action, and let $a = (a^1, \dots, a^N)$. Outcomes are generated by a kernel $O(s, a) \in \Delta(\mathcal{O})$, and the principal posts a vector of contracts $b^i(o) \geq 0$. Agent i's payoff becomes

$$R_A^i(s, a, b, o) = r^i(s, a) + b^i(o),$$

allowing for action externalities through $r^i(s, a)$ and through $O(s, a)$.

The first conceptual choice is the equilibrium notion. A *Bayes–Nash* or Markov–perfect analysis ties agent i's incentives to beliefs about other agents' responses; robust guarantees then depend on joint deviations and strategic feedback loops. If our goal is a conservative guarantee—and if the principal can commit to contracts that do not condition on unobserved actions—a more attractive target is a *dominant-strategy* style condition: for each i, the recommended action $a_p^i(s)$ should be optimal for agent i irrespective of what other agents do. In a one-shot version, a robust dominant-strategy incentive constraint takes the form

$$\inf_{p \in \mathcal{U}_{s,(a_p^i, a^{-i})}} \mathbb{E}_p[b^i(o)] + \widehat{Q}^i\big(s, a_p^i, a^{-i}\big) - \varepsilon_s^i \;\geq\; \sup_{q \in \mathcal{U}_{s,(a^i, a^{-i})}} \mathbb{E}_q[b^i(o)] + \widehat{Q}^i\big(s, a^i, a^{-i}\big) + \varepsilon_s^i$$

for all $a^i \neq a_p^i$ and all a^{-i}. This is stringent: it asks the principal to insure agent i's incentives against the worst-case play of others *and* worst-case outcome perturbations. The benefit is interpretability and deployment stability: if the constraint holds, agent i does not need to forecast other agents' behavior for the mechanism to work.

A direct generalization of our separability lever exists, but it is weaker in multi-agent environments because deviations can be masked by others' actions. Fixing a recommended profile $a_p(s)$, we can define an *agent-specific robust separability*

$$\bar{d}_i\big(s, a_p^i \to a^i; a^{-i}\big) = \max_{o \in \mathcal{O}} \left\{ \inf_{p \in \mathcal{U}_{s,(a_p^i, a^{-i})}} p(o) - \sup_{q \in \mathcal{U}_{s,(a^i, a^{-i})}} q(o) \right\},$$

and then a conservative worst-case leverage $\bar{d}_i^{\min}(s) = \min_{a^i \neq a_p^i} \inf_{a^{-i}} \bar{d}_i(s, a_p^i \to a^i; a^{-i})$. When $\bar{d}_i^{\min}(s) > 0$, a single-diagnostic-outcome contract for agent i again exists, with required payments scaling as $\varepsilon_s^i / \bar{d}_i^{\min}(s)$. Economically, the message is sharper than in the single-agent case: *externalities reduce auditability*. If other agents can "explain away" the diagnostic event, then no limited-liability transfer can cheaply isolate a unilateral deviation.

A second issue is collusion and coalitional deviations. If agents can coordinate off-platform, the relevant constraint is not unilateral dominant-strategy IC but a coalitional notion: no subset $C \subseteq \{1, \ldots, N\}$ should gain by deviating jointly. The separability object then becomes a *set* comparison between outcome distributions under a_p and under (a^C, a_p^{-C}), and the required transfers can scale with the smallest separation across coalitions, which may be prohibitively small. This limitation is not merely technical: it highlights when contractual robustness must be complemented by *organizational* interventions (anti-collusion monitoring, randomized audits, or structural separation of duties) that increase effective separability.

Finally, even in the single-agent environment, the agent in our validation analysis is an "oracle" best responder with correct beliefs about O and its own continuation values. In many deployments, agents *learn* the mapping from actions to outcomes over time and may behave according to evolving beliefs. This motivates a weaker but behaviorally plausible robustness notion that sits between ex post IC and full rationality: *learning-stable IC*. One formalization is to require that, for every belief p in a set of plausible beliefs $\mathcal{B}_{s,a}$ (e.g., the same uncertainty set $\mathcal{U}_{s,a}$ or a confidence region derived from the agent's data), the recommended action remains optimal:

$$\mathbb{E}_{o \sim p(\cdot|s,a_p)}[b_s(o)] + \widehat{Q}(s, a_p) - \varepsilon_s \;\geq\; \mathbb{E}_{o \sim p(\cdot|s,a)}[b_s(o)] + \widehat{Q}(s, a) + \varepsilon_s \quad \forall a \neq a_p.$$

This condition can be easier to satisfy than worst-case robust IC if $\mathcal{B}$ is smaller than $\mathcal{U}$, but it can also be harder if the agent's learning dynamics transiently concentrate on "wrong" models. The broader point is that robustness is ultimately relative to a *declared* misspecification class; different operational assumptions about what agents know and learn map into different classes.

5.2 Partial observability and information design

Our baseline assumes the principal conditions contracts on the current state s and the realized outcome o. In many settings the state is only partially observed: a regulator observes a noisy signal of firm quality; a platform observes engagement metrics but not true user welfare; a manager observes performance indicators but not project difficulty. Let y_t denote an observation generated by $y_t \sim Z(\cdot \mid s_t)$, and suppose the principal can condition on (y_t, o_t) but not on s_t directly. Then the principal's control problem becomes a POMDP in the belief state $\mu_t \in \Delta(\mathcal{S})$.

The robust contracting layer extends cleanly if we redefine objects on beliefs. The nominal model becomes $\hat{O}(\mu, a)$ induced by $\hat{O}(s, a)$ and μ, uncertainty sets become belief-dependent (e.g., $\mathcal{U}_{\mu,a}$ as the set of mixtures of $\mathcal{U}_{s,a}$ weighted by μ), and the per-period contract is $b_\mu(o)$ or $b_\mu(o, y)$ depending on what is contractible. The key change is that separability can collapse when the observation blurs distinctions between states in which the diagnostic outcome is informative and states in which it is not. Formally, if contracts cannot condition on y, separability is computed under the *marginal* distribution of o, which is typically less separated than the conditional distribution of o given y. This gives an economic interpretation of "information design" in our setting: making richer observables contractible (better logging, verifiable milestones, third-party attestations) increases $\bar{d}$ and can strictly expand the set of implementable policies under limited liability.

Two caveats arise. First, richer observables can create new manipulation channels: agents may strategically affect y if it is itself influenced by hidden actions or reporting. Second, when the principal's belief updates are misspecified, the mapping from (y, o) to μ becomes another source of ε-type error; robustness then must include uncertainty over filtering, not only over outcome generation. Both issues suggest that partial observability is not merely a technical extension: it is the locus where robustness, measurement, and strategic gaming interact.

5.3 Budget constraints and feasibility under limited liability

Our per-state construction highlights that robustness is "paid for" through higher expected transfers, scaling as $\varepsilon_s/\bar{d}$. In practice, principals rarely have unlimited subsidy capacity. A simple way to incorporate budget limits is to add an explicit constraint to the per-state problem:

$$\sup_{p \in \mathcal{U}_{s,a_p}} \mathbb{E}_p[b_s(o)] \le B_s,$$

where B_s is a state-dependent spending cap (or a uniform cap B). This immediately yields a feasibility bound. If we restrict attention to single-diagnostic-outcome contracts, the largest robust margin achievable at state

s is on the order of $B_s \cdot \bar{d}_{\min}(s)$; therefore a sufficient condition for robust implementation with margin $2\varepsilon_s$ is

$$B_s \gtrsim \frac{2\varepsilon_s}{\bar{d}_{\min}(s)}.$$

When this inequality fails, the model delivers a sharp operational conclusion: *robust implementation is impossible under the declared robustness envelope and the declared liability constraint.* The principal must then change something structural—choose a different recommended action with higher separability, enrich observables to increase $\bar{d}$, relax robustness (smaller uncertainty set or smaller ε_s through better value estimation), or accept a probabilistic notion of compliance.

Dynamic budget constraints complicate matters further. If the principal has an episode-wide budget $\bar{B}$, then high payments early can crowd out future robustness where it matters more. This naturally leads to a joint dynamic program in which the "shadow price" of budget enters the contract computation, and may rationalize deliberately tolerating small deviation risks in early, low-stakes states to preserve funds for later bottlenecks. Importantly, this is not a departure from our logic: it is the same robustness tradeoff, but with an additional resource constraint that couples states across time.

5.4 Policy implications for platforms and regulation

The most immediate policy implication is that robustness can be improved either by paying more or by measuring better, and these two levers are substitutes. Our separability parameter $\bar{d}$ makes this substitution concrete: when outcomes are weakly informative about hidden actions, limited-liability incentives become expensive or infeasible, even for a well-intentioned principal. This provides an economic rationale for interventions that increase verifiability: standardized reporting requirements, audit rights, tamper-evident logs, or the design of outcome categories that are hard to manipulate and highly diagnostic of effort. In platform settings, this points toward product and instrumentation design as part of the mechanism: the platform can sometimes increase $\bar{d}$ more cheaply by changing what it observes than by increasing subsidies.

Regulators also shape the robustness envelope indirectly. Legal constraints on contracting (minimum payments, restrictions on penalties, limits on contingent pay) effectively tighten limited-liability constraints and can make $\varepsilon/\bar{d}$ bottlenecks binding. Conversely, regulation can reduce uncertainty by mandating disclosure and data access, shrinking $\mathcal{U}$ and improving separability. Our framework suggests a "safe harbor" interpretation: if a principal commits to an incentive scheme that is robust with respect to a regulator-approved uncertainty model, then observed deviations are more plausibly

attributable to unmodeled manipulation or to failures of monitoring rather than to predictable distribution shift.

We close with limitations. Worst-case robustness can be conservative when the true shift is benign or structured; per-state constraints ignore the possibility of correlating incentives across time to economize on payments; and our uncertainty sets treat Nature as adversarial rather than statistically grounded unless they are carefully calibrated. Nevertheless, these are not reasons to abandon robustness; they are reasons to connect it more tightly to data (how to choose $\mathcal{U}$) and to system design (how to increase $\bar{d}$). In our view, the central tradeoff remains: robust incentive alignment is feasible when observables are sufficiently diagnostic, and costly when they are not.