WHEN DECISIONS INDUCE POLICIES: PERFORMATIVE PREDICTION WITH LATENT RESPONSES

Anonymous authors
Paper under double-blind review

ABSTRACT

When a learned decision rule is deployed in an adaptive environment, future data are shaped by agents who respond to it. Performative prediction captures this feedback through induced distribution shift, but similar observable shifts can arise from different latent policy responses, which may require different interventions. We formulate this problem as a bilevel multi-agent reinforcement learning framework in which strategic users adapt to the incentives induced by a routing rule, while the institution updates the rule from aggregate feedback and behavior-derived evidence. To infer latent responses, we compare behavior windows against reference policies representing benign and targeted behavior. Theoretically, we connect user adaptation and platform learning through a governance-loss bound, and derive finite-time guarantees for predictive, stable user adaptation under explicit interaction and prediction-error conditions. Across competitive Riichi Mahjong and non-competitive platform environments, policy-evidence-based learning better distinguishes latent strategic responses than aggregate response signals, leading to more stable long-run regimes, fewer harmful users, and improved benign-user outcomes.

1 INTRODUCTION

Machine learning models are increasingly used in decision-making domains in which model-driven decisions shape future observations (Perdomo et al., 2020; Hardt & Mendler-Dünner, 2023). In public-benefit programs, benefit decisions can affect how applicants document income, employment status, or household needs (Hardt et al., 2015; Milli et al., 2018; Kleinberg & Raghavan, 2018). Similarly, in content ranking, recommendations can change user consumption and preferences, shaping the engagement data observed later (Chaney et al., 2017; Mansoury et al., 2020; Dean & Morgenstern, 2022). These examples are not ordinary dataset shift: future observations change because the decision rule itself changes how individuals respond (Perdomo et al., 2020; Mendler-Dünner et al., 2020). Performative prediction studies this feedback by modeling a chosen parameter θ as inducing a future distribution $\mathcal{D}(\theta)$ over observations, making the dependence between model choice and future data explicit (Perdomo et al., 2020).

Performative prediction is built around the response map (Perdomo et al., 2020; Hardt & Mendler-Dünner, 2023). Performative stability and performative optimality evaluate a model on the distribution it induces, so the choice of response model determines the risk being optimized (Perdomo et al., 2020; Mendler-Dünner et al., 2020). Prior work makes this object tractable in several ways: by imposing regularity assumptions on $\mathcal{D}(\cdot)$ (Perdomo et al., 2020; Mendler-Dünner et al., 2020; Izzo et al., 2021a), by deriving it from structural models of feature response (Hardt et al., 2015; Dong et al., 2017; Miller et al., 2019; Jagadeesan et al., 2021), or by replacing the stateless map $\mathcal{D}(\theta)$ with a stateful transition such as $d_{t+1} = \mathrm{Tr}(d_t, \theta_t)$ (Brown et al., 2020; Li & Wai, 2021; Izzo et al., 2021b). These formulations make performative feedback analyzable, and in some settings computable (Jagadeesan et al., 2022; Hardt et al., 2022; Mofakhami et al., 2023). Yet aggregate-level response models need not identify the mechanism that generated the observed shift. Consider a testing rule that rewards higher scores: an increase could reflect better learning or adaptation to the scoring rule itself. This illustrates the distinction between improving an underlying outcome and optimizing its measured proxy (Goodhart, 1984; Hardt et al., 2015; Milli et al., 2018). These mechanisms can generate similar aggregate improvements while implying very different next decisions. This observation suggests a complementary perspective: when the responding population consists of

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

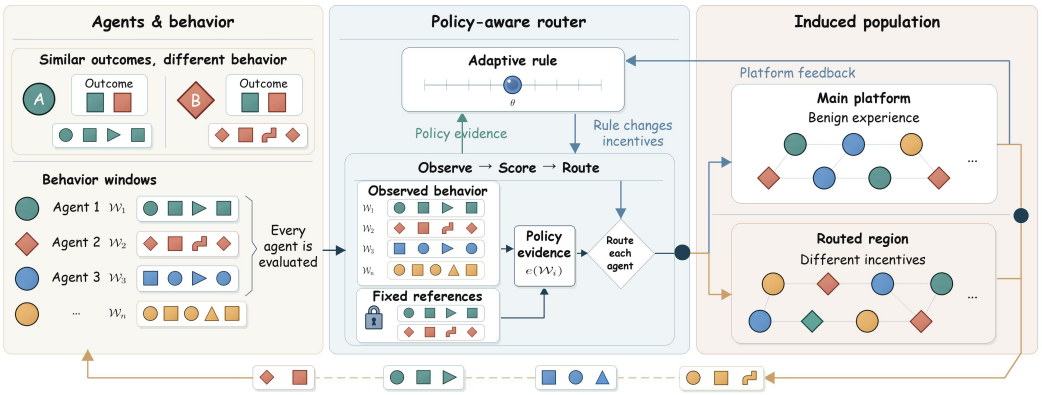


Figure 1: Policy-mediated routing in an adaptive population. Similar outcomes can arise from distinct behavior, motivating evidence from individual behavior windows. A common router evaluates each agent against fixed reference policies and applies the platform's current threshold to select the main or routed region. The resulting incentives shape subsequent user behavior, which returns for evaluation in the next round. Behavioral evidence and population feedback inform the platform's next rule.

sequential agents, the aggregate transition can be viewed as the result of agents learning from feedback, interacting with one another, and choosing actions under the current model θ_t (Buşoniu et al., 2008; Zhang et al., 2019). We take this perspective and model the performative response with MARL, where the current rule changes agent policies and the induced policies shape future observations and aggregate state (Zhang et al., 2019; Syrgkanis et al., 2015).

We study this policy-mediated response in platform regulation, a setting where the learner is not only predicting under performative feedback but also choosing rules that control the population response it induces (Perdomo et al., 2020; Jagadeesan et al., 2022; Ensign et al., 2017). We formulate the problem as a bilevel decision process: at the inner level, strategic agents respond to the current rule through their policies; at the outer level, a platform regulator updates the rule using behavior-derived evidence and aggregate feedback (Zhang et al., 2019; Hazan, 2016). This connects the response-map abstraction in performative prediction with a control problem in which the induced response is generated by adaptive agent policies (Brown et al., 2020; Li & Wai, 2021).

Contributions.

1. Policy-mediated performative response. We refine the response-map abstraction by modeling the induced distribution as the consequence of adaptive agent policies. Conditional on the interaction history, the response follows $\theta_t \mapsto \Pi_t^{\text{real}} \mapsto \mathcal{D}_t^{\text{pol}}(\theta_t)$.

2. Bilevel learning for institutional decision rules. We formulate the resulting problem as a bilevel learning process: the inner level models policy response through adaptive user learning, while the outer level updates a continuous institutional threshold under policy-mediated feedback. Our analysis connects user and platform regret to governance loss, quantifying how predictive learning and policy-update errors shape the guarantee.

3. Observable evidence for latent policy response. Because Π_t^{real} is not directly observed, we construct behavior-derived evidence through reference policies anchored to benign and targeted behavior. This evidence is an observation interface for the latent policy response, not the response mechanism itself.

2 Related Work

Our work is related to performative prediction, multi-agent reinforcement learning, and Mahjong AI. Performative prediction studies learning under deployment-induced distribution shift (Perdomo et al., 2020; Mendler-Dünner et al., 2020; Jagadeesan et al., 2022; Brown et al., 2020), while strategic classification and related decision-dependent learning models study how agents react to deployed rules (Hardt et al., 2015; Dong et al., 2017; Milli et al., 2018; Kleinberg & Raghavan, 2018). MARL models adaptive interacting policies (Buşoniu et al., 2008; Zhang et al., 2019), while learning in

games connects predictable utility variation to regret and aggregate performance (Syrgkanis et al., 2015; Capitaine et al., 2025). Our Mahjong experiment is also connected to prior work on Mahjong AI and deep reinforcement learning for imperfect-information play (Li et al., 2020; Equim, 2022; Zheng & Li, 2020; Zhao & Holden, 2022). Unlike these lines of work individually, we use MARL to instantiate the latent response mechanism in performative prediction: a platform rule induces policy adaptation, and the induced policy response shapes the future observations. A detailed discussion is provided in Appendix A.

3 PRELIMINARIES

3.1 BEHAVIOR, ROUTING, AND USER INCENTIVES

We study a platform whose decisions change users' incentives and future behavior, as in performative prediction (Perdomo et al., 2020). Consider a credit platform that routes applications to standard processing (the main region) or additional verification (the routed region). Applicants adapt their submissions to gain access while avoiding verification costs. The platform balances unnecessary verification against undetected misrepresentation. Throughout, b denotes benign users and h denotes users pursuing the behavior targeted by governance. In this example, they are legitimate applicants and applicants who misrepresent information. Figure 1 illustrates the resulting feedback loop.

The abstraction below follows the objects in this loop. The deployed threshold θ_t is the platform's decision; an executable behavior mode k is one way a strategic user can respond; the mixture $\pi_{i,t}$ records how often user i selects among these modes; and the loss vector $\ell_{i,t}$ records the incentives created by the current rule and the current behavior of others. Thus the platform and the responding population adapt on two coupled time scales. The notation does not require a person to run an explicit online-learning algorithm: $\pi_{i,t}$ may be interpreted as a reduced-form description of an individual's or population's evolving response frequencies. Later learning assumptions constrain the resulting adaptation trajectory rather than the user's internal cognition.

Behavioral evidence and routing. A behavior window $\mathcal{W} = ((s_\tau, a_\tau))_{\tau=1}^m$ records observations and actions. Fix reference policies $\pi_{\mathrm{h}}(a \mid s)$ and $\pi_{\mathrm{b}}(a \mid s)$ for targeted and benign behavior, respectively, with positive probabilities on scored actions. For a threshold θ in a nonempty set $\Theta \subset \mathbb{R}$, define

$$e(\mathcal{W}) = \sum_{\tau=1}^{m} \log \frac{\pi_{\mathrm{h}}(a_\tau \mid s_\tau)}{\pi_{\mathrm{b}}(a_\tau \mid s_\tau)}, \qquad \delta_\theta(\mathcal{W}) = \mathbf{1}\{e(\mathcal{W}) \geq \theta\}. \tag{1}$$

A larger score indicates greater compatibility with the targeted reference. The indicator $\delta_\theta = 1$ sends the interaction to the routed region; $\delta_\theta = 0$ retains it in the main region. The references stay fixed while thresholds and user behavior adapt.

Behavior modes and adaptive strategies. We represent the latent policy response Π_t^{real} at the window level. Each strategic user $i \in \{1, \ldots, N\}$ chooses among $K_i \geq 2$ modes, indexed by $k = 0, \ldots, K_i - 1$. An active mode $k \geq 1$ specifies a fixed executable policy that generates a window with law $\Pr_{i,k}$. Mode 0 is a safe outside option, such as not applying. At round t, user i samples a mode from its adaptive mixture $\pi_{i,t} \in \Delta_i$, where $\Delta_i = \{x \in \mathbb{R}_+^{K_i} : \sum_k x_k = 1\}$. Changes in these selection probabilities represent behavioral adaptation and thereby change the distribution of observed windows. The references score actions; the mixture describes which executable response is selected. In experiments this mixture can be induced by a learned policy, while in the theory it is simply the observable strategic state whose evolution we analyze.

Let $\Pr_{\mathrm{b}}$ be the benign window law. The routing probabilities are

$$r_{i,k}(\theta) = \Pr_{i,k}\{e(\mathcal{W}) \geq \theta\} \quad (k \geq 1), \qquad F_{\mathrm{b}}(\theta) = \Pr_{\mathrm{b}}\{e(\mathcal{W}) \geq \theta\}. \tag{2}$$

Thus $r_{i,k}$ is the probability that mode k is routed, and F_{b} is the benign false-routing probability. These quantities connect evidence to incentives: they determine each behavior's routing risk and the burden on benign users. For active modes, harmful exposure is $h_{i,k}(\theta) = 1 - r_{i,k}(\theta)$. The safe mode has $r_{i,0} = h_{i,0} = 0$. Write r_i, h_i for the corresponding vectors.

Routing-induced user incentives. Write $\pi = (\pi_1, \ldots, \pi_N)$ for the user mixtures and π_{-i} for the other users' mixtures. We model user loss as a routing penalty minus a behavior benefit. Other users' routing can also affect that benefit:

$$m_i(\pi, \theta) = \sum_j w_{ij} \langle r_j(\theta), \pi_j \rangle,$$

$$\ell_i(k, \pi_{-i}; \theta) = \lambda_i r_{i,k}(\theta) - a_{i,k} - b_{i,k} m_i(\pi, \theta), \quad k \geq 1. \tag{3}$$

Here $\lambda_i > 0$ is the routing penalty, $a_{i,k}$ the base benefit, and m_i the weighted routing rate of other users. The signed coefficient $b_{i,k}$ describes their effect on user i's benefit. For example, with limited lending capacity, verification of competing applications can change the benefit of applying. Utility is $-\ell_i$, and the safe-mode loss is zero.

The next assumption isolates the source of performativity. Users may change how often they select each behavior mode, but the within-mode behavior law stays fixed; otherwise a routing change could reflect either strategic adaptation or unmodeled drift inside a mode. In the credit example, modes are stable submission strategies and the mixture over them adapts.

Assumption 1 (Fixed mode laws and bounded routing incentives). *The window laws $\mathrm{Pr}_{i,k}$ for $k \geq 1$ and Pr_{b} are fixed and independent of opponents' current mixtures. The payoff parameters in (3) are fixed, with $w_{ij} \geq 0$, $w_{ii} = 0$, and $\sum_j w_{ij} \leq 1$. The safe mode is always available and has zero user loss, routing cost, and harmful exposure. Uniformly over feasible profiles and thresholds, $\|\ell_i(\cdot, \pi_{-i}; \theta)\|_\infty \leq 1$.*

These conditions give $m_i \in [0, 1]$ and make routing probabilities functions of the threshold. Their role is structural rather than behavioral: the menu of executable responses stays fixed while strategic adaptation appears as movement of $\pi_{i,t}$ over that menu. If the executable policies themselves evolve, the mode state must be enlarged; that case lies outside the fast-rate result.

3.2 Interaction protocol, governance loss, and user regret

Interaction protocol. Let $\mathcal{H}_{t-1}$ contain the observations, controls, and feedback available before round t. A controller $\mathcal{A} \in \mathcal{C}$ selects $\theta_t = \mathcal{A}_t(\mathcal{H}_{t-1}, \xi_t) \in \Theta$, where $\mathcal{C}$ is the class of measurable rules and ξ_t is internal randomness independent of future innovations. User learners $\boldsymbol{\mathcal{B}} = (\mathcal{B}_i)_i$ choose their mixtures from pre-action information. They then generate new windows, which are scored and routed under θ_t. Returns and routing outcomes provide feedback for subsequent decisions. The adaptive prediction result in Section 4.3 also makes θ_t available to users before they choose their modes.

Platform governance loss. Let $p \in (0, 1)$ be the fixed strategic-population share and $\omega_i > 0$, $\sum_i \omega_i = 1$, its user weights. Costs $C_{\mathrm{FP}}, C_{\mathrm{FN}} > 0$ price benign false routing and harmful exposure, respectively. The conditional expected governance loss and its average are

$$L_t = (1-p)C_{\mathrm{FP}}F_{\mathrm{b}}(\theta_t) + pC_{\mathrm{FN}}\sum_i \omega_i \langle h_i(\theta_t), \pi_{i,t}\rangle, \qquad \overline{L}_T = \frac{1}{T}\sum_{t=1}^{T} L_t. \tag{4}$$

To compare very different response mechanisms—a human population, multiplicative-weights-like updates, or a neural policy—we need a common trajectory-level measure of whether adaptation keeps up with the incentives it faces. External regret provides that interface without assuming per-round best responses.

Definition 1 (User regret). *For the realized mode-loss vector $\ell_{i,t} = \ell_i(\cdot, \pi_{-i,t}; \theta_t)$, user i's external regret is*

$$R_i(T) = \sum_{t=1}^{T} \langle \ell_{i,t}, \pi_{i,t}\rangle - \min_{u \in \Delta_i} \sum_{t=1}^{T} \langle \ell_{i,t}, u\rangle. \tag{5}$$

The comparator u is the best fixed mode mixture in hindsight, evaluated against the thresholds and opponents of the same run.

Small user regret means that, over the realized sequence of platform rules and opponents, the adaptation process does nearly as well as the best single mode mixture that could have been chosen in hindsight. It is therefore a measure of how quickly the responding population learns a useful behavioral response, rather than an assumption that users solve each round's optimization problem exactly. The platform seeks small $\overline{L}_T$. Methods connects these two learning processes through routing incentives.

4 Methods

Our framework couples behavioral evidence, routing, and learning (Figure 2). Evidence determines which behavior is routed; routing changes user returns; users and the platform adapt from the resulting feedback. We first convert user regret into a governance-loss bound (Theorem 1). We then control the user-regret term under fixed and adaptive thresholds (Theorems 2 and 3).

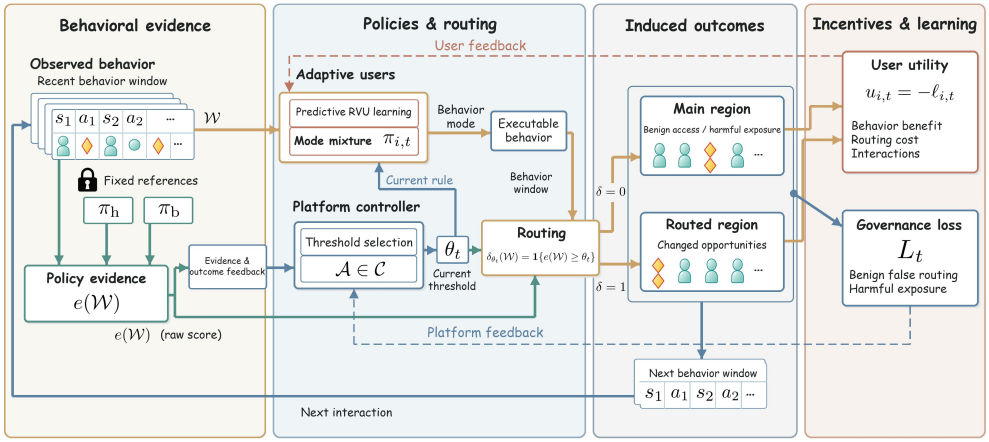


Figure 2: Theoretical framework for evidence-based governance. The platform selects θ_t from available evidence and outcome feedback. Users choose behavior modes, and fixed references score the resulting windows for routing. User utility guides subsequent behavior updates; governance feedback guides subsequent threshold updates. The next interaction repeats this loop. Solid arrows carry interactions and observations; dashed arrows carry learning feedback.

4.1 Connecting user and platform regret

Routing supports governance when harmful exposure carries a cost for the user. We express this link through the incentive inequality

$$\ell_i(k, \pi_{-i}; \theta) \geq \delta_i h_{i,k}(\theta) - \varepsilon_i(\theta). \tag{6}$$

Here $\delta_i > 0$ is an analysis scale converting exposure into user loss. The incentive gap ε_i covers the part of harmful exposure that routing does not make costly. To obtain a bound valid for every mode and opponent mixture, substitute (3) and use $0 \leq m_i \leq 1$. With $(b)_+ = \max\{b, 0\}$, this gives

$$\varepsilon_i(\theta) = \max\left\{0, \max_{1 \leq k < K_i} \left[a_{i,k} + (b_{i,k})_+ + \delta_i - (\lambda_i + \delta_i) r_{i,k}(\theta)\right]\right\}, \tag{7}$$

$$U(\theta) = (1-p) C_{\mathrm{FP}} F_{\mathrm{b}}(\theta) + p C_{\mathrm{FN}} \sum_i \widetilde{\omega}_i \varepsilon_i(\theta), \qquad \widetilde{\omega}_i = \omega_i / \delta_i. \tag{8}$$

The routing objective U combines benign false-routing cost with the incentive gaps. It is the benchmark for the cumulative governance bound below.

The platform has an analogous learning problem: even if users adapted instantly, governance could remain poor because the platform has not yet found a useful threshold. We therefore compare its threshold sequence with the best fixed threshold in hindsight.

Definition 2 (Platform regret on the routing objective). *Let $U^\star = \inf_{\theta \in \Theta} U(\theta)$. For the deployed thresholds, define*

$$Q_T = \sum_{t=1}^{T} [U(\theta_t) - U^\star], \qquad \overline{U}_T = \frac{1}{T} \sum_{t=1}^{T} U(\theta_t). \tag{9}$$

The comparator is the best fixed threshold for the common objective U.

The quantities Q_T and $R_i(T)$ therefore measure two different adaptation clocks. Q_T/T measures how quickly the institution approaches a good fixed governance rule, whereas $R_i(T)/T$ measures how quickly user i approaches a good fixed behavioral response under the rules it actually encounters. The safe option gives $\sum_{t=1}^{T} \langle \ell_{i,t}, \pi_{i,t} \rangle \leq R_i(T)$. Combining this with (6) controls cumulative harmful exposure and yields the first theorem. Importantly, this first decomposition does not require RVU or any particular user-update mechanism.

The first theorem asks what can already be said about governance if we know only how well the platform and users learn on their own objectives. The answer is a transfer inequality: before any equilibrium or RVU assumption, the remaining governance error is exactly their two regret terms.

Theorem 1 (User–platform regret decomposition). *Under Assumption 1 and the interaction protocol in Section 3.2, every controller $\mathcal{A} \in \mathcal{C}$ and user-learner profile $\boldsymbol{\mathcal{B}}$ induce a run satisfying*

$$\overline{L}_T \le U^\star + \underbrace{\frac{Q_T}{T}}_{\textit{platform regret}} + \underbrace{\frac{pC_{\mathrm{FN}}}{T}\sum_i \widetilde{\omega}_i R_i(T)}_{\textit{weighted user regret}}. \tag{10}$$

The inequality holds pathwise for the conditional expected losses, including randomized controllers.

The proof bounds $\overline{L}_T$ by $\overline{U}_T$ plus weighted user regret, then uses $\overline{U}_T = U^\star + Q_T/T$. The decomposition exposes the two adaptation time scales directly: governance can be slow because the platform has not yet identified a good threshold, because users have not yet learned how to respond to the thresholds they encounter, or both. Evidence enters through the routing probabilities: stricter verification can deter more misrepresentation while burdening more legitimate applicants. Improving active-mode routing without increasing benign false routing reduces the incentive gaps and U. Appendix B.1 proves this comparison and the regret conversion.

4.2 Predictable and stable user adaptation

Theorem 1 applies to arbitrary user adaptation, but its usefulness depends on how quickly the user regrets $R_i(T)$ grow. Ordinary no-regret learning only requires $R_i(T) = o(T)$; it does not exploit whether the incentive sequence is easy to predict, nor does it by itself control the nonstationarity created when users change one another's losses. For the sharper guarantees below, we impose a stronger trajectory property: adaptation should be faster when the next loss vector is predictable, and sufficiently conservative changes in behavior should provide a stability credit. This is the predictive *Regret bounded by Variation in Utilities* (RVU) structure (Syrgkanis et al., 2015).

Conceptually, RVU compares two rates. The prediction-error term measures how quickly the incentive environment moves relative to what the user can anticipate; the negative path-length term rewards a stable proximal response. In an interacting population, that negative term is crucial because it can absorb the loss variation generated by other users' own adaptation. We therefore use RVU as a sufficient property of the induced response trajectory, not as a claim that real users explicitly execute a particular optimization algorithm.

Assumption 2 (Predictable and stable adaptation (predictive RVU)). *Let $v_{i,t} \in \mathbb{R}^{K_i}$ predict user i's mode-loss vector before its action, and set $P_i(T) = \sum_{t=2}^{T} \|\pi_{i,t} - \pi_{i,t-1}\|_1^2$. Each user learner satisfies, for constants $\mathcal{I}_i, \alpha_i, \sigma_i > 0$, on every admissible run and every horizon T,*

$$R_i(T) \le \mathcal{I}_i + \alpha_i \sum_{t=2}^{T} \|\ell_{i,t} - v_{i,t}\|_\infty^2 - \sigma_i P_i(T). \tag{11}$$

Here $\mathcal{I}_i$ is the initialization cost, α_i converts loss-prediction error into regret, and σ_i is the stability margin attached to strategy movement.

The vector $v_{i,t}$ predicts mode losses; the reference policies in (1) instead score observed actions. These are distinct objects. Taking $v_{i,t} = \ell_{i,t-1}$ recovers the usual RVU structure: if incentives change slowly, the positive prediction-error term is small, while the negative path-length term can cancel variation induced by interacting users. Under a fixed screening rule, yesterday's losses are informative precisely when the responding population itself does not move too abruptly.

An exact RVU learner. Optimistic mirror descent predicts a strategy and then corrects it using loss feedback. Let D_i be the Bregman divergence of a mirror function, differentiable at the iterates, with $D_i(x, z) \ge \frac{1}{2}\|x - z\|_1^2$. For an auxiliary strategy $z_{i,t}$ and fixed stepsize $\eta_i > 0$, the updates are

$$\pi_{i,t} = \arg\min_{x \in \Delta_i} \{\eta_i \langle v_{i,t}, x\rangle + D_i(x, z_{i,t})\}, \quad z_{i,t+1} = \arg\min_{x \in \Delta_i} \{\eta_i \langle \ell_{i,t}, x\rangle + D_i(x, z_{i,t})\}. \tag{12}$$

Assume full expected-loss feedback, $v_{i,1} = 0$, $\|v_{i,t}\|_\infty \le 1$, and well-defined updates. If $D_i^{\max} = \sup_{u \in \Delta_i} D_i(u, z_{i,1}) < \infty$, then

$$R_i(T) \le \frac{D_i^{\max}}{\eta_i} + \eta_i \sum_{t=1}^{T} \|\ell_{i,t} - v_{i,t}\|_\infty^2 - \frac{P_i(T)}{8\eta_i}. \tag{13}$$

Thus optimistic mirror descent is an exact mechanism satisfying Assumption 2, with $\mathcal{I}_i = D_i^{\max}/\eta_i + \eta_i$, $\alpha_i = \eta_i$, and $\sigma_i = 1/(8\eta_i)$ (Appendix B.2). A smaller stepsize strengthens the negative variation

term. Related optimistic or predictive regularized online learners admit the same RVU logic. By contrast, generic no-regret learning does not automatically provide the negative path-length term. For the neural actor–critic and PPO-style policies used in the experiments, we do not assume that the training algorithm is exactly OMD: Appendix B.10 derives an approximate RVU inequality for the induced mode-probability trajectory, with an explicit policy-update defect that records departures from the ideal predictive proximal step.

Let $\beta_i = \max_{k\geq 1} |b_{i,k}|$, $B_{ij} = \beta_i^2 w_{ij}/4$, and $A_0 = \sum_i \widetilde{\omega}_i \mathcal{I}_i$. The coefficient B_{ij} measures how user j's strategy variation affects user i's loss. For mirror updates, write $A_\eta = \sum_i \widetilde{\omega}_i (D_i^{\max}/\eta_i + \eta_i)$.

A fixed rule is the cleanest stress test of the population dynamics: the institution contributes no nonstationarity, so any remaining incentive movement comes from users reacting to one another. The next theorem asks whether RVU stability can absorb precisely that endogenous variation, allowing governance loss to approach the fixed rule's certificate.

Theorem 2 (Fixed-rule governance). *Under Assumptions 1 and 2, let $\theta_t = \theta$ and $v_{i,t} = \ell_{i,t-1}$ for $t \geq 2$. If $\sum_i \widetilde{\omega}_i \alpha_i B_{ij} \leq \widetilde{\omega}_j \sigma_j$ for every j, then $\sum_i \widetilde{\omega}_i R_i(T) \leq A_0$ and the governance bound below holds. The exposure bound additionally requires $\varepsilon_i(\theta) = 0$ for all i:*

$$\overline{L}_T \leq U(\theta) + \frac{pC_{\mathrm{FN}} A_0}{T}, \qquad \sum_{t=1}^{T} \sum_i \omega_i \langle h_i(\theta), \pi_{i,t} \rangle \leq A_0. \tag{14}$$

The coupling condition has a direct interpretation: each user's stabilizing response must be strong enough to absorb the incentive variation transmitted to it through the population. When this holds, mutual adaptation does not accumulate into persistent governance error; after the initialization transient, the average loss approaches $U(\theta)$ at rate $O(T^{-1})$. In concrete online mechanisms, optimistic/entropic mirror descent (a multiplicative-weights-like proximal update) supplies the required RVU stability; a generic no-regret rule need not. Zero incentive gaps additionally bound total harmful exposure. Appendix B.3 proves both claims and gives feasible mirror stepsizes.

4.3 Prediction-aware governance under adaptive control

When the platform changes its rule, a third time scale appears: the rule itself may move faster than users can track the resulting incentives. With θ_t available before action, define the current-rule benchmark $q_{i,t} = \ell_i(\cdot, \pi_{-i,t-1}; \theta_t)$ (Capitaine et al., 2025) and prediction error $\mathcal{E}_i(T) = \sum_{t=2}^{T} \|v_{i,t} - q_{i,t}\|_\infty^2$. The benchmark holds other users at their previous mixtures and evaluates the newly deployed rule. Their subsequent adaptation accounts for the difference between $q_{i,t}$ and the realized loss, while $\mathcal{E}_i(T)$ measures how well the user's pre-action prediction tracks the immediate incentive change caused by the new rule. Thus Q_T describes how quickly the platform improves its own threshold choice, whereas $\mathcal{E}_i(T)$ describes whether those policy changes occur on a time scale that the responding population can anticipate.

A real platform usually updates its rule. This creates a failure mode absent above: θ_t may improve the platform objective while moving users' incentive landscape faster than they can track. The current-rule benchmark separates this institutional movement from opponent-induced movement, so the unified theorem exposes the platform-learning term Q_T and user tracking error $\mathcal{E}_i(T)$ separately.

Theorem 3 (Unified governance guarantee). *Under Assumptions 1 and 2 and the prediction timing above, suppose $2\sum_i \widetilde{\omega}_i \alpha_i B_{ij} \leq \widetilde{\omega}_j \sigma_j$ for every j. For every $\mathcal{A} \in \mathcal{C}$, its induced run satisfies $\sum_i \widetilde{\omega}_i R_i(T) \leq A_0 + 2\sum_i \widetilde{\omega}_i \alpha_i \mathcal{E}_i(T)$ and*

$$\overline{L}_T \leq U^\star + \frac{Q_T}{T} + \frac{pC_{\mathrm{FN}} A_0}{T} + \frac{2pC_{\mathrm{FN}}}{T} \sum_i \widetilde{\omega}_i \alpha_i \mathcal{E}_i(T). \tag{15}$$

For fixed learning constants,

$$Q_T = o(T), \quad \mathcal{E}_i(T) = o(T)\ (\forall i) \quad \Longrightarrow \quad \limsup_{T\to\infty} \overline{L}_T \leq U^\star. \tag{16}$$

Exact benchmark predictions, $v_{i,t} = q_{i,t}$ for $t \geq 2$, give $\sum_{t=1}^{T} [L_t - U(\theta_t)] \leq pC_{\mathrm{FN}} A_0$.

The bound separates three sources of transient error: the platform's own threshold-selection regret Q_T, user initialization, and the tracking error $\mathcal{E}_i(T)$ created when the incentive landscape moves faster than users can predict. The proof absorbs opponent-induced variation into the RVU terms (Appendix B.4). For previous-loss predictions, Lipschitz routing bounds $\mathcal{E}_i(T)$ by threshold variation, making the time-scale interpretation explicit: rapidly changing thresholds are costly unless current-rule information permits accurate prediction. A rate for Q_T depends on the controller's update and feedback objective; Appendix B.12 derives this connection. Section 5 evaluates the resulting evidence–routing–adaptation loop through platform outcomes.

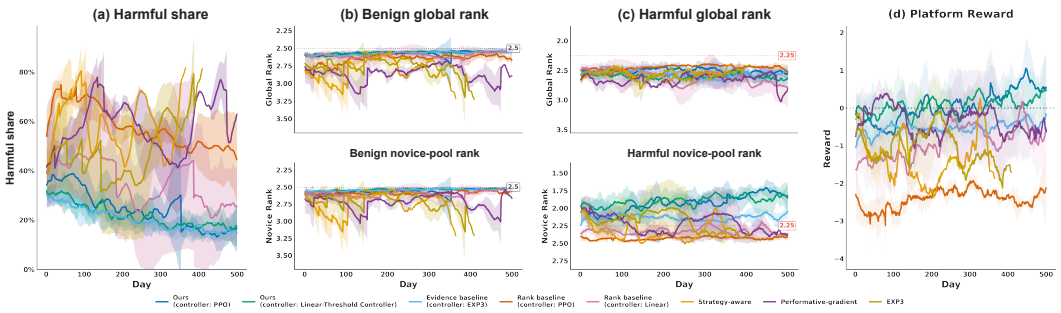


Figure 3: Competitive Mahjong learning dynamics. Lines show cross-seed means; bands show one standard deviation. Reward is the simulator-specific platform objective.

Table 1: Mahjong outcomes: cross-seed means (first line) and standard deviations (second line) of final-window averages. Ev/Rank denote Evidence/GRP; G/N denote global/novice-pool ranks. Arrows indicate preferred directions.

Method	$p_{\mathrm{h}} \downarrow$	$\bar{R}_{\mathrm{b}}^{\mathrm{G}} \downarrow$	$\bar{R}_{\mathrm{b}}^{\mathrm{N}} \downarrow$	$\bar{R}_{\mathrm{h}}^{\mathrm{G}}$	$\bar{R}_{\mathrm{h}}^{\mathrm{N}}$
Ev+PPO	20.77% ±11.96%	2.596 ±0.114	2.585 ±0.114	2.502 ±0.110	1.824 ±0.116
Ev+Linear	17.48% ±5.60%	**2.550** **±0.023**	**2.522** **±0.015**	2.644 ±0.071	1.850 ±0.150
Ev+EXP3	**15.38%** **±2.72%**	2.553 ±0.014	2.532 ±0.015	2.551 ±0.075	2.112 ±0.063
Rank+PPO	55.99% ±15.03%	2.690 ±0.099	2.569 ±0.024	2.491 ±0.047	2.429 ±0.027
Rank+Linear	46.10% ±26.40%	2.652 ±0.130	2.649 ±0.132	2.601 ±0.254	2.379 ±0.031
Strategy-aware	61.46% ±3.62%	2.969 ±0.061	2.850 ±0.033	2.535 ±0.029	2.237 ±0.082
Perf-gradient	67.76% ±5.24%	3.035 ±0.128	2.879 ±0.133	2.629 ±0.117	2.228 ±0.099
EXP3	60.53% ±7.98%	2.909 ±0.084	2.851 ±0.073	2.494 ±0.023	2.261 ±0.096

5 EXPERIMENTS

We test the governance mechanism of Theorem 1 in Mahjong and a service platform. Fixed references score behavior through $e(\mathcal{W})$, users adapt to routing incentives, and the platform updates θ_t. Harmful-user presence and benign-user outcomes measure the protection and exposure objectives of L_t.

5.1 COMPETITIVE RIICHI MAHJONG PLATFORM

Setting. The novice pool is the main region; the high-rank pool is the routed region. Benign users (b) are novices; harmful users (h) use low-rank accounts to exploit novice matches. Routing removes this advantage but costs benign users access to suitable matches.

We use 31 policies pretrained on public Tenhou records and retained at different reinforcement-learning stages. Multiplayer Elo orders their strengths as R1–R31; R3 and R29 provide the fixed references π_{h} and π_{b}. Harmful users adapt among 21 executable behavior modes with PPO (Schulman et al., 2017), balancing rank gains and survival under routing. Platform feedback combines capture, false routing, benign ranks, and threshold-movement costs.

Comparisons and metrics. EVIDENCE denotes our reference-policy score; Ev+PPO, Ev+Linear, and Ev+EXP3 combine it with three threshold controllers. Rank+PPO and Rank+Linear replace Evidence with a frozen Global Rank Predictor (GRP), holding the controller family fixed. Strategy Aware, Performative Gradient, and EXP3 also use GRP. We measure harmful-user share p_{h} in the current novice pool and benign-user mean rank globally and within that pool; lower is better. Harmful-user ranks describe the resulting operating conditions.

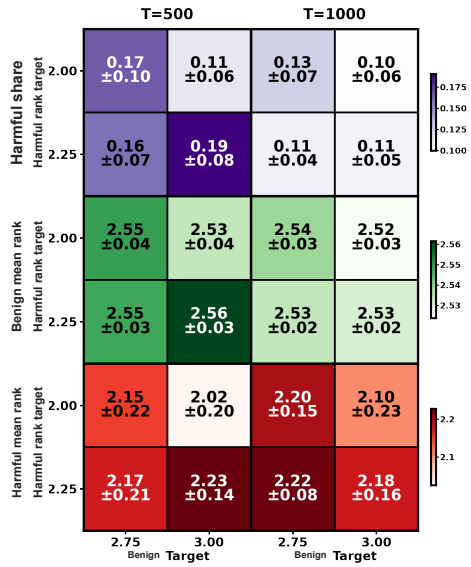


Figure 4: Mahjong outcomes at default and 2× horizons (Ev+PPO).

Table 2: Harmful-behavior detection in Mahjong.

Method	Recall ↑	FN ↓
Clustering	0.582	0.418
RNN	0.642	0.358
Transformer	0.547	0.453
EVIDENCE	**0.789**	**0.211**

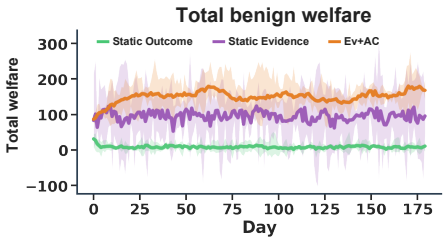


Figure 5: Non-competitive platform: total benign welfare (higher is better); bands show rollout variability.

Governance under adaptation. Evidence reduces harmful-user share from 55.99% to 20.77% with PPO and from 46.10% to 17.48% with Linear (Table 1). Benign global rank improves in both comparisons. Ev+EXP3 reaches 15.38%, extending the benefit to a finite-arm controller. Figure 3 shows that these gains persist as users adapt.

Evidence quality. Table 2 tests detection of harmful behavior windows. Evidence achieves recall 0.789, compared with 0.642 for the strongest baseline, RNN. Its false-negative rate is 0.211 versus 0.358. The router thus detects more targeted behavior as users seek to remain in the novice pool.

Persistence. For Ev+PPO, we vary both rank targets and extend $T = 500$ to $T = 1000$. Similar outcomes at both horizons (Figure 4) show that governance benefits persist during continued user learning.

5.2 NON-COMPETITIVE PLATFORM

Benign users (b) consume a service normally; harmful users (h) pursue hidden harmful behavior. Routing affects their payoffs, entry, exit, and benign welfare. Fixed references score action sequences with $e(\mathcal{W})$, testing governance beyond competitive matches.

Design and results. STATIC OUTCOME applies a fixed threshold to an outcome statistic; STATIC EVIDENCE uses a fixed threshold on Evidence. EV+AC uses Evidence and learns the threshold with actor–critic updates. We measure total benign welfare $\mathcal{W}_t = N_{\mathrm{b},t} W_t$, where $N_{\mathrm{b},t}$ is the active benign-user count and W_t is per-user welfare.

Figure 5 shows higher welfare with STATIC EVIDENCE than with STATIC OUTCOME, and a further gain with EV+AC. These gains support both parts of the mechanism: behavioral evidence improves routing, and adaptive control sustains benign-user benefit.

6 CONCLUSIONS

We study performative prediction when decisions induce adaptive policies. Reference-policy evidence links platform decisions to latent behavior, while the analysis separates governance error into platform learning and user adaptation. Under RVU and the coupling condition, a fixed rule has an $O(T^{-1})$ remainder; adaptive thresholds add platform regret and current-rule prediction error. Across both experiments, evidence improves governance and adaptive routing preserves benign-user benefit.

Limitations and outlook. The fast-rate result assumes fixed mode-conditioned window laws and a safe alternative. Extending it to evolving modes, broader behavioral diversity, measurement noise, and deployment constraints is an important next step.

Reproducibility Statement

Appendix B provides detailed proofs and specifies the learning assumptions and controller guarantees. Appendix C describes the Mahjong policy bank, reference-policy evidence, behavior modes, population dynamics, reward functions, baselines, and training hyperparameters. It also reports supplementary detection comparisons and routing ablations. Appendix D specifies the non-competitive simulator, routing rules, welfare and population dynamics, learning configurations, random seeds, and training settings. Together, these details specify the theoretical model and experimental procedures underlying the reported results.

AI Use Statement

Large language models were used to improve the clarity and fluency of the manuscript, including sentence polishing, wording refinement, and grammar checking. They were not involved in developing the research ideas, designing the model architecture, constructing the dataset, conducting experiments, or analyzing the results. All manuscript content, experimental results, figures, and references were thoroughly reviewed and verified by the authors, who assume full responsibility for the submitted work.

References

Michael Bowling, Neil Burch, Michael Bradley Johanson, and Oskari Tammelin. Heads-up limit hold'em poker is solved. *Science*, 347:145 – 149, 2015. URL https://api.semanticscholar.org/CorpusID:3796371.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. *ArXiv*, abs/2011.03885, 2020. URL https://api.semanticscholar.org/CorpusID:226281759.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365:885 – 890, 2019. URL https://api.semanticscholar.org/CorpusID:195892791.

Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *J. Mach. Learn. Res.*, 13:2617–2654, 2012. URL https://api.semanticscholar.org/CorpusID:119150.

Lucian Buşoniu, Robert Babuška, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38:156–172, 2008. URL https://api.semanticscholar.org/CorpusID:206794869.

Aymeric Capitaine, Etienne Boursier, Eric Moulines, Michael I. Jordan, and Alain Oliviero Durmus. Prediction-aware learning in multi-agent systems. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6699–6725. PMLR, 2025. URL https://proceedings.mlr.press/v267/capitaine25a.html.

Allison June-Barlow Chaney, Brandon M Stewart, and Barbara E. Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. *Proceedings of the 12th ACM Conference on Recommender Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:39558129.

Nilesh N. Dalvi, Pedro M. Domingos, Mausam, Sumit K. Sanghai, and Deepak Verma. Adversarial classification. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004. URL https://api.semanticscholar.org/CorpusID:14134514.

Sarah Dean and Jamie H. Morgenstern. Preference dynamics under personalized recommendations. *Proceedings of the 23rd ACM Conference on Economics and Computation*, 2022. URL https://api.semanticscholar.org/CorpusID:249097459.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. *Proceedings of the 2018 ACM Conference on Economics and Computation*, 2017. URL https://api.semanticscholar.org/CorpusID:34070350.

Daniel L. Ensign, Sorelle A. Friedler, Scott Neville, Carlos Eduardo Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *FAT*, 2017. URL https://api.semanticscholar.org/CorpusID:3318760.

Equim. Mortal: A mahjong ai based on deep reinforcement learning. https://github.com/Equim-chan/Mortal, 2022. Accessed: 2025-11-20.

Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *AAAI Conference on Artificial Intelligence*, 2017. URL https://api.semanticscholar.org/CorpusID:19141434.

Charles A. E. Goodhart. Problems of monetary management: The uk experience. 1984. URL https://api.semanticscholar.org/CorpusID:168522062.

Moritz Hardt and Celestine Mendler-Dünner. Performative prediction: Past and future. *ArXiv*, abs/2310.16608, 2023. URL https://api.semanticscholar.org/CorpusID:264451701.

Moritz Hardt, Nimrod Megiddo, Christos Harilaos Papadimitriou, and Mary Wootters. Strategic classification. *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, 2015. URL https://api.semanticscholar.org/CorpusID:15443770.

Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. Performative power. *ArXiv*, abs/2203.17232, 2022. URL https://api.semanticscholar.org/CorpusID:247839528.

Elad Hazan. Introduction to online convex optimization. *Found. Trends Optim.*, 2:157–325, 2016. URL https://api.semanticscholar.org/CorpusID:30482768.

He He, Jordan Lee Boyd-Graber, Kevin Kwok, and Hal Daumé. Opponent modeling in deep reinforcement learning. *ArXiv*, abs/1609.05559, 2016. URL https://api.semanticscholar.org/CorpusID:3393747.

Zachary Izzo, Lexing Ying, and James Y. Zou. How to learn when data reacts to your model: Performative gradient descent. In *International Conference on Machine Learning*, 2021a. URL https://api.semanticscholar.org/CorpusID:231924755.

Zachary Izzo, James Y. Zou, and Lexing Ying. How to learn when data gradually reacts to your model. *ArXiv*, abs/2112.07042, 2021b. URL https://api.semanticscholar.org/CorpusID:245131621.

Meena Jagadeesan, Celestine Mendler-Dünner, and Moritz Hardt. Alternative microfoundations for strategic classification. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:235129599.

Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:246441904.

Jon M. Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *Proceedings of the 2019 ACM Conference on Economics and Computation*, 2018. URL https://api.semanticscholar.org/CorpusID:49868894.

Joel Z. Leibo, Vinícius Flores Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. *ArXiv*, abs/1702.03037, 2017. URL https://api.semanticscholar.org/CorpusID:580203.

Jun-Jie Li, Sotetsu Koyamada, Qiwei Ye, Guoqing Liu, Chao Wang, Ruihan Yang, Li Zhao, Tao Qin, Tie-Yan Liu, and Hsiao-Wuen Hon. Suphx: Mastering mahjong with deep reinforcement learning. *ArXiv*, abs/2003.13590, 2020. URL https://api.semanticscholar.org/CorpusID:214713995.

Qiang Li and Hoi-To Wai. State dependent performative prediction with stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, 2021. URL https://api.semanticscholar.org/CorpusID:238259151.

Daniel Lowd and Christopher Meek. Adversarial learning. In *Knowledge Discovery and Data Mining*, 2005. URL https://api.semanticscholar.org/CorpusID:6259400.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, P. Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *ArXiv*, abs/1706.02275, 2017. URL https://api.semanticscholar.org/CorpusID:26419660.

Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and R. Burke. Feedback loop and bias amplification in recommender systems. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020. URL https://api.semanticscholar.org/CorpusID:220846905.

Celestine Mendler-Dünner, Juan C. Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *ArXiv*, abs/2006.06887, 2020. URL https://api.semanticscholar.org/CorpusID:219636100.

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, 2019. URL https://api.semanticscholar.org/CorpusID:211146670.

Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. The social cost of strategic classification. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2018. URL https://api.semanticscholar.org/CorpusID:52099203.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Kirkeby Fidjeland, Georg Ostrovski, Stig Petersen, Charlie Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015. URL https://api.semanticscholar.org/CorpusID:205242740.

Mehrnaz Mofakhami, Ioannis Mitliagkas, and Gauthier Gidel. Performative prediction with neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2023. URL https://api.semanticscholar.org/CorpusID:253180829.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. *ArXiv*, abs/2002.06673, 2020. URL https://api.semanticscholar.org/CorpusID:211133055.

John D. Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and P. Abbeel. High-dimensional continuous control using generalized advantage estimation. *CoRR*, abs/1506.02438, 2015. URL https://api.semanticscholar.org/CorpusID:3075448.

John D. Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017. URL https://api.semanticscholar.org/CorpusID:28695052.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, L. Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 2016. URL https://api.semanticscholar.org/CorpusID:515925.

Richard S. Sutton and Andrew G. Barto. Reinforcement learning: An introduction. *IEEE Trans. Neural Networks*, 9:1054–1054, 1998. URL https://api.semanticscholar.org/CorpusID:60035920.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert Schapire. Fast convergence of regularized learning in games. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/7fea637fd6d02b8f0adf6f7dc36aed93-Paper.pdf.

Ardi Tampuu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, and Raul Vicente. Multiagent cooperation and competition with deep reinforcement learning. *PLoS ONE*, 12, 2015. URL https://api.semanticscholar.org/CorpusID:12046082.

K. Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *ArXiv*, abs/1911.10635, 2019. URL https://api.semanticscholar.org/CorpusID:208268127.

Xiangyun Zhao and S. Holden. Building a 3-player mahjong ai using deep reinforcement learning. *ArXiv*, abs/2202.12847, 2022. URL https://api.semanticscholar.org/CorpusID:247154924.

Yajun Zheng and Shu-Qin Li. A review of mahjong ai research. *Proceedings of the 2020 2nd International Conference on Robotics, Intelligent Control and Artificial Intelligence*, 2020. URL https://api.semanticscholar.org/CorpusID:229380028.

A Additional Related Work

A.1 Performative prediction and strategic response

Performative prediction studies learning problems in which deploying a model changes the distribution on which the model is later evaluated (Perdomo et al., 2020). The central object is the response map $\mathcal{D}(\theta)$, which maps a deployed parameter to the induced data distribution. This framework gives rise to notions such as performative stability and performative optimality, and it has motivated algorithms for retraining, stochastic optimization, and gradient-based learning under deployment feedback (Mendler-Dünner et al., 2020; Izzo et al., 2021a;b). Regret-based formulations further study repeated learning when the learner receives feedback from decisions that themselves affect the environment (Jagadeesan et al., 2022). More recent variants consider stateful or state-dependent performative effects, where the current population state and the deployed model jointly determine future data (Brown et al., 2020; Li & Wai, 2021; Hardt & Mendler-Dünner, 2023).

A related literature studies strategic classification and adversarial response, where agents manipulate observable features in response to a deployed classifier (Hardt et al., 2015; Dong et al., 2017). This line of work analyzes the social cost of strategic behavior, the role of effort and investment, causal interpretations of strategic manipulation, and heterogeneous microfoundations for agent response (Milli et al., 2018; Kleinberg & Raghavan, 2018; Miller et al., 2019; Jagadeesan et al., 2021). Earlier adversarial-learning models similarly view prediction as an interaction between a learner and agents who can adapt to the decision rule (Dalvi et al., 2004; Lowd & Meek, 2005; Brückner et al., 2012). These works provide important models of decision-induced behavior, but they often focus on static feature response or one-step manipulation.

Our work keeps the performative-prediction viewpoint but makes the response mechanism explicitly policy-mediated. Instead of treating deployment feedback only as an aggregate response map $\theta \mapsto \mathcal{D}(\theta)$, we model the response as a latent sequential policy adaptation. Conditional on the interaction history, an institutional rule θ_t changes the environment faced by strategic agents, induces a policy response Π_t^{real}, and thereby generates the future observations:

$$\theta_t \mapsto \Pi_t^{\mathrm{real}} \mapsto \mathcal{D}_t^{\mathrm{pol}}(\theta_t).$$

This distinction matters because similar aggregate outcomes can arise from different latent policies. Our method therefore uses behavior-derived policy evidence as an observation interface for the latent response.

A.2 Multi-agent learning and strategic adaptation

Multi-agent reinforcement learning studies sequential decision making with multiple interacting agents (Buşoniu et al., 2008; Zhang et al., 2019). Compared with single-agent reinforcement learning (Sutton & Barto, 1998; Mnih et al., 2015), MARL must handle non-stationarity, strategic interaction, credit assignment, and the rapid growth of joint state-action spaces. Modern MARL methods address these challenges through centralized critics, policy-gradient methods, opponent modeling, and cooperative or competitive training objectives (Lowe et al., 2017; Foerster et al., 2017; He et al., 2016; Leibo et al., 2017; Tampuu et al., 2015).

Learning in games offers a complementary way to analyze strategic adaptation along its actual trajectory. Optimistic regularized algorithms admit regret bounds with both a positive payoff-variation term and a negative strategy-variation term; these terms can cancel across interacting learners (Syrgkanis et al., 2015). Prediction-aware learning extends this perspective to changing games in which agents forecast payoff-relevant states (Capitaine et al., 2025). Our governance analysis uses this structure with heterogeneous directed interactions and a current-rule loss prediction. A routing incentive inequality then converts weighted user regret into a bound on benign false routing and harmful exposure.

The experimental implementations use reinforcement learning to generate adaptive user behavior and platform control. User policies generate behavior windows and routing outcomes; their aggregate statistics provide feedback for platform control. The theoretical model connects user-specific strategies, incentives, and learning geometries to average governance performance during continued adaptation.

A.3 MAHJONG AI AND BEHAVIOR-LEVEL EVIDENCE

Riichi Mahjong is a challenging imperfect-information multiplayer game. It involves hidden hands, stochastic tile draws, asymmetric public and private information, opponent modeling, and long-horizon risk management. Prior Mahjong AI research has mainly focused on building stronger playing agents and reviewing algorithmic approaches for Mahjong decision making (Zheng & Li, 2020). Suphx demonstrated that deep reinforcement learning, global reward prediction, oracle guidance, and run-time policy adaptation can produce expert-level Mahjong play (Li et al., 2020). Other work and open-source systems have also explored deep reinforcement learning for Mahjong agents (Equim, 2022; Zhao & Holden, 2022). More broadly, strong game-playing systems in Go and poker show how learning, search, and game-theoretic methods can handle complex strategic domains (Silver et al., 2016; Bowling et al., 2015; Brown & Sandholm, 2019).

Our Mahjong experiment studies platform regulation when experienced players strategically enter a novice population. Final rank is informative but noisy: tile draws can cause a strong player to lose or a novice to obtain a favorable outcome. Behavior windows supply additional evidence of the policy generating these outcomes.

We therefore use reference-policy evidence. Behavior windows are compared against high-rank and low-rank Mahjong reference policies, and the resulting evidence statistic is used by the platform's outer learner. This connects Mahjong AI to our broader performative-prediction setting: trained Mahjong policies provide reference behavior models, while the platform uses evidence derived from those policies to observe latent strategic response. The experiment tests whether policy-facing evidence improves long-run platform regulation relative to coarse outcome-level signals.

B THEORETICAL ANALYSIS

The proof follows the two-regret decomposition in Theorem 1. Section B.1 converts user losses into harmful exposure and combines the two regrets. Section B.2 establishes RVU for the mirror learner and bounds how users affect one another. Sections B.3 and B.4 combine these ingredients for fixed and adaptive thresholds. Later sections give individual learning rates, inexact-update extensions, and controller-specific regret comparisons.

B.1 FROM USER REGRET TO PLATFORM GOVERNANCE LOSS

The safe mode connects learning to governance: a user with small regret cannot accumulate much more loss than that zero-loss choice. The routing incentive inequality makes harmful exposure costly to the user, up to the gap ε_i. Combining these two facts yields the platform bound.

Proof of Theorem 1. Step 1: relate each mode's loss to its harmful exposure. Fix a controller $\mathcal{A}$ and user learners $\boldsymbol{\mathcal{B}}$ under the interaction protocol in Section 3.2. Consider their induced threshold and strategy trajectory. For $k \geq 1$,

$$\begin{aligned}\delta_i h_{i,k} - \ell_i(k, \pi_{-i}; \theta) &= a_{i,k} + b_{i,k} m_i + \delta_i - (\lambda_i + \delta_i) r_{i,k} \\ &\leq a_{i,k} + (b_{i,k})_+ + \delta_i - (\lambda_i + \delta_i) r_{i,k} \leq \varepsilon_i(\theta). \end{aligned} \tag{17}$$

The inequality also holds at $k = 0$.

Step 2: compare user learning with the safe mode and aggregate. Let e_0 be the point mass on mode 0. The first line below averages the incentive inequality; the second uses the definition of user regret; the third adds benign false-routing cost and substitutes the definition of platform regret:

$$\delta_i \sum_t \langle h_i(\theta_t), \pi_{i,t} \rangle \leq \sum_t \langle \ell_{i,t}, \pi_{i,t} \rangle + \sum_t \varepsilon_i(\theta_t), \tag{18}$$

$$\sum_t \langle \ell_{i,t}, \pi_{i,t} \rangle = R_i(T) + \min_u \sum_t \langle \ell_{i,t}, u \rangle \leq R_i(T) + \sum_t \langle \ell_{i,t}, e_0 \rangle = R_i(T), \tag{19}$$

$$\overline{L}_T \leq \overline{U}_T + \frac{pC_{\mathrm{FN}}}{T} \sum_i \widetilde{\omega}_i R_i(T) = U^\star + \frac{Q_T}{T} + \frac{pC_{\mathrm{FN}}}{T} \sum_i \widetilde{\omega}_i R_i(T).$$

The argument applies to each realization of controller and learner randomness. Taking expectations also gives

$$\mathbb{E}\overline{L}_T \le U^\star + \frac{\mathbb{E}Q_T}{T} + \frac{pC_{\mathrm{FN}}}{T}\sum_i \widetilde{\omega}_i \mathbb{E}R_i(T). \tag{20}$$

□

The conversion uses $R_i(T)$ on the same induced run as the platform loss. The platform term Q_T enters through the exact identity $\overline{U}_T = U^\star + Q_T/T$ from Definition 2.

When routing removes the incentive gap. The zero-gap condition is

$$\varepsilon_i(\theta) = 0 \quad \Longleftrightarrow \quad r_{i,k}(\theta) \ge \frac{a_{i,k} + (b_{i,k})_+ + \delta_i}{\lambda_i + \delta_i} \quad (\forall k \ge 1). \tag{21}$$

Meaning of the reference-policy score. Let $\rho(\cdot \mid s)$ be a user's action distribution and D_{KL} the Kullback–Leibler divergence, with positive reference probabilities on the support of ρ. Then

$$\mathbb{E}_{a\sim\rho(\cdot|s)} \log \frac{\pi_{\mathrm{h}}(a \mid s)}{\pi_{\mathrm{b}}(a \mid s)} = D_{\mathrm{KL}}(\rho\|\pi_{\mathrm{b}}) - D_{\mathrm{KL}}(\rho\|\pi_{\mathrm{h}}). \tag{22}$$

Both divergences condition on s. The score measures relative compatibility with the two reference policies. Routing then compares it with θ in (1). To derive the identity,

$$\begin{aligned} D_{\mathrm{KL}}(\rho\|\pi_{\mathrm{b}}) - D_{\mathrm{KL}}(\rho\|\pi_{\mathrm{h}}) &= \sum_a \rho(a \mid s)\left[\log\frac{\rho(a \mid s)}{\pi_{\mathrm{b}}(a \mid s)} - \log\frac{\rho(a \mid s)}{\pi_{\mathrm{h}}(a \mid s)}\right] \\ &= \sum_a \rho(a \mid s)\log\frac{\pi_{\mathrm{h}}(a \mid s)}{\pi_{\mathrm{b}}(a \mid s)}. \end{aligned}$$

The terms involving $\log\rho$ cancel, proving (22). The window score accumulates these comparisons along the observed state-action sequence. Fixed references provide a common scale for comparing actions across user adaptations.

Evidence comparison. Consider scores e, e' and operating thresholds θ, θ' on the same mode laws and payoff parameters. Primes mark the probabilities, gaps, and objective obtained with e'. Their explicit threshold-dependent comparison is

$$\begin{aligned} &F'_{\mathrm{b}}(\theta') \le F_{\mathrm{b}}(\theta), \quad r'_{i,k}(\theta') \ge r_{i,k}(\theta) \quad (\forall i, k \ge 1) \\ &\Longrightarrow \quad \varepsilon'_i(\theta') \le \varepsilon_i(\theta)\ (\forall i), \qquad U'(\theta') \le U(\theta). \end{aligned} \tag{23}$$

This compares the resulting operating points: changing a score is useful for the bound when it improves these routing probabilities. Suppressing the two thresholds only in the following display, the same argument gives a sensitivity bound:

$$\begin{aligned} |F'_{\mathrm{b}} - F_{\mathrm{b}}| \le d_{\mathrm{b}}, \quad \|r'_i - r_i\|_\infty \le d_i \quad \Longrightarrow \quad &|\varepsilon'_i - \varepsilon_i| \le (\lambda_i + \delta_i)d_i, \\ &|U' - U| \le (1-p)C_{\mathrm{FP}}d_{\mathrm{b}} + pC_{\mathrm{FN}}\sum_i \widetilde{\omega}_i(\lambda_i + \delta_i)d_i. \end{aligned} \tag{24}$$

The sensitivity bound follows from $|\max_j x_j - \max_j y_j| \le \max_j |x_j - y_j|$. It quantifies how errors in the routing probabilities affect the incentive gaps and U.

Probabilistic routing. For a routing probability $d_\theta(\mathcal{W}) \in [0,1]$, set $r_{i,k}(\theta) = \mathbb{E}_{i,k}d_\theta(\mathcal{W})$ for active modes and $F_{\mathrm{b}}(\theta) = \mathbb{E}_{\mathrm{b}}d_\theta(\mathcal{W})$. The incentive conversion uses these probabilities and $h_{i,k} = 1 - r_{i,k}$, so the same proof applies. The hard threshold in (1) and the sigmoid router in the non-competitive simulator are two ways of generating the routing probabilities.

B.2 RVU FOR OPTIMISTIC MIRROR DESCENT

This section establishes the learning condition used in the main theorems. We first derive the mirror learner's regret bound, then control the loss variation caused by other users. These are separate ingredients: RVU describes the learner; the interaction bound describes the game.

Mirror geometry and the RVU inequality. For $D_i(x,z) = \psi_i(x) - \psi_i(z) - \langle \nabla\psi_i(z), x - z\rangle$, two admissible geometries are

$\psi_i(x)$	$z_{i,1}$	$D_i^{\max}$
$\sum_k x_k \log x_k$	$\mathbf{1}/K_i$	$\log K_i$
$K_i\|x\|_2^2/2$	$\mathbf{1}/K_i$	$(K_i-1)/2$

$$D_i(x,z) \geq \tfrac{1}{2}\|x - z\|_1^2. \tag{25}$$

Bounded losses preserve interior iterates for entropy; boundary comparators are admissible. Write $P_i(T) = \sum_{t=2}^{T} \|\pi_{i,t} - \pi_{i,t-1}\|_1^2$. The optimistic mirror inequality (Syrgkanis et al., 2015) is

$$R_i(T) \leq \frac{D_i^{\max}}{\eta_i} + \eta_i \sum_{t=1}^{T} \|\ell_{i,t} - v_{i,t}\|_\infty^2 - \frac{P_i(T)}{8\eta_i}. \tag{26}$$

Proof. Step 1: apply optimality to the prediction and correction updates. Suppress i and write $x_t = \pi_{i,t}$. For any fixed comparator $u \in \Delta_i$, the two optimality conditions give

$$\begin{aligned} \eta\langle \ell_t, z_{t+1} - u\rangle &\leq D(u, z_t) - D(u, z_{t+1}) - D(z_{t+1}, z_t), \\ \eta\langle v_t, x_t - z_{t+1}\rangle &\leq D(z_{t+1}, z_t) - D(z_{t+1}, x_t) - D(x_t, z_t). \end{aligned} \tag{27}$$

Step 2: isolate prediction error and retain the negative movement terms. Using $\eta\langle \ell_t - v_t, x_t - z_{t+1}\rangle \leq \eta^2\|\ell_t - v_t\|_\infty^2 + \|x_t - z_{t+1}\|_1^2/4$,

$$\begin{aligned} \langle \ell_t, x_t - u\rangle \leq \frac{D(u, z_t) - D(u, z_{t+1})}{\eta} + \eta\|\ell_t - v_t\|_\infty^2 \\ - \frac{\|z_{t+1} - x_t\|_1^2}{4\eta} - \frac{\|x_t - z_t\|_1^2}{2\eta}. \end{aligned} \tag{28}$$

Step 3: sum over time and recover strategy variation. The triangle inequality relates consecutive strategies to the auxiliary iterates; the divergence terms telescope. Together they yield

$$P_i(T) \leq 2\sum_{t=1}^{T} \big(\|x_t - z_t\|_1^2 + \|z_{t+1} - x_t\|_1^2\big),$$

$$\sum_t \langle \ell_t, x_t - u\rangle \leq \frac{D(u, z_1)}{\eta} + \eta \sum_t \|\ell_t - v_t\|_\infty^2 - \frac{P_i(T)}{8\eta}.$$

Maximize over $u \in \Delta_i$. □

Since $v_{i,1} = 0$ and $\|\ell_{i,1}\|_\infty \leq 1$, equation (26) proves that the mirror learner satisfies the predictive RVU property in Assumption 2, with the constants stated after (12).

Relation to RVU learning. The classical Regret bounded by Variation in Utilities (RVU) analysis bounds regret by variation in successive utility vectors and retains a negative strategy-variation term (Syrgkanis et al., 2015). Writing utilities as $u_{i,t} = -\ell_{i,t}$ preserves these variation norms. Choosing $v_{i,t} = \ell_{i,t-1}$ in (26) gives this bound structure; a general predictor gives the predictive version used here. Theorems 2 and 3 state the weighted coefficient conditions that cancel interaction-induced variation. The mirror calculation establishes the required learning inequality, while those coupling conditions determine when it yields the fast governance bounds.

How one user's learning changes another's losses. The coefficient B_{ij} measures the influence of user j's strategy variation on user i's squared loss variation. The following estimate is what allows the negative RVU terms to absorb these interactions. For probability vectors x, y and $r \in [0,1]^K$,

$$\begin{aligned} |\langle r, x - y\rangle| = |\langle r - \tfrac{1}{2}\mathbf{1}, x - y\rangle| \leq \tfrac{1}{2}\|x - y\|_1, \\ |m_i(\pi, \theta) - m_i(\pi', \theta)| \leq \tfrac{1}{2}\sum_j w_{ij}\|\pi_j - \pi_j'\|_1. \end{aligned}$$

Consequently, with $d_j = \|\pi_j - \pi_j'\|_1$,

$$\begin{aligned} \|\ell_i(\cdot, \pi_{-i}; \theta) - \ell_i(\cdot, \pi'_{-i}; \theta)\|_\infty^2 &\leq \frac{\beta_i^2}{4}\Big(\sum_j w_{ij} d_j\Big)^2 \\ &\leq \frac{\beta_i^2}{4}\Big(\sum_j w_{ij}\Big)\sum_j w_{ij} d_j^2 \leq \sum_j B_{ij} d_j^2. \end{aligned} \tag{29}$$

B.3 Proof of fixed-rule governance

The threshold is fixed, so with the previous-loss predictor all prediction error comes from users changing their mixtures. The coupling condition makes the resulting positive variation terms no larger than the negative RVU terms.

Proof of Theorem 2. Step 1: substitute the interaction bound into RVU and sum across users. At fixed θ, equations (11) and (29) give

$$R_i(T) \le \mathcal{I}_i + \alpha_i \sum_j B_{ij} P_j(T) - \sigma_i P_i(T),$$

$$\sum_i \widetilde{\omega}_i R_i(T) \le A_0 + \sum_j \underbrace{\left(\sum_i \widetilde{\omega}_i \alpha_i B_{ij} - \widetilde{\omega}_j \sigma_j \right)}_{\le 0} P_j(T) \le A_0. \tag{30}$$

Step 2: convert weighted user regret into governance loss. Substitution into (10) proves (14). If all gaps vanish, equations (18)–(19) also give

$$\sum_{t,i} \omega_i \langle h_i(\theta), \pi_{i,t} \rangle \le \sum_i \widetilde{\omega}_i R_i(T) \le A_0.$$

□

From user adaptation to aggregate governance. At each round, the strategies induce a probability distribution over user–mode pairs, $\mu_t(i,k) = \omega_i \pi_{i,t}(k)$. Its routed mass is $\sum_{i,k} \mu_t(i,k) r_{i,k}(\theta_t)$ and its harmful mass is $\sum_{i,k} \mu_t(i,k) h_{i,k}(\theta_t)$, exactly the strategic quantities used in the governance loss. These quantities aggregate the current policies at each round. Theorem 2 bounds average governance loss by $U(\theta) + pC_{\mathrm{FN}} A_0/T$; adaptive control adds platform regret and prediction error. The aggregate bound follows directly from the weighted regrets, with each user's comparator retained as in Definition 1.

Feasible stepsizes. For the mirror realization and fixed $\kappa_i > 0$, $\eta_i = s\kappa_i$ satisfies the coupling condition with coefficient χ whenever

$$0 < s^2 \le \min_{j:\, \sum_i \widetilde{\omega}_i \kappa_i B_{ij} > 0} \frac{\widetilde{\omega}_j}{8\chi \kappa_j \sum_i \widetilde{\omega}_i \kappa_i B_{ij}}, \qquad \chi = \begin{cases} 1, & \text{fixed rule,} \\ 2, & \text{dynamic rule,} \\ 3, & \text{estimated feedback.} \end{cases} \tag{31}$$

The minimum over an empty set is $+\infty$.

B.4 Combining platform regret with predictive user regret

The current-rule benchmark $q_{i,t}$ separates two changes: opponents move from their previous mixtures, and the user's predictor may differ from the benchmark. RVU absorbs the first change under the coupling condition; $\mathcal{E}_i(T)$ records the second.

Proof of Theorem 3. Step 1: split loss-prediction error into interaction and benchmark error. For $t \ge 2$, both opponent profiles are evaluated at θ_t:

$$\begin{aligned} \|\ell_{i,t} - v_{i,t}\|_\infty^2 &\le 2\|\ell_{i,t} - q_{i,t}\|_\infty^2 + 2\|q_{i,t} - v_{i,t}\|_\infty^2 \\ &\le 2 \sum_j B_{ij} \|\pi_{j,t} - \pi_{j,t-1}\|_1^2 + 2\|q_{i,t} - v_{i,t}\|_\infty^2. \end{aligned} \tag{32}$$

Step 2: apply RVU and cancel interaction-induced variation. Hence

$$R_i(T) \le \mathcal{I}_i + 2\alpha_i \sum_j B_{ij} P_j(T) + 2\alpha_i \mathcal{E}_i(T) - \sigma_i P_i(T), \tag{33}$$

$$\sum_i \widetilde{\omega}_i R_i(T) \le A_0 + 2\sum_i \widetilde{\omega}_i \alpha_i \mathcal{E}_i(T) + \sum_j \underbrace{\left(2\sum_i \widetilde{\omega}_i \alpha_i B_{ij} - \widetilde{\omega}_j \sigma_j\right)}_{\le 0} P_j(T). \tag{34}$$

Step 3: combine the two regrets in the platform bound. Theorem 1 yields

$$\begin{aligned} \overline{L}_T &\le \overline{U}_T + \frac{pC_{\mathrm{FN}}}{T}\left[A_0 + 2\sum_i \widetilde{\omega}_i \alpha_i \mathcal{E}_i(T)\right] \\ &= U^\star + \frac{Q_T}{T} + \frac{pC_{\mathrm{FN}}}{T}\left[A_0 + 2\sum_i \widetilde{\omega}_i \alpha_i \mathcal{E}_i(T)\right]. \end{aligned} \tag{35}$$

Taking upper limits proves (16); setting $\mathcal{E}_i(T) = 0$ gives the constant cumulative excess bound. □

For the mirror realization, (33) becomes

$$R_i(T) \le \frac{D_i^{\max}}{\eta_i} + \eta_i + 2\eta_i \sum_j B_{ij} P_j(T) + 2\eta_i \mathcal{E}_i(T) - \frac{P_i(T)}{8\eta_i}. \tag{36}$$

Threshold-variation specialization. Suppose $|r_{i,k}(\theta) - r_{i,k}(\theta')| \le M_i|\theta - \theta'|$ uniformly over modes and feasible thresholds. Define $G_i = \lambda_i M_i + \beta_i \sum_j w_{ij} M_j$ and $P_\theta(T) = \sum_{t=2}^T |\theta_t - \theta_{t-1}|^2$. For a fixed profile π,

$$|m_i(\pi, \theta) - m_i(\pi, \theta')| \le \sum_j w_{ij} M_j |\theta - \theta'|,$$

$$\|\ell_i(\cdot, \pi_{-i}; \theta) - \ell_i(\cdot, \pi_{-i}; \theta')\|_\infty \le \Big(\lambda_i M_i + \beta_i \sum_j w_{ij} M_j\Big)|\theta - \theta'|.$$

Taking $v_{i,t} = \ell_i(\cdot, \pi_{-i,t-1}; \theta_{t-1})$ gives

$$\mathcal{E}_i(T) \le G_i^2 P_\theta(T), \qquad \overline{L}_T \le \overline{U}_T + \frac{pC_{\mathrm{FN}}}{T}\left[A_0 + 2\sum_i \widetilde{\omega}_i \alpha_i G_i^2 P_\theta(T)\right]. \tag{37}$$

A score density bounded by M_i implies the Lipschitz premise. Two useful path regimes are

$$|\theta_t - \theta_{t-1}| \le \Delta \implies P_\theta(T)/T \le \Delta^2, \qquad P_\theta(T) = o(T) \implies (\overline{L}_T - \overline{U}_T)_+ \to 0.$$

The second implication uses the previous-loss predictor and fixed learning constants.

B.5 Individual regret and stepsize selection

The main governance bounds require a weighted sum of user regrets. This section additionally bounds each user's regret and explains how a stepsize chosen before the run yields an individual learning rate. The derivation keeps the same comparator as Definition 1.

Define $Q_{i,z}(v) = \arg\min_{x \in \Delta_i}\{\eta_i \langle v, x\rangle + D_i(x, z)\}$. For $x = Q_{i,z}(v)$ and $y = Q_{i,z}(v')$, optimality and strong convexity imply

$$\begin{aligned} \|x - y\|_1^2 &\le \eta_i \langle v' - v, x - y\rangle \le \eta_i \|v - v'\|_\infty \|x - y\|_1, \\ \|\pi_{i,t} - \pi_{i,t-1}\|_1 &\le \|Q_{i,z_{i,t}}(v_{i,t}) - Q_{i,z_{i,t}}(0)\|_1 + \|Q_{i,z_{i,t-1}}(\ell_{i,t-1}) - Q_{i,z_{i,t-1}}(v_{i,t-1})\|_1 \\ &\le \eta_i \|v_{i,t}\|_\infty + \eta_i \|\ell_{i,t-1} - v_{i,t-1}\|_\infty \le 3\eta_i, \qquad P_i(T) \le 9T\eta_i^2. \end{aligned} \tag{38}$$

Theorem 4 (Individual learning rates). *Under the predictive mirror model, arbitrary fixed stepsizes $\eta_i > 0$ satisfy*

$$R_i(T) \le \frac{D_i^{\max}}{\eta_i} + \eta_i + 18\eta_i T \sum_j B_{ij}\eta_j^2 + 2\eta_i \mathcal{E}_i(T). \tag{39}$$

For fixed ratios $\kappa_i > 0$, let

$$d = \max_i \frac{D_i^{\max}}{\kappa_i} > 0, \qquad c = 18 \max_i \left(\kappa_i \sum_j B_{ij}\kappa_j^2\right), \qquad e_T = \max_i \kappa_i[1 + 2\mathcal{E}_i(T)].$$

If a bound $\bar{e}_T \ge e_T$ is available before the run and valid for its selected stepsizes, then

$$\eta_i = s^\star \kappa_i, \qquad s^\star = \min\left\{\left(\frac{d}{cT}\right)^{1/4}, \sqrt{\frac{d}{\bar{e}_T}}\right\}, \tag{40}$$

$$\max_i R_i(T) \le 2d^{3/4}(cT)^{1/4} + 2\sqrt{d\bar{e}_T}. \tag{41}$$

Set $(d/(cT))^{1/4} = +\infty$ when $c = 0$.

Proof. Insert (38) into (36) and drop $-P_i/(8\eta_i)$. For $\eta_i = s\kappa_i$,

$$\max_i R_i(T) \le \frac{d}{s} + cTs^3 + e_T s.$$

At $s = s^\star$,

$$\frac{d}{s^\star} = \max\{d^{3/4}(cT)^{1/4}, \sqrt{d\bar{e}_T}\} \le d^{3/4}(cT)^{1/4} + \sqrt{d\bar{e}_T},$$
$$cT(s^\star)^3 \le d^{3/4}(cT)^{1/4}, \qquad e_T s^\star \le \sqrt{d\bar{e}_T}.$$

Summation proves (41), including $c = 0$. □

For fixed model constants, the pre-run prediction guarantees yield

$\max_i \mathcal{E}_i(T)$	cumulative regret bound	average regret bound
$O(T^{1/2})$	$O(T^{1/4})$	$O(T^{-3/4})$
$\le 4(T-1)$	$O(T^{1/2})$	$O(T^{-1/2})$

(42)

The stepsize is fixed within each run. A fixed positive cap enforcing (31) preserves these asymptotic bounds.

B.6 Empirical equilibrium under a fixed rule

Individual regret also has an equilibrium interpretation when the rule is fixed. At fixed θ, define the empirical joint distribution and its coarse correlated equilibrium error by

$$\nu_T = \frac{1}{T}\sum_{t=1}^{T} \bigotimes_i \pi_{i,t}, \qquad \epsilon_{\mathrm{CCE}}(\nu_T) = \max\left\{0, \max_i \frac{R_i(T)}{T}\right\}. \tag{43}$$

The affine game in (3) is the independent mixed extension of its pure-mode game. Therefore the fixed unilateral deviation gain equals $R_i(T)/T$. With $v_{i,t} = \ell_{i,t-1}$,

$$\epsilon_{\mathrm{CCE}}(\nu_T) \le \max_i \left\{\frac{D_i^{\max}}{\eta_i T} + \frac{\eta_i}{T} + 9\eta_i \sum_j B_{ij}\eta_j^2\right\}, \tag{44}$$
$$\eta_i = \kappa_i T^{-1/4} \quad \Longrightarrow \quad \epsilon_{\mathrm{CCE}}(\nu_T) = O(T^{-3/4}).$$

The first line follows from (26), (29), and (38). The empirical-distribution rate uses horizon-dependent stepsizes; Theorem 2 uses fixed learning constants for its $O(T^{-1})$ governance remainder.

B.7 Discrete payoff-state prediction

A predictor may forecast a discrete payoff state instead of a full loss vector. If a correct state prediction reconstructs the benchmark loss, prediction mistakes control the same $\mathcal{E}_i(T)$ that appears in Theorem 3.

Under (12) and the coupling condition of Theorem 3, let S_t be a finite payoff state and $\widehat{S}_{i,t}$ its prediction. Under the reconstruction property $\widehat{S}_{i,t} = S_t \Rightarrow v_{i,t} = q_{i,t}$,

$$\mathrm{Mist}_i(T) = \sum_{t=2}^{T} \mathbf{1}\{\widehat{S}_{i,t} \neq S_t\}, \qquad \mathcal{E}_i(T) \le 4\,\mathrm{Mist}_i(T),$$
$$\overline{L}_T \le \overline{U}_T + \frac{pC_{\mathrm{FN}}}{T}\left[A_\eta + 8\sum_i \widetilde{\omega}_i \eta_i\,\mathrm{Mist}_i(T)\right]. \tag{45}$$

The factor 4 follows from $\|v_{i,t} - q_{i,t}\|_\infty \le 2$. The comparator remains the fixed mixture in (5); the reconstructed object is the full current-rule mode-loss vector.

B.8 Learning guarantees and residuals

The core conversion and the RVU rate play different roles. The conversion needs a bound on user regret; RVU supplies the sharper bound through variation cancellation. This section tracks how approximation errors enter that same chain. Theorem 1 accepts any upper bounds on the user learners' regrets:

$$R_i(T) \le \mathcal{R}_i(T) \quad (\forall i) \quad \Longrightarrow \quad \overline{L}_T \le U^\star + \frac{Q_T}{T} + \frac{pC_{\mathrm{FN}}}{T}\sum_i \widetilde{\omega}_i \mathcal{R}_i(T). \tag{46}$$

An additive error in the predictive RVU property propagates linearly. Under the coupling condition of Theorem 3,

$$R_i(T) \le \mathcal{I}_i + \alpha_i \sum_{t=2}^{T} \|\ell_{i,t} - v_{i,t}\|_\infty^2 - \sigma_i P_i(T) + \Xi_i(T), \quad \Xi_i(T) \ge 0$$
$$\Longrightarrow \quad \overline{L}_T \le U^\star + \frac{Q_T}{T} + \frac{pC_{\mathrm{FN}}}{T}\left[A_0 + 2\sum_i \widetilde{\omega}_i \alpha_i \mathcal{E}_i(T) + \sum_i \widetilde{\omega}_i \Xi_i(T)\right]. \tag{47}$$

This follows by retaining $\sum_i \widetilde{\omega}_i \Xi_i(T)$ in (34).

B.9 Estimated feedback and inexact policy updates

Estimated losses and imperfect optimization introduce two distinct errors. Feedback error changes the losses used for learning; update residuals measure how far a policy step falls short of mirror-update optimality. We define both before deriving their effect on governance.

Definition 3 (User-update residuals). *For feasible iterates $\pi_{i,t}, z_{i,t}, z_{i,t+1}$ with defined mirror gradients, predictions $\|v_{i,t}\|_\infty \le 1$, $v_{i,1} = 0$, and estimates $\|\widehat{\ell}_{i,t}\|_\infty \le 1$, define*

$$f^x_{i,t} = \eta_i v_{i,t} + \nabla\psi_i(\pi_{i,t}) - \nabla\psi_i(z_{i,t}), \qquad \rho^x_{i,t} = \sup_{u\in\Delta_i} \langle f^x_{i,t}, \pi_{i,t} - u\rangle,$$
$$f^z_{i,t} = \eta_i \widehat{\ell}_{i,t} + \nabla\psi_i(z_{i,t+1}) - \nabla\psi_i(z_{i,t}), \qquad \rho^z_{i,t} = \sup_{u\in\Delta_i} \langle f^z_{i,t}, z_{i,t+1} - u\rangle,$$
$$\mathcal{M}_i(T) = \sum_{t=1}^{T}(\rho^x_{i,t} + \rho^z_{i,t}), \qquad d_{i,t} = \|\widehat{\ell}_{i,t} - \ell_{i,t}\|_\infty,$$
$$D_i^{\mathrm{fb}}(T) = \sum_{t=1}^{T} d_{i,t}, \qquad E_i^{\mathrm{fb}}(T) = \sum_{t=2}^{T} d_{i,t}^2. \tag{48}$$

The residuals are nonnegative; exact mirror updates give $\rho^x_{i,t} = \rho^z_{i,t} = 0$.

Theorem 5 (Governance with inexact learning). *Under the routing model, prediction timing, mirror geometry, and Definition 3, suppose* $3\sum_i \widetilde{\omega}_i \eta_i B_{ij} \le \widetilde{\omega}_j/(8\eta_j)$ *for every* j*. Then*

$$\overline{L}_T \le \overline{U}_T + \frac{pC_{\mathrm{FN}}}{T}\left[A_\eta + \sum_i \widetilde{\omega}_i \left\{3\eta_i[\mathcal{E}_i(T) + E_i^{\mathrm{fb}}(T)] + 2D_i^{\mathrm{fb}}(T) + \frac{\mathcal{M}_i(T)}{\eta_i}\right\}\right]. \tag{49}$$

Proof. Definition 3 adds $\rho_{i,t}^x + \rho_{i,t}^z$ to the right side of the summed three-point inequalities (27). Thus, for the estimated-loss regret $\widehat{R}_i(T)$,

$$\begin{aligned}
\widehat{R}_i(T) &\le \frac{D_i^{\max}}{\eta_i} + \eta_i \sum_t \|\widehat{\ell}_{i,t} - v_{i,t}\|_\infty^2 - \frac{P_i(T)}{8\eta_i} + \frac{\mathcal{M}_i(T)}{\eta_i},\\
R_i(T) &\le \widehat{R}_i(T) + 2D_i^{\mathrm{fb}}(T),\\
\|\widehat{\ell}_{i,t} - v_{i,t}\|_\infty^2 &\le 3\|\ell_{i,t} - q_{i,t}\|_\infty^2 + 3\|q_{i,t} - v_{i,t}\|_\infty^2 + 3d_{i,t}^2 \quad (t \ge 2).
\end{aligned}$$

Since $\|\widehat{\ell}_{i,1}\|_\infty \le 1$, equation (29) yields

$$\begin{aligned}
R_i(T) \le \frac{D_i^{\max}}{\eta_i} + \eta_i + 3\eta_i \sum_j B_{ij} P_j(T) - \frac{P_i(T)}{8\eta_i}\\
+ 3\eta_i[\mathcal{E}_i(T) + E_i^{\mathrm{fb}}(T)] + 2D_i^{\mathrm{fb}}(T) + \frac{\mathcal{M}_i(T)}{\eta_i}.
\end{aligned}$$

Weighting by $\widetilde{\omega}_i$ cancels the P_j terms; (10) completes the proof. □

For positive entropic iterates, these residuals have the explicit form

$$\begin{aligned}
f_{i,t}^x &= \eta_i v_{i,t} + \log(\pi_{i,t}/z_{i,t}), \qquad \rho_{i,t}^x = \langle f_{i,t}^x, \pi_{i,t}\rangle - \min_k f_{i,t}^x(k),\\
f_{i,t}^z &= \eta_i \widehat{\ell}_{i,t} + \log(z_{i,t+1}/z_{i,t}), \quad \rho_{i,t}^z = \langle f_{i,t}^z, z_{i,t+1}\rangle - \min_k f_{i,t}^z(k).
\end{aligned} \tag{50}$$

Costly alternatives and model error. For a fixed feasible comparator u_i^0, write $B_i^0(T) = \sum_t \langle \ell_{i,t}, u_i^0\rangle$. More generally,

$$\begin{aligned}
\ell_{i,t}(k) &\ge \delta_i h_{i,t}(k) - \varepsilon_{i,t} - \zeta_{i,t}, \qquad \zeta_{i,t} \ge 0,\\
\overline{L}_T &\le \frac{1}{T}\sum_t \left[(1-p)C_{\mathrm{FP}} F_{\mathrm{b},t} + pC_{\mathrm{FN}} \sum_i \widetilde{\omega}_i \varepsilon_{i,t}\right]\\
&\quad + \frac{pC_{\mathrm{FN}}}{T}\sum_i \widetilde{\omega}_i \left[R_i(T) + B_i^0(T) + \sum_t \zeta_{i,t}\right].
\end{aligned} \tag{51}$$

This follows from $\sum_t \langle \ell_{i,t}, \pi_{i,t}\rangle \le R_i(T) + B_i^0(T)$. The term $B_i^0(T)$ retains the cost of an available reference behavior; $\zeta_{i,t}$ retains deviations from the modeled incentive inequality. For changing window laws or batch-dependent decisions, use the conditional routing probabilities in $F_{\mathrm{b},t}$ and $h_{i,t}$. The same algebra then applies with the corresponding losses and incentive gaps.

B.10 RVU BOUNDS FOR IMPLEMENTED ACTOR–CRITIC POLICIES

The mirror analysis can be applied to the policies actually produced by a user learner. We reconstruct its update in probability space and charge the discrepancy from a predictive mirror step. This yields a finite-time regret bound for decoded neural policies, including policies with zero-probability modes, and a corresponding governance bound.

From network outputs to mode probabilities. Let $\pi_{i,t}$ be the conditional distribution over executable modes induced by the learner and its decoder, as in (63). For a fixed analysis parameter $0 \le \tau < 1$, set $\bar{\pi}_{i,t} = (1-\tau)\pi_{i,t} + \tau \mathbf{1}/K_i$. Choose $\tau > 0$ when the decoder excludes modes; otherwise $\tau = 0$ is allowed when all probabilities are positive. This smoothing is used only in the proof; the deployed policy remains $\pi_{i,t}$. It costs at most $2\tau T$ in regret.

For a vector x, define $\mathrm{sp}(x) = (\max_k x_k - \min_k x_k)/2$. Fix $\eta_i > 0$ and pre-action predictions $\|v_{i,t}\|_\infty \le 1$ with $v_{i,1} = 0$. The observed policy sequence defines

$$\begin{aligned} g_{i,t} &= v_{i,t} - v_{i,t+1} - \ell_{i,t} - \frac{1}{\eta_i}\log\frac{\bar\pi_{i,t+1}}{\bar\pi_{i,t}}, \quad c_{i,t} = \mathrm{sp}(g_{i,t}), \\ \mathcal{D}_i(T) &= 2\sum_{t=1}^{T}(c_{i,t} + \eta_i c_{i,t}^2) + 2\tau T, \qquad\qquad D_i^0 = \log\frac{1}{\min_k \bar\pi_{i,1}(k)}. \end{aligned} \tag{52}$$

Logarithms and ratios are componentwise. The span removes the additive constants that leave a probability update unchanged. The quantity $c_{i,t}$ measures how far the actual change in mode probabilities departs from the predictive update on the modeled user losses. It includes the effect of changing observations and decoder state, as well as parameter updates.

Theorem 6 (RVU with an explicit policy-update error). *For any sequence of deployed mode mixtures, $\|\ell_{i,t}\|_\infty \le 1$, and the positive smoothed probabilities above, define $\mathcal{I}_i^{\mathrm{act}} = D_i^0/\eta_i + 2\eta_i$, $\alpha_i^{\mathrm{act}} = 2\eta_i$, and $\sigma_i^{\mathrm{act}} = (1-\tau)^2/(8\eta_i)$. Then the regret of the actual learner satisfies*

$$R_i(T) \le \mathcal{I}_i^{\mathrm{act}} + \alpha_i^{\mathrm{act}}\sum_{t=2}^{T}\|\ell_{i,t} - v_{i,t}\|_\infty^2 - \sigma_i^{\mathrm{act}} P_i(T) + \mathcal{D}_i(T). \tag{53}$$

Under the routing model and prediction timing of Theorem 3, if $2\sum_i \widetilde\omega_i \alpha_i^{\mathrm{act}} B_{ij} \le \widetilde\omega_j \sigma_j^{\mathrm{act}}$ for every j, then

$$\overline{L}_T \le U^\star + \frac{Q_T}{T} + \frac{pC_{\mathrm{FN}}}{T}\sum_i \widetilde\omega_i\left[\mathcal{I}_i^{\mathrm{act}} + 2\alpha_i^{\mathrm{act}}\mathcal{E}_i(T) + \mathcal{D}_i(T)\right]. \tag{54}$$

Proof. Reconstruct a mirror trajectory from the actual policy sequence. Suppress i and let $b_t = (\max_k g_t(k) + \min_k g_t(k))/2$, $e_t = g_t - b_t\mathbf{1}$. Then $\|e_t\|_\infty = c_t$. Define $z_t(k) \propto \bar\pi_t(k)\exp(\eta v_t(k))$. The identity in (52) gives

$$\bar\pi_t(k) \propto z_t(k)e^{-\eta v_t(k)}, \qquad z_{t+1}(k) \propto z_t(k)e^{-\eta[\ell_t(k)+e_t(k)]}. \tag{55}$$

These are exactly the entropic mirror updates for losses $\ell_t + e_t$. The auxiliary z_t is constructed for analysis; no mirror step is added to the experimental learner. The three-point proof in Appendix B.2 applies to any finite loss vector, so for the regret $\widehat{R}(T)$ of $\bar\pi_t$ on $\ell_t + e_t$,

$$\widehat{R}(T) \le \frac{D^0}{\eta} + \eta\sum_t \|\ell_t + e_t - v_t\|_\infty^2 - \frac{P_{\bar\pi}(T)}{8\eta}.$$

Transfer regret and variation back to the deployed policy. Let $R_{\bar\pi}(T)$ use losses ℓ_t. For every fixed comparator u, $|\langle e_t, \bar\pi_t - u\rangle| \le 2c_t$, while $\|\pi_t - \bar\pi_t\|_1 \le 2\tau$. Hence

$$\begin{aligned} R_i(T) &\le R_{\bar\pi}(T) + 2\tau T \le \widehat{R}(T) + 2\sum_t c_t + 2\tau T, \\ P_{\bar\pi}(T) &= (1-\tau)^2 P_i(T), \\ \|\ell_t + e_t - v_t\|_\infty^2 &\le 2\|\ell_t - v_t\|_\infty^2 + 2c_t^2. \end{aligned}$$

Use $v_1 = 0$ and $\|\ell_1\|_\infty \le 1$ to obtain (53). Substituting it into the variation cancellation in (34) retains $\sum_i \widetilde\omega_i \mathcal{D}_i(T)$; Theorem 1 gives (54). □

Exact and approximate RVU learning. Equation (53) is an RVU bound with a specified implementation error, derived for the deployed policy sequence. If $\tau = 0$ and $c_{i,t} = 0$, it recovers a predictive RVU learner. More generally, uniform bounds on D_i^0 and $\mathcal{D}_i(T)$ over admissible runs and horizons give a fixed initialization constant, yielding Assumption 2. For general updates, the finite-time governance bound retains $\mathcal{D}_i(T)$. With fixed constants, $Q_T = o(T)$, $\mathcal{E}_i(T) = o(T)$, and $\sum_t (c_{i,t} + c_{i,t}^2) = o(T)$, it gives

$$\limsup_T \overline{L}_T \le U^\star + 2\tau p C_{\mathrm{FN}}\sum_i \widetilde\omega_i.$$

The last term is the explicit cost of comparing with modes outside the learner's support through analysis smoothing.

Absorbing relative update error. Uniformly bounded cumulative error is one sufficient case. A broader condition allows implementation error to grow with prediction error and strategy variation. Write $V_i(T) = \sum_{t=2}^{T} \|\ell_{i,t} - v_{i,t}\|_\infty^2$. Use a uniform bound on D_i^0 in $\mathcal{I}_i^{\mathrm{act}}$, and suppose fixed nonnegative constants d_i, κ_i, ρ_i, uniform over admissible runs and horizons, satisfy

$$\mathcal{D}_i(T) \le d_i + \kappa_i V_i(T) + \rho_i P_i(T), \qquad \rho_i < \sigma_i^{\mathrm{act}}. \tag{56}$$

Substitution into (53) gives

$$R_i(T) \le (\mathcal{I}_i^{\mathrm{act}} + d_i) + (\alpha_i^{\mathrm{act}} + \kappa_i) V_i(T) - (\sigma_i^{\mathrm{act}} - \rho_i) P_i(T).$$

Thus Assumption 2 holds with $\mathcal{I}_i = \mathcal{I}_i^{\mathrm{act}} + d_i$, $\alpha_i = \alpha_i^{\mathrm{act}} + \kappa_i$, and $\sigma_i = \sigma_i^{\mathrm{act}} - \rho_i > 0$. Theorems 2–3 apply when their prediction and coupling conditions hold with these adjusted constants. In particular, adaptive governance requires $2\sum_i \widetilde{\omega}_i(\alpha_i^{\mathrm{act}} + \kappa_i) B_{ij} \le \widetilde{\omega}_j(\sigma_j^{\mathrm{act}} - \rho_j)$ for every j. This expresses the available stability margin: implementation error may consume part of the negative variation term while the remainder absorbs interactions among learners.

Relating the error to a neural update. Write $F_{i,o}(w)$ for the smoothed mode probabilities obtained from parameters w at information state o, including the decoder's state. Let $\Delta w_{i,t} = w_{i,t+1} - w_{i,t}$ be the actual optimizer increment and $J_{i,t} = \nabla_w \log F_{i,o_{i,t}}(w_{i,t})$. Suppose each component of $\log F_{i,o_{i,t}}$ has Hessian norm at most $H_{i,t}$ along this increment. Taylor expansion gives

$$\begin{aligned} \log \frac{\bar{\pi}_{i,t+1}}{\bar{\pi}_{i,t}} &= J_{i,t}\Delta w_{i,t} + a_{i,t} + r_{i,t}^{\mathrm{net}}, & \|r_{i,t}^{\mathrm{net}}\|_\infty &\le \tfrac{1}{2} H_{i,t} \|\Delta w_{i,t}\|_2^2, \\ a_{i,t} &= \log F_{i,o_{i,t+1}}(w_{i,t+1}) - \log F_{i,o_{i,t}}(w_{i,t+1}). \end{aligned} \tag{57}$$

Here $a_{i,t}$ isolates the change in observations and decoder state. For a mode-loss estimate $\widehat{\ell}_{i,t}$ derived from the learner's feedback, subadditivity of sp yields

$$\begin{aligned} c_{i,t} \le\ & \mathrm{sp}\left(v_{i,t} - v_{i,t+1} - \widehat{\ell}_{i,t} - \frac{J_{i,t}\Delta w_{i,t}}{\eta_i} \right) + \mathrm{sp}(\widehat{\ell}_{i,t} - \ell_{i,t}) \\ & + \frac{\mathrm{sp}(a_{i,t})}{\eta_i} + \frac{H_{i,t}}{2\eta_i} \|\Delta w_{i,t}\|_2^2. \end{aligned} \tag{58}$$

The four terms quantify the implemented update direction, feedback error, information-state change, and nonlinear parameterization. An advantage baseline common to all modes disappears under the span. Differences between discounted or shaped returns and the modeled window loss remain in the feedback term. This calculation uses the actual increment, so multiple optimizer steps, entropy regularization, clipping, and shared actor–critic parameters are included in $\Delta w_{i,t}$.

Proximal geometry and PPO. The predictive update has an exact proximal interpretation. Let $d_{i,t}^{\mathrm{prox}} = v_{i,t} - v_{i,t+1} - \ell_{i,t}$. For positive $\bar{\pi}_{i,t}$, the KL-regularized local step is

$$\pi_{i,t}^{+} = \arg\max_{x \in \Delta_i} \left\{ \langle d_{i,t}^{\mathrm{prox}}, x \rangle - \frac{1}{\eta_i} \mathrm{KL}(x \| \bar{\pi}_{i,t}) \right\}, \qquad \pi_{i,t}^{+}(k) \propto \bar{\pi}_{i,t}(k) e^{\eta_i d_{i,t}^{\mathrm{prox}}(k)}. \tag{59}$$

The first-order condition makes $d_{i,t}^{\mathrm{prox}} - \eta_i^{-1} \log(\pi_{i,t}^{+}/\bar{\pi}_{i,t})$ constant across modes. Hence $\bar{\pi}_{i,t+1} = \pi_{i,t}^{+}$ gives $c_{i,t} = 0$. The negative variation term comes from the curvature of the KL penalty; its coefficient $\sigma_i^{\mathrm{act}} = (1-\tau)^2/(8\eta_i)$ increases as the proximal stepsize decreases. The direction $d_{i,t}^{\mathrm{prox}}$ includes the predictive correction $v_{i,t} - v_{i,t+1}$ as well as the utility $-\ell_{i,t}$.

PPO pursues conservative return improvement through a clipped on-policy surrogate (Schulman et al., 2017). For a sampled observation–action pair (o, a), let $\varrho(w) = \mu_w(a \mid o)/\mu_{w_{\mathrm{old}}}(a \mid o)$, where μ_w is the behavior-control policy. With estimated advantage $\widehat{A}$ and clip parameter $\epsilon_{\mathrm{PPO}} > 0$, the actor objective is

$$\widehat{\mathbb{E}}\Big[\min\Big\{ \varrho(w)\widehat{A}, \mathrm{clip}(\varrho(w), 1 - \epsilon_{\mathrm{PPO}}, 1 + \epsilon_{\mathrm{PPO}})\widehat{A} \Big\} \Big]. \tag{60}$$

Near the rollout policy, this agrees with the usual return-improvement surrogate; clipping removes additional surrogate gain from sufficiently large changes in the advantage-favored direction. It modifies the optimization objective rather than imposing a hard constraint on every action probability. The connection to RVU is the combination of a utility-directed update and proximal stability. Equations (52) and (58) quantify the predictive, feedback, optimization, and decoder contributions for the actual policy sequence. Condition (56) specifies when these contributions preserve standard RVU through adjusted constants.

Application to the experimental learners. The Mahjong user uses clipped PPO with GAE and Adam, followed by a decoder into executable behavior modes. The non-competitive users use actor–critic updates of camouflage policies. These implementations supply the mode probabilities and feedback for (52) and (58). Their governance bound follows from (54); relative errors satisfying (56) are absorbed into the RVU constants. This condition concerns the induced learning trajectory, while the PPO clip parameter specifies its training surrogate. For continuous controls, representation and within-mode approximation errors enter (51), together with any payoff or routing-model discrepancy. This connects the implemented conservative adaptation to the regret analysis and, through it, to the platform's governance objective.

B.11 Controller and user-learner realization

Under the routing model, each controller is evaluated through its generated thresholds and user-mode mixtures:

$$(\mathcal{A}, \boldsymbol{\mathcal{B}}) \longmapsto (\theta_t, \pi_t)_{t=1}^{T} \longmapsto \overline{L}_T^{\mathcal{A},\boldsymbol{\mathcal{B}}} \le U^\star + \frac{Q_T^{\mathcal{A},\boldsymbol{\mathcal{B}}}}{T} + \frac{pC_{\mathrm{FN}}}{T} \sum_i \widetilde{\omega}_i R_i^{\mathcal{A},\boldsymbol{\mathcal{B}}}(T). \tag{61}$$

Platform implementations. Let o_t contain the observed evidence summaries and current threshold, and let $\widetilde{o}_t$ append the configured bias coordinate. Write Z_t for the simulator state, including population counts, user behavior, and actuator memory. The Mahjong controllers produce

$\mathcal{A}$	threshold selection
PPO	$A_t \sim \pi_{w_n}^{\mathrm{plat}}(\cdot \mid o_t), \quad \theta_t = g_{\mathrm{PPO}}(Z_t, A_t)$
Linear	$\theta_t = \mathsf{L}_{Z_t}(w_t^\top \widetilde{o}_t + \epsilon_t), \quad \epsilon_t \sim \mathcal{N}(0, \sigma_{\mathrm{lin}}^2)$
EXP3	$J_t \sim p_t, \quad \theta_t = \vartheta_{J_t}, \quad \{\vartheta_j\}_{j=1}^{K_{\mathrm{arm}}} \subset \Theta$

(62)

Here g_{PPO} is the action decoder and L_{Z_t} applies clipping, step limits, and smoothing. EXP3 updates arm probabilities from the reward of the selected threshold; routing uses the same reference-policy evidence. The PPO and Linear controllers also condition their actions on evidence summaries. Rollout targets and training settings are specified in Appendix C.16.

User implementations. Let $o_{i,t}$ be the information available to a behavior-control policy, including decoder state. Its action distribution $\mu_{i,t}(\cdot \mid o_{i,t})$ and decoder d_i induce

$$\pi_{i,t}(k) = \int \mathbf{1}\{\mathsf{d}_i(a; o_{i,t}) = k\}\, \mu_{i,t}(da \mid o_{i,t}), \qquad \pi_{i,t} \in \Delta_i. \tag{63}$$

In Mahjong, the morphology policy and KP decoder select executable behavior modes. The integral above gives their selection probabilities, which are the mode mixtures used in the analysis. The index i identifies a decision-making learner: a shared policy receives feedback from the interactions it governs, while independently trained policies define separate learners. Given the model's user losses, these mixtures determine $R_i(T)$ in Theorem 1. The simulator supplies rank, survival, and stability feedback to the learned behavior policy. Theorem 6 derives a predictive RVU bound directly from these induced mode probabilities, with an explicit policy-update error. Equation (58) connects that error to the actor–critic increment and feedback; (12) is the exact RVU realization.

Available behavior and decision timing. The Mahjong decoder selects among 21 active behavior modes. The zero-loss option belongs to the theoretical routing model. For a decoder restricted to active modes, a fixed available mode supplies the comparator in (51), with its cumulative cost retained as $B_i^0(T)$.

In the theoretical protocol, the threshold precedes the windows used to evaluate routing; Theorem 3 also makes it visible before user action. The Mahjong rollout instead computes evidence summaries within a batch, selects a threshold for that batch, and applies the next user-mode choice to subsequent simulation. Appendix C.16 gives this order explicitly. Batch-dependent routing uses conditional routing rates in (51). The current-rule predictor in Section 4.3 specifies the timing used for the adaptive RVU rate.

Routing loss and population outcomes. We retain the governance loss L_t from (4). To express it in routing rates, define $q_t^{\mathrm{b}} = F_{\mathrm{b}}(\theta_t)$, $q_t^{\mathrm{h}} = \sum_i \omega_i \langle r_i(\theta_t), \pi_{i,t} \rangle$, and $q_t^{\mathrm{safe}} = \sum_i \omega_i \pi_{i,t}(0)$. These are the model's benign false-routing rate, strategic routing rate, and safe-mode share, respectively. Since $h_{i,0} = r_{i,0} = 0$ and $h_{i,k} = 1 - r_{i,k}$ for $k \geq 1$, the same loss is

$$L_t = (1-p)C_{\mathrm{FP}} q_t^{\mathrm{b}} + pC_{\mathrm{FN}}\big(1 - q_t^{\mathrm{safe}} - q_t^{\mathrm{h}}\big). \tag{64}$$

Thus harmful exposure excludes both routed behavior and the safe alternative. The expression $1 - q_t^{\mathrm{h}}$ is valid only when $q_t^{\mathrm{safe}} = 0$. The coefficients $(1-p)C_{\mathrm{FP}}$ and pC_{FN} retain the population shares and error costs from Preliminaries. Theorem 1 bounds the time average of this loss through $\overline{U}_T$ and cumulative user regret.

Experimental comparisons. For observation choice S and controller $\mathcal{A}$, let $J_{\mathrm{exp}}(S, \mathcal{A})$ denote an evaluated platform cost. The experimental design separates controller and evidence effects:

$$\begin{aligned} &\{J_{\mathrm{exp}}(\mathrm{Ev}, \mathcal{A}) : \mathcal{A} \in \{\mathrm{PPO}, \mathrm{Linear}, \mathrm{EXP3}\}\}, \\ &\Delta_{\mathrm{evidence}}(\mathcal{A}) = J_{\mathrm{exp}}(\mathrm{Rank}, \mathcal{A}) - J_{\mathrm{exp}}(\mathrm{Ev}, \mathcal{A}), \quad \mathcal{A} \in \{\mathrm{PPO}, \mathrm{Linear}\}. \end{aligned} \tag{65}$$

The first comparison varies the controller under shared evidence; the second fixes the controller family and varies its observation signal.

The experiments examine these two design choices through service outcomes. Harmful exposure in the model is weighted over the fixed potential population. Mahjong harmful-user share is measured within the current novice pool, whose composition changes through routing, arrival, return, and exit. Benign-user ranks measure the service consequences of these changes; the non-competitive simulator uses penetration and welfare. The shaped loss $\widehat{\ell}_t^{\mathrm{plat}}$ supplies the controller with feedback on routing, population occupancy, benign experience, and threshold movement. Its role is to generate control decisions; the incentive conversion explains how a threshold trajectory and user learning determine L_t under the routing model.

The score $e(\mathcal{W})$ measures action-policy compatibility, the behavior decoder selects executable modes, and $v_{i,t}$ predicts mode losses in the learning analysis. The fast bounds use Assumption 2; the algorithm-independent conversion uses regret on the specified user losses. For applications with a costly alternative or a different payoff model, (51) states the corresponding conversion explicitly.

B.12 Controller choice and feasible thresholds

The controller $\mathcal{A}$ selects thresholds, while the user learners $\boldsymbol{\mathcal{B}}$ generate the mode mixtures. Their joint trajectory determines two different quantities: Q_T is the platform regret on U (Definition 2), and $R_i(T)$ is user regret on the mode losses (Definition 1). PPO, Linear, EXP3, Strategy-Aware, and Performative Gradient supply different threshold-selection rules. Their observations, updates, and action decoders determine the resulting trajectories. The incentive conversion in Theorem 1 then applies under the routing model. The predictive RVU property in Assumption 2 concerns the user learners, regardless of which controller selects the platform thresholds.

For a controller whose thresholds lie in a nonempty $\Theta_{\mathcal{A}} \subseteq \Theta$, set

$$\begin{aligned} U_{\mathcal{A}}^\star &= \inf_{\theta \in \Theta_{\mathcal{A}}} U(\theta), & a_{\mathcal{A}} &= U_{\mathcal{A}}^\star - U^\star \geq 0, \\ Q_T^{\mathcal{A},\boldsymbol{\mathcal{B}}} &= \sum_t [U(\theta_t) - U_{\mathcal{A}}^\star] + T a_{\mathcal{A}} =: Q_{T,\mathcal{A}}^{\mathrm{within}} + T a_{\mathcal{A}}. \end{aligned} \tag{66}$$

Consequently, Theorem 1 gives

$$\overline{L}_T^{\mathcal{A},\boldsymbol{\mathcal{B}}} \leq U^\star + a_{\mathcal{A}} + \frac{Q_{T,\mathcal{A}}^{\mathrm{within}}}{T} + \frac{pC_{\mathrm{FN}}}{T} \sum_i \widetilde{\omega}_i R_i^{\mathcal{A},\boldsymbol{\mathcal{B}}}(T). \tag{67}$$

The same U evaluates continuous-threshold controllers and finite-arm selectors. Domain approximation and within-domain selection appear as separate terms.

Grid approximation. If F_{b} is M_{b}-Lipschitz and $r_{i,k}$ is M_i-Lipschitz, then

$$|U(\theta) - U(\theta')| \le L_U|\theta - \theta'|,$$
$$L_U = (1-p)C_{\mathrm{FP}}M_{\mathrm{b}} + pC_{\mathrm{FN}}\sum_i \widetilde{\omega}_i(\lambda_i + \delta_i)M_i. \tag{68}$$

Indeed, $|\varepsilon_i(\theta) - \varepsilon_i(\theta')| \le (\lambda_i + \delta_i)M_i|\theta - \theta'|$ by the maximum inequality. For a grid $\Theta_{\mathcal{A}}$ covering Θ within distance h,

$$\forall\theta \in \Theta\ \exists\vartheta \in \Theta_{\mathcal{A}} : |\theta - \vartheta| \le h \quad\Longrightarrow\quad 0 \le a_{\mathcal{A}} \le L_U h.$$

Apply the Lipschitz bound to an arbitrary minimizing sequence of U and take its limit.

Selection through estimated guarantees. Suppose a selector satisfies, for functions $\widehat{U}_t$ on $\Theta_{\mathcal{A}}$,

$$\sup_{\theta\in\Theta_{\mathcal{A}}} |\widehat{U}_t(\theta) - U(\theta)| \le \zeta_t, \qquad \sum_t \widehat{U}_t(\theta_t) - \inf_{\theta\in\Theta_{\mathcal{A}}}\sum_t \widehat{U}_t(\theta) \le S_T. \tag{69}$$

For every $u \in \Theta_{\mathcal{A}}$,

$$\sum_t [U(\theta_t) - U(u)] \le \sum_t [\widehat{U}_t(\theta_t) - \widehat{U}_t(u)] + 2\sum_t \zeta_t \le S_T + 2\sum_t \zeta_t.$$

Taking the supremum over u yields

$$Q_T \le S_T + 2\sum_t \zeta_t + Ta_{\mathcal{A}}. \tag{70}$$

Here $\widehat{U}_t$ estimates the routing guarantee (8). This calculation identifies the quantity that a threshold-selection guarantee must control.

Relating controller feedback to the routing guarantee. The same argument accommodates a controller trained on another loss scale. Let $g_t(\theta)$ be its round-t loss function on $\Theta_{\mathcal{A}}$, with lower values preferred, and suppose a valid selection bound for this loss is

$$\sum_t g_t(\theta_t) - \inf_{\theta\in\Theta_{\mathcal{A}}}\sum_t g_t(\theta) \le S_T^g. \tag{71}$$

For a fixed scale $a > 0$ and offsets b_t independent of θ, suppose $\sup_{\theta\in\Theta_{\mathcal{A}}} |g_t(\theta) - aU(\theta) - b_t| \le d_t$. For every fixed $u \in \Theta_{\mathcal{A}}$, the offsets cancel and

$$a\sum_t [U(\theta_t) - U(u)] \le \sum_t [g_t(\theta_t) - g_t(u)] + 2\sum_t d_t \le S_T^g + 2\sum_t d_t.$$

Taking the supremum over u and adding the domain gap gives

$$Q_T \le \frac{S_T^g + 2\sum_t d_t}{a} + Ta_{\mathcal{A}}. \tag{72}$$

Equation (70) is the case $a = 1$, $b_t = 0$, and $g_t = \widehat{U}_t$. The quantities d_t account for the difference between the controller's feedback objective and the routing guarantee. In the experiments, the shaped loss in (111) supplies feedback on routing, occupancy, benign experience, and threshold movement. Any use of (72) with that loss requires the corresponding feedback-selection and discrepancy bounds. Thus a controller rate combines its selection bound with the discrepancy between its feedback loss and U.

Shared guarantees and algorithm-dependent rates. For learners satisfying Theorem 3,

$$\overline{L}_T^{\mathcal{A},\boldsymbol{\mathcal{B}}} \le U^\star + \frac{Q_T^{\mathcal{A},\boldsymbol{\mathcal{B}}}}{T} + \frac{pC_{\mathrm{FN}}}{T}\left[A_0 + 2\sum_i \widetilde{\omega}_i\alpha_i\mathcal{E}_i^{\mathcal{A},\boldsymbol{\mathcal{B}}}(T)\right]. \tag{73}$$

For fixed model and learning constants and $0 \le q, e < 1$,

$$\begin{aligned} &Q_T^{\mathcal{A},\boldsymbol{\mathcal{B}}} = O(T^q), \quad \max_i \mathcal{E}_i^{\mathcal{A},\boldsymbol{\mathcal{B}}}(T) = O(T^e) \\ \Longrightarrow\qquad &\overline{L}_T^{\mathcal{A},\boldsymbol{\mathcal{B}}} - U^\star \le O(T^{q-1} + T^{e-1} + T^{-1}). \end{aligned} \tag{74}$$

Each controller is evaluated along its own induced user-learning trajectory. More generally, if $R_i(T) \le \mathcal{R}_i(T)$ and the feedback-selection conditions above hold, substitution into (67) gives

$$\overline{L}_T \le U^\star + a_{\mathcal{A}} + \frac{S_T^g + 2\sum_t d_t}{aT} + \frac{pC_{\mathrm{FN}}}{T} \sum_i \widetilde{\omega}_i \mathcal{R}_i(T). \tag{75}$$

Under the predictive RVU and coupling conditions of Theorem 3, its weighted regret bound sharpens this to

$$\begin{aligned} \overline{L}_T \le U^\star + a_{\mathcal{A}} + \frac{S_T^g + 2\sum_t d_t}{aT} \\ + \frac{pC_{\mathrm{FN}}}{T} \left[A_0 + 2 \sum_i \widetilde{\omega}_i \alpha_i \mathcal{E}_i(T) \right]. \end{aligned} \tag{76}$$

These formulas separate four sources of cost: the feasible threshold domain, controller selection, feedback-objective discrepancy, and user learning. For fixed constants, $S_T^g = o(T)$, $\sum_t d_t = o(T)$, and $\mathcal{E}_i(T) = o(T)$ imply $\limsup_T \overline{L}_T \le U_{\mathcal{A}}^\star$. The derivation uses the selected thresholds and induced user losses, and imposes no particular parameterization of the controller. Controller-specific rates enter through valid bounds for its actual update and feedback protocol. Exact and inexact user-learning realizations are supplied by (12) and Theorem 5. Thus the same analysis organizes all controller implementations while retaining their distinct learning dynamics.

B.13 Comparison with independent deployment

The routing guarantee and independent deployment evaluate different objects. Here we define the deployment comparison, derive its relation to the realized-response benchmark, and specialize the difference to the governance model.

Two response trajectories. Let Z_t collect the response and population variables used to evaluate the cost of the adaptive run at round t, and let Z_t^ϑ denote their values when threshold ϑ is deployed throughout a separate run with the same user learners and initial law. For a common round-cost function c_t, define $f_t(\vartheta) = c_t(\vartheta, Z_t)$, $h_t(\vartheta) = c_t(\vartheta, Z_t^\vartheta)$, and $C_t = c_t(\theta_t, Z_t) = f_t(\theta_t)$. Thus f_t changes the candidate rule while holding the adaptive run's response fixed, whereas h_t includes the response induced by deploying that candidate throughout. We first work pathwise under a common coupling of initial conditions and environmental randomness, with finite comparison values. For expected-cost comparisons, replace C_t, $f_t(\vartheta)$, and $h_t(\vartheta)$ by their expectations, with the candidate ϑ fixed inside each expectation and the comparator infima taken afterwards.

For a fixed comparator ϑ, write $R_T^{\mathrm{real}}(\vartheta) = \sum_t [C_t - f_t(\vartheta)]$ and $R_T^{\mathrm{cf}}(\vartheta) = \sum_t [C_t - h_t(\vartheta)]$. Adding and subtracting the cost on the realized response gives the exact decomposition

$$R_T^{\mathrm{cf}}(\vartheta) = R_T^{\mathrm{real}}(\vartheta) + \sum_t [f_t(\vartheta) - h_t(\vartheta)]. \tag{77}$$

The second term measures the change in cost caused by the different response trajectory at the same comparator threshold.

Comparison with the best fixed deployment. Taking the best comparator separately in each criterion gives

$$R_T^{\mathrm{real}} = \sum_t C_t - \inf_{\vartheta \in \Theta} \sum_t f_t(\vartheta), \qquad R_T^{\mathrm{cf}} = \sum_t C_t - \inf_{\vartheta \in \Theta} \sum_t h_t(\vartheta),$$

$$R_T^{\mathrm{cf}} - R_T^{\mathrm{real}} = \inf_{\vartheta \in \Theta} \sum_t f_t(\vartheta) - \inf_{\vartheta \in \Theta} \sum_t h_t(\vartheta), \tag{78}$$

$$|R_T^{\mathrm{cf}} - R_T^{\mathrm{real}}| \le \sup_{\vartheta \in \Theta} \left| \sum_t [f_t(\vartheta) - h_t(\vartheta)] \right| =: \Xi_T. \tag{79}$$

To verify the inequality without assuming a minimizing threshold exists, set $F_T(\vartheta) = \sum_t f_t(\vartheta)$ and $H_T(\vartheta) = \sum_t h_t(\vartheta)$. When $\Xi_T < \infty$, the pointwise bounds $H_T - \Xi_T \le F_T \le H_T + \Xi_T$ imply $\inf H_T - \Xi_T \le \inf F_T \le \inf H_T + \Xi_T$. The bound is immediate if $\Xi_T = \infty$.

Explicit response discrepancy for routing. For the governance model in (4), take $Z_t = \pi_t$, $Z_t^\vartheta = \pi_t^\vartheta$, and

$$c(\vartheta, \pi) = (1-p)C_{\mathrm{FP}}F_{\mathrm{b}}(\vartheta) + pC_{\mathrm{FN}}\sum_i \omega_i \langle h_i(\vartheta), \pi_i\rangle.$$

The mode laws, population weights, and cost parameters are the same in both deployments. Evaluating both responses at ϑ cancels the benign false-routing term, so

$$f_t(\vartheta) - h_t(\vartheta) = pC_{\mathrm{FN}}\sum_i \omega_i \langle h_i(\vartheta), \pi_{i,t} - \pi_{i,t}^\vartheta\rangle. \tag{80}$$

Here h_t denotes the deployment cost, while the vector h_i denotes modewise harmful exposure. Since $h_i(\vartheta) \in [0,1]^{K_i}$ and both mixtures sum to one,

$$\left|\langle h_i(\vartheta), \pi_{i,t} - \pi_{i,t}^\vartheta\rangle\right| = \left|\langle h_i(\vartheta) - \tfrac{1}{2}\mathbf{1}, \pi_{i,t} - \pi_{i,t}^\vartheta\rangle\right| \le \tfrac{1}{2}\|\pi_{i,t} - \pi_{i,t}^\vartheta\|_1.$$

Define the cumulative response discrepancy $D_\Pi(T) = \sup_{\vartheta\in\Theta}\sum_{t=1}^T\sum_i \omega_i\|\pi_{i,t} - \pi_{i,t}^\vartheta\|_1$. The preceding inequalities yield

$$\Xi_T \le \frac{pC_{\mathrm{FN}}}{2}D_\Pi(T), \qquad R_T^{\mathrm{cf}} \le R_T^{\mathrm{real}} + \frac{pC_{\mathrm{FN}}}{2}D_\Pi(T). \tag{81}$$

Consequently, a bound $R_T^{\mathrm{real}} \le B_T$ and $D_\Pi(T) = o(T)$ give $R_T^{\mathrm{cf}}/T \le B_T/T + o(1)$. This identifies the additional response quantity needed to transfer a realized-response rate to independently deployed rules. For expected-cost comparisons, the same bound uses $\sup_\vartheta \sum_{t,i}\omega_i\mathbb{E}\|\pi_{i,t} - \pi_{i,t}^\vartheta\|_1$ under a common coupling. For population-dependent simulator costs, equations (77)–(79) retain the full state Z_t, including population changes.

Relation to the main guarantee and experiments. The platform regret Q_T in Definition 2 compares values of the common routing guarantee U, with $U^\star = \inf_{\theta\in\Theta}U(\theta)$. This is a benchmark for the loss bound. The independently deployed comparator instead minimizes $\sum_t h_t(\vartheta)$, the actual cost generated by each rule and its response. Theorems 1–3 bound governance loss along each induced trajectory; equation (81) concerns the distinct comparison between deployments. In particular, Q_T and R_T^{real} have different loss functions and are not interchangeable. Appendix B.14 gives a realized-response optimization benchmark under its stated convexity and update conditions.

The grouped routing experiments in Appendix C.17 run fixed and adaptive rules separately. Their measured outcomes therefore include the user adaptation and population dynamics generated by each deployment. Together, the theory and experiments connect the incentive mechanism that controls governance loss with the service outcomes obtained by deploying the complete learning loop.

B.14 The realized-response convex benchmark

The fixed-state diagnostic evaluates candidate thresholds on a common realized response, as in Appendix B.13.

Theorem 7 (Projected-gradient benchmark). *Let Θ be nonempty, compact, and convex, with Euclidean diameter D. Suppose $f_t : \Theta \to \mathbb{R}$ is convex and $g_t \in \partial f_t(\theta_t)$ satisfies $\|g_t\|_2 \le G$. Then*

$$\theta_{t+1} = \mathrm{proj}_\Theta(\theta_t - \eta g_t), \qquad \eta > 0, \tag{82}$$

$$R_T^{\mathrm{real}} \le \frac{D^2}{2\eta} + \frac{\eta G^2 T}{2}. \tag{83}$$

For $D, G > 0$, $\eta = D/(G\sqrt{T})$ gives $R_T^{\mathrm{real}} \le DG\sqrt{T}$.

Proof. For every $u \in \Theta$, projection nonexpansiveness and convexity give

$$\begin{aligned} f_t(\theta_t) - f_t(u) &\le \langle g_t, \theta_t - u\rangle \\ &\le \frac{\|\theta_t - u\|_2^2 - \|\theta_{t+1} - u\|_2^2}{2\eta} + \frac{\eta G^2}{2}, \\ \sum_t [f_t(\theta_t) - f_t(u)] &\le \frac{\|\theta_1 - u\|_2^2}{2\eta} + \frac{\eta G^2 T}{2} \le \frac{D^2}{2\eta} + \frac{\eta G^2 T}{2}. \end{aligned}$$

Maximize over u. □

B.15 Convex surrogates for clipped costs

On a compact convex active domain, suppose a convex surrogate has bounded subgradients and

$$\ell_t^{\mathrm{clip}}(\theta) \le \widetilde{\ell}_t(\theta) \le \ell_t^{\mathrm{clip}}(\theta) + \epsilon_t \quad (\forall \theta), \qquad \epsilon_t \ge 0. \tag{84}$$

Projected updates on $\widetilde{\ell}_t$ yield

$$\begin{aligned}\sum_t [\ell_t^{\mathrm{clip}}(\theta_t) - \ell_t^{\mathrm{clip}}(u)] &\le \sum_t [\widetilde{\ell}_t(\theta_t) - \widetilde{\ell}_t(u)] + \sum_t \epsilon_t \\ &\le DG\sqrt{T} + \sum_t \epsilon_t.\end{aligned}$$

The comparator ranges over the active domain; additional clipped thresholds are covered when they admit equivalent active-domain losses. Here clipping acts on the threshold cost.

C Detailed Riichi Mahjong Competitive Platform Experiment

This appendix gives the detailed construction of the Competitive Riichi Mahjong platform used in Section 5.1. The simulator instantiates the open-platform regulation loop in a competitive environment where genuine newcomers and adaptive harmful users interact over a long horizon. Riichi Mahjong is a four-player imperfect-information game with hidden hands, stochastic tile draws, risk management, opponent modeling, and long-horizon planning. These properties make it a useful testbed for regulation: final rank is informative but noisy, while policy traces contain more stable evidence about whether a user behaves like a genuine novice.

The platform contains two matchmaking regions. The novice pool is the normal service region where benign users should play. The high-rank pool is the quarantine region used for routed users. Here b denotes a genuine novice, while h denotes an experienced player exploiting novice matches through a low-rank account. These labels describe the platform roles: the high-rank region is the appropriate destination for the latter users. The platform observes policy-evidence statistics from behavior windows, applies a threshold-based routing rule, and moves selected users from the novice pool to the high-rank region. The experiment asks whether an adaptive platform controller can preserve newcomer experience when harmful users also adapt their behavior to avoid routing. The construction follows the three components of Methods: reference policies score behavior, a shared threshold routes interactions, and routing-dependent returns train the user and platform policies. The comparisons vary evidence, threshold control, and user adaptation to identify which parts of this loop account for the governance outcomes.

C.1 Player model construction

The simulator uses 31 Riichi Mahjong policies produced by a two-stage training pipeline built on `mortal` (Equim, 2022). All models first undergo offline pretraining with Conservative Q-Learning on public Tenhou Phoenix Table records from 2023 to 2024, followed by online self-play reinforcement learning. We retain policies at different stages of this second phase to obtain novice, intermediate, and expert levels of play. The collection therefore shares a pretraining source while representing different amounts of subsequent reinforcement learning. Each retained model specifies an action distribution conditioned on the game state; its playing strength is assessed through the matchup protocol below.

Each game state s is encoded as a multi-channel one-dimensional tensor over the 34 tile types,

$$\mathbf{X}(s) \in \mathbb{R}^{C \times 34}. \tag{85}$$

The channels encode private hand tiles, melds, visible discards, game progress, and public state features. The action space contains 46 actions: 37 discard actions and 9 non-discard operations, including Chi, Pon, Daiminkan, Kakan, Ankan, Riichi, Ron, Tsumo, and Pass. The policy backbone is a pre-activation one-dimensional ResNet with channel attention. After global aggregation, the network outputs a 1024-dimensional hidden representation used by value and rank-prediction heads.

We also train a frozen Global Rank Predictor (GRP). Given a game state s, GRP outputs a distribution over final ranks,

$$\mathbf{p}_{\mathrm{grp}}(s) = [p_1(s), p_2(s), p_3(s), p_4(s)]^\top. \tag{86}$$

With rank utility vector $\mathbf{u} = [3, 1, -1, -3]^\top$, the expected game potential is

$$\mathrm{EGP}(s) = \mathbf{p}_{\mathrm{grp}}(s)^\top \mathbf{u}. \tag{87}$$

During agent training, the dense local reward is the potential difference

$$r_t = \mathrm{EGP}(s_{t+1}) - \mathrm{EGP}(s_t), \tag{88}$$

which reduces the sparsity of final-rank supervision while remaining aligned with rank outcomes. In the platform-level baselines below, GRP is frozen and used only to produce a rank-predictive statistic.

C.2 Evaluation protocol and player bank

We evaluate trained checkpoints with a permutation-invariant protocol. Because seat position affects Riichi Mahjong outcomes, each model comparison is averaged over all $4!$ seat assignments. We then compute multiplayer Elo scores. The expected score of model i against model j is

$$\mathbb{E}_{ij} = \frac{1}{1 + 10^{(R_j - R_i)/c_{\mathrm{elo}}}}, \tag{89}$$

where c_{elo} is the Elo scale parameter listed in Table 8. The Elo values are iteratively updated until convergence.

The resulting Elo ordering assigns the identifiers R1–R31 and the strength tiers in Table 3. These identifiers describe evaluated playing strength; training stage describes how each policy was produced, and realized player rank describes placement in an individual game. Elo summarizes performance across matchups, while the evidence score evaluates action choices along a behavior window.

We select R3 as the high-rank reference policy, denoted π_{h}, and R29 as the low-rank reference policy, denoted π_{b}. π_{h} represents expert-like behavior and the latent ability of a harmful user. π_{b} represents novice-like behavior and is used to simulate benign users. Their evaluated playing strengths anchor the reference family to expert-like and novice-like behavior. We freeze this pair before computing policy evidence and running platform governance.

Fixed-reference evaluation across learned policies. Table 5 evaluates R4/R5, R4/R19, R4/R28, and R4/R30 with R3/R29 held fixed. None of the tested policies is a reference model, and we do not refit the references for any matchup. The pairs cover policies in the same Elo tier and in different tiers, testing whether the reference score preserves behavioral distinctions across policies obtained through reinforcement learning. Their common pretraining source specifies the shared task and training pipeline; their different subsequent training stages supply the policy variation being evaluated. In particular, the R4/R5 comparison tests behavior-source information among policies assigned to the same strength tier.

C.3 Basic player types

The platform experiment uses benign users and three harmful-user behavior configurations. A benign user is simulated by π_{b} or by policies sampled from the low-skill tier. Benign users are genuine newcomers who should remain in the novice pool; their target experience is evaluated by expected rank, with the fair four-player value denoted by R_{fair}.

An unrestricted harmful player is represented by π_{h} without behavioral restriction in the novice pool. This type represents an experienced player who makes no attempt to hide skill. It provides a sanity check: such users are easy to identify from rank-predictive statistics.

A camouflaged harmful player has high skill but uses hand-crafted constraints to suppress visible dominance. The controlled camouflage constraints are written as
C1: Riichi declarations are allowed only when fewer than T_{riichi} tiles remain in the wall.
C2: Ron calls are disabled before the endgame region.
C3: Red Dora tiles are retained in hand to disrupt tile efficiency and hand value.
These constraints reduce apparent rank dominance while preserving enough strategic structure for the player to remain stronger than a genuine novice.

An adaptive harmful user is used in the dynamic platform simulation. Its behavior-control policy chooses a low-dimensional morphology coordinate. This coordinate is mapped to an executable harmful-user behavior mode by the KP predictor described in Section C.10. Adaptive harmful users therefore search over a spectrum between novice-like and expert-like behavior.

Table 3: Elo rankings of 31 Riichi Mahjong policies retained at different reinforcement-learning stages after public-record pretraining. Rank identifiers and strength tiers are assigned from matchup performance. R3 and R29 serve as the fixed high-rank and low-rank references.

Tier	Rank	Elo
Tier 1 (≥ 1900)	1	2202.62
Tier 1 (≥ 1900)	2	2170.59
Tier 1 (≥ 1900)	3	2120.97
Tier 1 (≥ 1900)	4	2118.47
Tier 1 (≥ 1900)	5	2108.46
Tier 1 (≥ 1900)	6	2102.70
Tier 1 (≥ 1900)	7	2096.81
Tier 1 (≥ 1900)	8	2087.40
Tier 1 (≥ 1900)	9	2081.58
Tier 1 (≥ 1900)	10	2041.91
Tier 1 (≥ 1900)	11	2022.31
Tier 1 (≥ 1900)	12	2008.36
Tier 1 (≥ 1900)	13	2002.85
Tier 1 (≥ 1900)	14	1988.49
Tier 1 (≥ 1900)	15	1978.78
Tier 1 (≥ 1900)	16	1973.87
Tier 1 (≥ 1900)	17	1963.88
Tier 1 (≥ 1900)	18	1957.04
Tier 2 (1700–1899)	19	1894.76
Tier 2 (1700–1899)	20	1889.94
Tier 2 (1700–1899)	21	1843.25
Tier 2 (1700–1899)	22	1833.00
Tier 2 (1700–1899)	23	1822.71
Tier 2 (1700–1899)	24	1820.23
Tier 2 (1700–1899)	25	1799.45
Tier 2 (1700–1899)	26	1761.52
Tier 2 (1700–1899)	27	1732.66
Tier 2 (1700–1899)	28	1700.64
Tier 4 (< 1500)	29	1363.35
Tier 4 (< 1500)	30	1279.68
Tier 4 (< 1500)	31	100.00

C.4 Controlled rank-predictive detection experiments

Before running the full platform simulation, we evaluate two controlled player settings. In Scenario A, one unrestricted harmful player is matched with three benign users for N_{ctrl} simulated games. The unrestricted harmful player achieves average rank 1.68 and Top-2 rate 85.0%, while benign users in the same games have average rank 2.77 and Top-2 rate 39.16%. This setting shows that an unrestricted harmful user is easy to identify from rank-predictive statistics.

In Scenario B, one camouflaged harmful player is matched with three benign users under the same protocol. The camouflage constraints push the camouflaged harmful user's average rank to 2.275 and reduce its Top-2 rate to 62.50%. Benign users partially recover to average rank 2.575 and Top-2 rate 45.83%. From the perspective of expected rank, the camouflaged harmful player resembles a talented beginner. This motivates using policy-evidence statistics rather than relying only on rank-predictive signals.

C.5 Reference-policy evidence statistic

The Mahjong evidence score in (1) uses fixed high-rank and low-rank references, π_{h} and π_{b}. For a window $\mathcal{W} = ((s_\tau, a_\tau))_{\tau=1}^{m}$, compute

$$S_{\mathrm{h}}(\mathcal{W}) = \sum_{\tau=1}^{m} \log \pi_{\mathrm{h}}(a_\tau \mid s_\tau), \qquad S_{\mathrm{b}}(\mathcal{W}) = \sum_{\tau=1}^{m} \log \pi_{\mathrm{b}}(a_\tau \mid s_\tau). \tag{90}$$

Their difference is $e(\mathcal{W}) = S_{\mathrm{h}}(\mathcal{W}) - S_{\mathrm{b}}(\mathcal{W})$. A large positive score indicates greater compatibility with expert-like than novice-like behavior. Equation (22) gives this comparison its action-policy interpretation. Equation (23) relates the resulting routing probabilities to the incentive gap.

The platform compares $e(\mathcal{W})$ with θ_t to route users from the novice pool to the high-rank region. This removes harmful users' opportunity to exploit novice matches. Routing a benign user imposes a service cost. Here m denotes the length of a scored window, while t and T index platform learning rounds and their horizon.

C.6 Details of platform-level baselines

Information interfaces for baselines. We distinguish native baselines from hybrid diagnostic ablations. A native baseline is implemented with the feedback interface assumed by the corresponding method family. Strategy-Aware is a strategic-classification controller over reported features and labels; Performative Gradient updates a deployed parameter using samples from the distribution induced by that deployment; EXP3 treats each threshold or rule as a bandit arm and updates from realized arm reward. These methods do not define a reference-policy likelihood model or a within-game policy-compatibility statistic.

Accordingly, in the main native-baseline comparison, these methods receive an outcome-level expected-rank statistic produced by the Global Rank Predictor. This is the natural analogue of their original feedback models in the Mahjong platform: the regulator observes how the deployed rule affects rank-related outcomes and updates the rule accordingly. In contrast, our method constructs a policy-evidence statistic from the sequence of in-game decisions using the fixed high-rank and low-rank reference policies. This evidence layer is part of the proposed method.

For diagnostic purposes, one can combine a prior controller with the proposed evidence layer. We refer to these as hybrid ablations, e.g., Ev+EXP3 or Performative-Gradient+Evidence. Such variants test whether gains come mainly from the evidence representation, from the adaptive controller, or from their combination; they are not treated as native instantiations of the prior baselines.

The three platform-level baselines use the same frozen GRP statistic. GRP maps in-game behavior and state information to a predicted final-rank distribution and returns the expected global rank. This keeps the baselines competitive: they use a learned rank-predictive model of player performance, not raw win rate or a hand-crafted outcome rule. The baselines differ in how they update the routing rule from this rank-predictive statistic.

Our method uses the reference-policy evidence score $e(\mathcal{W})$ with PPO, Linear, or EXP3 threshold control. The Rank+PPO and Rank+Linear comparisons in Section 5.1 isolate the observation signal within matched controller families; the evidence-based controller comparison tests the benefit across update mechanisms.

C.7 Supplementary fixed-reference and sequence-transfer diagnostics

These diagnostics examine behavioral discrimination with fixed references and transfer of trained sequence-model baselines. The policy-source test compares scores on designated checkpoint pairs; the sequence-transfer test evaluates classifiers trained on Standard and frozen before testing on Extra. Each table reports its own evaluation protocol.

Policy-source discrimination with fixed references. We evaluate the fixed R3/R29 evidence score on four other policy pairs from the reinforcement-learning pipeline. Each pair excludes both reference models: the first policy supplies positive seats and the second negative seats. Evidence and realized player rank are evaluated on the same examples without reference refitting. Elo assigns each policy's

Table 4: Platform-level baselines in the Competitive Mahjong experiment. All three baselines use the same frozen GRP statistic and differ only in the update mechanism.

Baseline	**Regulation statistic**	**Update mechanism**
Strategy Aware	Expected rank from GRP	Updates the routing rule after accounting for player response to the deployed rule, following the strategic-classification view.
Performative Gradient	Expected rank from GRP	Treats the deployed routing rule as changing the next observed population distribution and updates on the induced distribution.
EXP3	Expected rank from GRP	Treats candidate routing rules as bandit arms and updates sampling weights using the realized platform reward of the selected rule.

strength tier; realized rank records its placement in an individual game and differs from the frozen GRP used in the controller comparisons.

Table 5: Policy-source AUROC with fixed R3/R29 references. Elo tiers: R4/R5 are Tier 1; R19/R28 are Tier 2; R30 is Tier 4.

Matchup	Realized rank	Evidence
R4/R5	0.489	0.713
R4/R19	0.505	0.688
R4/R28	0.577	0.989
R4/R30	0.691	1.000

Table 5 shows higher source-discrimination AUROC for Evidence in all four matchups. For the same-tier R4/R5 pair, realized-rank AUROC is 0.489, whereas Evidence reaches 0.713 with the reference pair unchanged. A common pretraining source and shared Elo tier can therefore coexist with distinguishable behavior: reinforcement learning produces policies whose action traces retain information beyond terminal placement.

The fixed references capture these distinctions across policies from the training pipeline, consistent with the policy-compatibility interpretation in (22). Together, Tables 2 and 5 establish two useful properties of Evidence: sensitivity to the targeted behavior and behavioral discrimination with unchanged references. The matched controller comparisons in Table 1 establish the resulting governance benefit when this signal drives routing.

Table 6: Transfer of sequence-model baselines from Standard to Extra. Each entry follows the metric order in its column; thresholds are fixed on Standard validation data.

Model	Standard AUROC/BAcc	Extra AUROC/BAcc	Extra FPR/FNR
Bag-of-actions + LR	0.874/0.722	0.626/0.592	0.393/0.423
Isolation Forest	0.549/0.504	0.475/0.497	0.049/0.957
RNN	0.945/0.853	0.711/0.638	0.508/0.215
Transformer	0.913/0.769	0.675/0.630	0.367/0.372

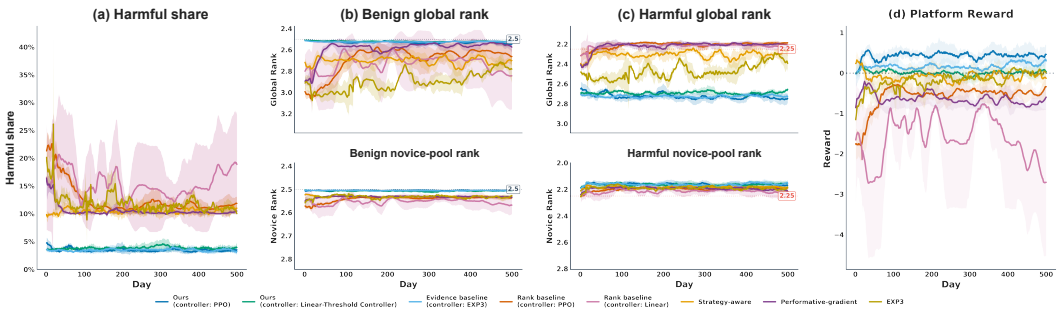


Figure 6: Additional Competitive Mahjong setting. Lines show cross-seed means and shaded bands show one standard deviation across seeds. Native rank-predictive baselines and controller-matched rank ablations fluctuate more strongly, while regulators operating on reference-policy evidence maintain lower harmful-user shares, more stable rank outcomes, and higher platform reward.

Sequence-model transfer. We fit the four sequence-model baselines in Table 6 once on the Standard suite, freeze their parameters and validation-selected thresholds, and evaluate them on the separate Extra suite. Bag-of-actions + LR uses logistic regression on action-count features. Table 6 reports area under the ROC curve (AUROC), balanced accuracy (BAcc), and false-positive/false-negative rates (FPR/FNR).

Table 6 tests whether a learned score and a fixed operating threshold transfer together. The RNN attains the highest Extra AUROC (0.711), but its frozen threshold yields FPR/FNR of $0.508/0.215$; Isolation Forest produces $0.049/0.957$. These operating points motivate separating behavioral evidence from adaptive threshold control. Our method fixes the reference score while adjusting the routing threshold to user responses. Its benefit is measured directly by the higher recall in Table 2, the fixed-reference discrimination in Table 5, and the lower harmful-user share under matched controllers in Table 1.

C.8 Additional Mahjong setting

We also evaluate the Mahjong regulator under another setting. The same agent pool, reference-policy evidence construction, reward terms, and controller implementations are used. This experiment tests whether the observed advantage of policy evidence persists beyond the main Mahjong setting reported in Section 5.1. Holding the evidence construction fixed while changing the population-flow setting tests whether its governance value survives a different induced population trajectory.

Figure 6 and Table 7 show the same qualitative pattern as the main experiment. Regulators using reference-policy evidence converge to low harmful-user shares and stable rank trajectories, while outcome-level rank baselines remain less stable and obtain lower platform rewards. Ev+EXP3 again performs competitively, suggesting that the evidence signal is useful beyond a specific controller.

C.9 Executable harmful-user behavior modes

The adaptive harmful-user policy does not output raw Mahjong actions directly. Raw action generation would entangle platform-level adaptation with low-level game mechanics. Instead, the harmful-user policy selects a behavior mode, and each mode is executed by resampling actions around the benign-user model.

For a game state s_t, let $\mathcal{A}_k^{\mathrm{b}}(s_t)$ denote the set of the top-k actions under the π_{b} policy. A harmful-user mode indexed by (k, p) samples its action from this top-k benign-user action set with variance level p. Thus k controls the size of the benign-user-compatible action support, while p controls how much the executed action deviates within that support. Smaller variance levels produce more stable novice-like behavior; larger variance levels allow stronger deviations while keeping the action inside the benign-user model's plausible output set.

We evaluate 21 harmful-user behavior modes. The set is

$$\mathcal{KP} = \{(k, p) : k \in \mathcal{K},\ p \in \mathcal{P}_k\}, \qquad |\mathcal{KP}| = 21. \tag{91}$$

Table 7: Additional Mahjong setting summary. Values are cross-seed mean ± standard deviation of per-seed averages over the final evaluation window. Ev denotes reference-policy evidence, Rank denotes the Global Rank Predictor statistic, b/h denote benign/harmful users, p_{h} is the harmful-user share in the novice pool, and G/N denote global/novice-pool mean rank. Lower harmful-user share and benign-user rank indicate better outcomes; harmful-user rank columns are reported as descriptive operating statistics.

Method	$p_{\mathrm{h}} \downarrow$	$\bar{R}_{\mathrm{b}}^{\mathrm{G}} \downarrow$	$\bar{R}_{\mathrm{b}}^{\mathrm{N}} \downarrow$	$\bar{R}_{\mathrm{h}}^{\mathrm{G}}$	$\bar{R}_{\mathrm{h}}^{\mathrm{N}}$
Ev+PPO	$\mathbf{3.35}\% \pm \mathbf{0.31}\%$	2.538 ± 0.026	$\mathbf{2.504} \pm \mathbf{0.002}$	2.750 ± 0.042	2.167 ± 0.021
Ev+Linear	$3.81\% \pm 0.47\%$	$\mathbf{2.521} \pm \mathbf{0.004}$	2.507 ± 0.004	2.678 ± 0.009	2.157 ± 0.022
Ev+EXP3	$3.54\% \pm 0.45\%$	2.522 ± 0.004	2.506 ± 0.003	2.717 ± 0.020	2.161 ± 0.012
Rank+PPO	$11.49\% \pm 1.34\%$	2.633 ± 0.087	2.535 ± 0.009	2.188 ± 0.017	2.188 ± 0.017
Rank+Linear	$15.48\% \pm 8.97\%$	2.822 ± 0.263	2.553 ± 0.040	2.294 ± 0.137	2.233 ± 0.039
Strategy-aware	$11.03\% \pm 0.40\%$	2.722 ± 0.040	2.534 ± 0.003	2.291 ± 0.025	2.183 ± 0.009
Perf-gradient	$10.19\% \pm 0.25\%$	2.547 ± 0.024	2.533 ± 0.007	2.204 ± 0.014	2.204 ± 0.014
EXP3	$11.79\% \pm 1.22\%$	2.890 ± 0.158	2.524 ± 0.010	2.492 ± 0.128	2.190 ± 0.023

For each (k, p), we simulate N_{ctrl} games and record the policy-evidence trajectory and expected rank. The 21 modes exhibit a clear weak-to-strong spectrum: some modes remain close to novice behavior, while others produce stronger harmful-user performance and larger policy-evidence values. This gives the adaptive harmful user a meaningful strategy space, because it can choose how much expert-like behavior to reveal while still producing executable Mahjong actions.

C.10 KP PREDICTOR FOR BEHAVIOR-MODE SELECTION

During dynamic training, the harmful-user behavior-control policy outputs a continuous morphology coordinate

$$\boldsymbol{\alpha}_t \in [0, 1]^2. \tag{92}$$

The KP predictor maps this coordinate, together with the resulting cumulative evidence features, to one of the 21 executable (k, p) behavior modes. Operationally, this means that harmful-user adaptation is realized by choosing a current (k, p) mode: a low-variance benign-like mode for safer camouflage, or a higher-variance mode for stronger rank performance.

The KP predictor receives two cumulative evidence sequences of length L_{KP}. Longer sequences are truncated and shorter sequences are padded. Values are scaled by c_{KP} for numerical stability. The network uses a dual-stream one-dimensional convolutional encoder, slope features, and a shared MLP trunk. It has three prediction heads:

H1: An evidence-intensity head predicts the induced policy-evidence level.
H2: A KP classification head predicts one of the 21 behavior modes.
H3: An expected-rank head predicts rank performance against benign users.

The training loss is

$$\mathcal{L}_{\mathrm{KP}} = \omega_{\mathrm{evi}}\mathcal{L}_{\mathrm{evi}} + \omega_{\mathrm{kp}}\mathcal{L}_{\mathrm{kp}} + \omega_{\mathrm{rank}}\mathcal{L}_{\mathrm{rank}}. \tag{93}$$

Here $\mathcal{L}_{\mathrm{evi}}$ and $\mathcal{L}_{\mathrm{rank}}$ are mean-squared errors, while $\mathcal{L}_{\mathrm{kp}}$ is cross entropy with label smoothing. This module maps continuous behavior control to executable Mahjong actions through a discrete (k, p) mode.

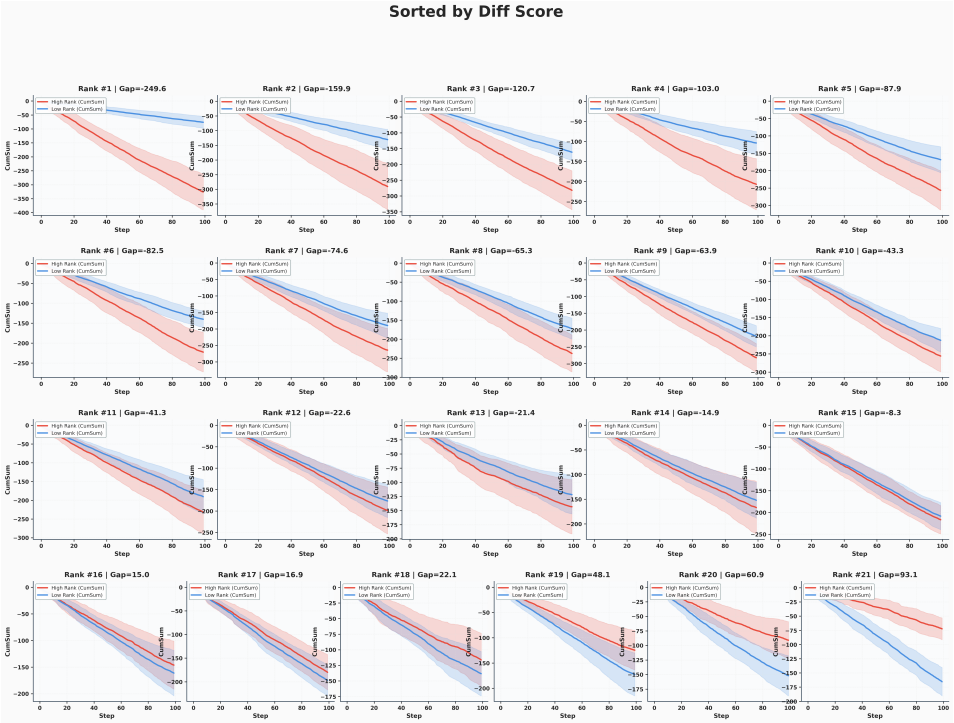


Figure 7: Mean cumulative reference log-likelihoods for the 21 harmful-user behavior modes. The difference between the high-rank and low-rank reference curves is Evidence. Each mode is indexed by (k, p), where k controls the top-k benign-user action support and p controls the sampling variance within that support.

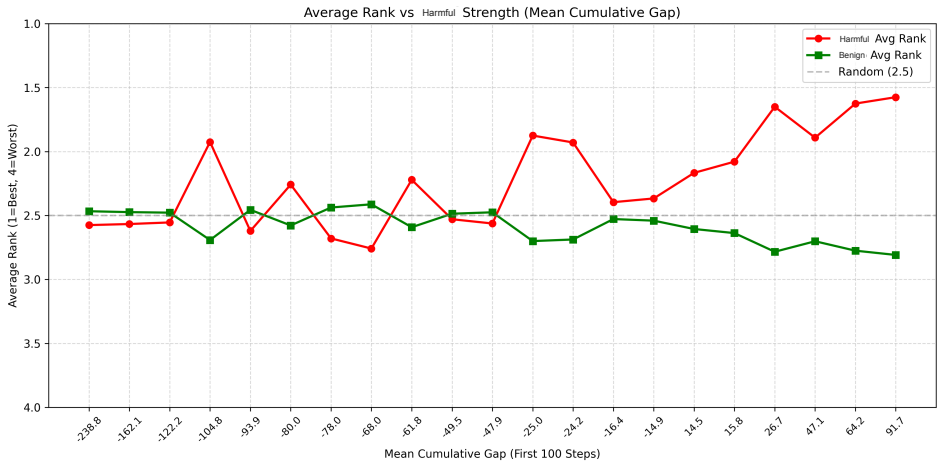


Figure 8: Expected rank against benign users as a function of mean policy evidence. The structured but nontrivial relationship shows that behavioral evidence and rank-predictive performance are related but not identical.

C.11 Fixed-state loss-shape diagnostic

This diagnostic examines how changing the threshold alters routing costs at a fixed response. It separates the immediate effect of the rule from subsequent user adaptation, complementing the closed-loop comparisons in the main text. Theorem 7 and Appendix B.15 provide the corresponding projected-gradient and convex-surrogate benchmarks.

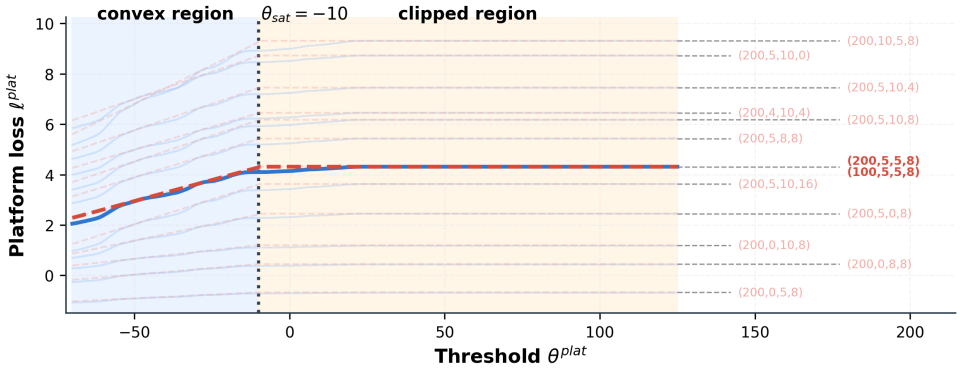


Figure 9: Fixed-state loss-shape diagnostic for the scalar Mahjong threshold loss. Each curve uses the displayed diagnostic weights $(\lambda_{\mathrm{FP}}, \lambda_{\mathrm{FN}}, \lambda_{\mathrm{occ}}, \lambda_{\mathrm{b}})$; the highlighted tuples are $(200, 5, 5, 8)$ and $(100, 5, 5, 8)$. These weights belong to the fixed-state diagnostic defined above. The simulator training coefficients are specified separately in Table 8. The sweep illustrates an active threshold range and saturated, loss-equivalent plateaus.

The fixed-state diagnostic combines routing error with residual harmful-user occupancy and benign-user experience. With the active population and its induced response held fixed, the scalar diagnostic loss takes the form

$$\ell^{\mathrm{plat}}(\theta) = \lambda_{\mathrm{FP}} q^{\mathrm{b}}(\theta) + \lambda_{\mathrm{FN}}\big(1 - q^{\mathrm{h}}(\theta)\big) + \lambda_{\mathrm{occ}} \rho^{\mathrm{h}}(\theta) - \lambda_{\mathrm{b}} \bar{r}^{\mathrm{b}}(\theta),$$

where q^{b} is the benign-user false-routing rate, q^{h} is the harmful-user routing rate, ρ^{h} is residual harmful-user occupancy in the novice pool, and $\bar{r}^{\mathrm{b}}$ summarizes benign-user experience. The sweep in Figure 9 holds the state and response fixed and varies only the platform threshold θ.

The sweep identifies an active threshold range in which routing outcomes change and a saturated range in which additional threshold movement leaves the cost unchanged. A convex upper surrogate describes the active-range costs in the plotted settings. This supplies a fixed-response view of threshold selection; the governance decomposition in Methods additionally accounts for how users adapt after the selected rule is deployed. The projected-gradient benchmark applies under its stated convexity or surrogate conditions.

C.12 Platform population state

The platform state tracks benign users and harmful users in both regions. Let $N^t_{\mathrm{b},1}$ and $N^t_{\mathrm{h},1}$ denote the numbers of benign users and harmful users in the novice pool on day t. Let $N^t_{\mathrm{b},2}$ and $N^t_{\mathrm{h},2}$ denote their counts in the high-rank region. The novice-pool harmful-user share is

$$p^t_{\mathrm{h}} = \frac{N^t_{\mathrm{h},1}}{N^t_{\mathrm{b},1} + N^t_{\mathrm{h},1}}. \tag{94}$$

This ratio determines how often benign users meet harmful users.

In a fair four-player game, the expected rank is R_{fair}. To compute a user's expected rank, fix that user's type and sample its $N_{\mathrm{seat}} - 1$ opponents. Under independent seat sampling, the number Y_- of harmful-user opponents satisfies

$$Y_- \sim \mathrm{Binomial}(N_{\mathrm{seat}} - 1, p^t_{\mathrm{h}}), \qquad P_t(y) = \binom{N_{\mathrm{seat}} - 1}{y} (p^t_{\mathrm{h}})^y (1 - p^t_{\mathrm{h}})^{N_{\mathrm{seat}} - 1 - y}. \tag{95}$$

For a configuration with x benign users and y harmful users, where $x + y = N_{\mathrm{seat}}$, Monte Carlo evaluation gives the conditional mean ranks

$$r'_{\mathrm{b}}(x, y) = \mathbb{E}[R_{\mathrm{b}} \mid x \text{ benign}, y \text{ harmful}], \qquad r'_{\mathrm{h}}(x, y) = \mathbb{E}[R_{\mathrm{h}} \mid x \text{ benign}, y \text{ harmful}], \tag{96}$$

defined when the corresponding player type is present. The focal benign-user contributes one benign-user seat, whereas the focal harmful-user contributes one harmful-user seat. Their novice-pool

expected ranks are therefore

$$\bar{R}^t_{\mathrm{b,nov}} = \sum_{y=0}^{N_{\mathrm{seat}}-1} P_t(y) r'_{\mathrm{b}}(N_{\mathrm{seat}} - y, y), \tag{97}$$

$$\bar{R}^t_{\mathrm{h,nov}} = \sum_{y=0}^{N_{\mathrm{seat}}-1} P_t(y) r'_{\mathrm{h}}(N_{\mathrm{seat}} - 1 - y, y + 1). \tag{98}$$

Let $r^{\mathrm{zone}}_{\mathrm{b}}$ and $r^{\mathrm{zone}}_{\mathrm{h}}$ denote expected ranks in the high-rank region. The platform-wide expected ranks are

$$\bar{R}^t_{\mathrm{b}} = \frac{N^t_{\mathrm{b},1}}{N^t_{\mathrm{b}}} \bar{R}^t_{\mathrm{b,nov}} + \frac{N^t_{\mathrm{b},2}}{N^t_{\mathrm{b}}} r^{\mathrm{zone}}_{\mathrm{b}}, \qquad N^t_{\mathrm{b}} = N^t_{\mathrm{b},1} + N^t_{\mathrm{b},2}, \tag{99}$$

$$\bar{R}^t_{\mathrm{h}} = \frac{N^t_{\mathrm{h},1}}{N^t_{\mathrm{h}}} \bar{R}^t_{\mathrm{h,nov}} + \frac{N^t_{\mathrm{h},2}}{N^t_{\mathrm{h}}} r^{\mathrm{zone}}_{\mathrm{h}}, \qquad N^t_{\mathrm{h}} = N^t_{\mathrm{h},1} + N^t_{\mathrm{h},2}. \tag{100}$$

C.13 Population flow

Each simulated day updates the population through arrival, routing, residence in the high-rank region, return or exit, and churn. We evaluate two flow regimes.

In the Reflow regime, the total population size is fixed. We initialize N^{tot}_0 users with harmful-user fraction p^{h}_0, so

$$N^0_{\mathrm{h},1} = p^{\mathrm{h}}_0 N^{\mathrm{tot}}_0, \qquad N^0_{\mathrm{b},1} = (1 - p^{\mathrm{h}}_0) N^{\mathrm{tot}}_0, \qquad N^0_{\mathrm{b},2} = N^0_{\mathrm{h},2} = 0. \tag{101}$$

No new users arrive and no routed user permanently leaves. Routed harmful users return to the novice pool after $T^{\mathrm{h}}_{\mathrm{return}}$ days, and falsely routed benign users return after $T^{\mathrm{b}}_{\mathrm{return}}$ days.

In the No Reflow regime, the platform has open population flow. New benign users enter at rate

$$\Delta N^t_{\mathrm{b,in}} = r^{\mathrm{b}}_{\mathrm{in}} N^0_{\mathrm{b}}. \tag{102}$$

Harmful-user inflow depends on the current harmful-user population size,

$$\Delta N^t_{\mathrm{h,in}} = \mathrm{clip}\left(\min\left(\beta_{\mathrm{net}} N^t_{\mathrm{h}}, \max(\alpha_{\mathrm{base}} N^0_{\mathrm{h}} r^{\mathrm{h}}_{\mathrm{in}}, N^t_{\mathrm{h}} r^{\mathrm{h}}_{\mathrm{in}})\right), N_{\min}, N_{\max}\right). \tag{103}$$

Benign-user churn increases when the novice-pool expected rank exceeds the tolerated target:

$$\Delta N^t_{\mathrm{b,churn}} = N^t_{\mathrm{b},1} \min\left(c_{\mathrm{b,max}}, c_{\mathrm{b},0} \exp\left(\kappa_{\mathrm{b}}(\bar{R}^t_{\mathrm{b,nov}} - R_{\mathrm{target}})\right)\right). \tag{104}$$

Harmful-user churn in the high-rank region is

$$\Delta N^t_{\mathrm{h,frustration}} = N^t_{\mathrm{h},2} \min\left(c_{\mathrm{h,max}}, c_{\mathrm{h},0} \max(0, r^{\mathrm{zone}}_{\mathrm{h}} - R_{\mathrm{fair}})\right). \tag{105}$$

Routed users leave after their residence time in the high-rank region: harmful users after $T^{\mathrm{h}}_{\mathrm{leave}}$ days and benign users after $T^{\mathrm{b}}_{\mathrm{leave}}$ days. The same residence values are used in Reflow; the only difference is that users return rather than leave.

C.14 Routing statistics and platform threshold

For each day, the platform simulates Mahjong games, computes policy-evidence statistics, and applies the current platform threshold. Let N^t_{caught} be the number of harmful users routed from the novice pool, and let N^t_{false} be the number of benign users falsely routed. The harmful-user capture rate, benign false-routing rate, and harmful-user survival rate are

$$\hat{q}^{\mathrm{h}}_t = \frac{N^t_{\mathrm{caught}}}{N^t_{\mathrm{h},1}}, \qquad \hat{q}^{\mathrm{b}}_t = \frac{N^t_{\mathrm{false}}}{N^t_{\mathrm{b},1}}, \qquad \sigma^t_{\mathrm{surv}} = 1 - \hat{q}^{\mathrm{h}}_t. \tag{106}$$

These are the simulator-level counterparts of the capture quantities. The platform threshold is updated as

$$\theta_t = \theta_{t-1} + \mathrm{clip}(a^{\mathrm{plat}}_t, \delta_{\min}, \delta_{\max}), \tag{107}$$

where a^{plat}_t is the controller action. A larger threshold makes routing more permissive; a smaller threshold increases capture pressure.

C.15 User and platform returns

User returns reward successful play and survival under routing, while platform returns reward benign experience and control of harmful presence. In the training implementation, harmful users share a behavior-control policy whose feedback is aggregated over the active harmful users. The two learners optimize the separate objectives below.

The harmful-user reward is

$$\begin{aligned} r_t^{\mathrm{h}} = \lambda_{\mathrm{rank}}^{\mathrm{h}}(R_{\mathrm{base}} - \bar{R}_{\mathrm{h}}^t) - \lambda_{\mathrm{surv}}^{\mathrm{h}} \max(0, \sigma_{\mathrm{target}} - \sigma_{\mathrm{surv}}^t) \\ - \lambda_{\mathrm{stab}}^{\mathrm{h}} \|\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_{t-1}\|_1. \end{aligned} \tag{108}$$

The first term rewards rank dominance, the second penalizes excessive capture, and the third discourages unstable jumps in behavior mode.

The platform reward is the negative of a shaped simulator loss aligned with the platform objective. Define the benign experience anchor

$$R_{\mathrm{anchor}}(\bar{R}_{\mathrm{b}}) = \begin{cases} b_{\mathrm{good}}, & \bar{R}_{\mathrm{b}} \le R_{\mathrm{fair}}, \\ b_{\mathrm{ok}}, & R_{\mathrm{fair}} < \bar{R}_{\mathrm{b}} \le R_{\mathrm{target}}, \\ -c_{\mathrm{rank}}(\bar{R}_{\mathrm{b}} - R_{\mathrm{target}}), & \bar{R}_{\mathrm{b}} > R_{\mathrm{target}}. \end{cases} \tag{109}$$

Define the benign false-routing penalty

$$L_{\mathrm{FP}}(\hat{q}_t^{\mathrm{b}}) = \begin{cases} c_{\mathrm{FP}} \hat{q}_t^{\mathrm{b}}, & \hat{q}_t^{\mathrm{b}} > 0, \\ -b_{\mathrm{safe}}, & \hat{q}_t^{\mathrm{b}} = 0. \end{cases} \tag{110}$$

The shaped platform loss is

$$\begin{aligned} \widehat{\ell}_t^{\mathrm{plat}} = L_{\mathrm{FP}}(\hat{q}_t^{\mathrm{b}}) + \lambda_{\mathrm{FN}}^{\mathrm{plat}} \left(q_{\mathrm{cap}}^{\mathrm{target}} - \min(\hat{q}_t^{\mathrm{h}}, q_{\mathrm{cap}}^{\mathrm{target}}) \right) \\ + \lambda_{\mathrm{occ}}^{\mathrm{plat}} p_{\mathrm{h}}^t - \lambda_{\mathrm{b}}^{\mathrm{plat}} R_{\mathrm{anchor}}(\bar{R}_{\mathrm{b}}^t) + \lambda_{\Delta\theta}^{\mathrm{plat}} |\theta_t - \theta_{t-1}|. \end{aligned} \tag{111}$$

The platform reward is

$$r_t^{\mathrm{plat}} = -\widehat{\ell}_t^{\mathrm{plat}}. \tag{112}$$

This training objective combines routing and operational feedback. The false-routing penalty includes a bonus for zero benign captures, and the missed-capture term measures shortfall from a target rather than total uncaptured harmful exposure. The nonnegative weights $\lambda_{\mathrm{FN}}^{\mathrm{plat}}$, $\lambda_{\mathrm{occ}}^{\mathrm{plat}}$, $\lambda_{\mathrm{b}}^{\mathrm{plat}}$, and $\lambda_{\Delta\theta}^{\mathrm{plat}}$ respectively price capture shortfall, novice-pool harmful-user occupancy, benign-rank reward, and threshold movement. Equation (64) instead restates the theoretical governance loss using routing rates and the safe-mode share. The shaped reward specifies the experimental controller's feedback; Theorem 1 evaluates its induced trajectory through L_t and user regret.

C.16 Training protocol

Each day collects K_{day} platform transitions, usually five, using the current harmful-user behavior mode. Each transition generates a behavior batch and its evidence. The platform controller uses five evidence quantiles and the current threshold to sample a Gaussian action. Its decoder squashes that action and limits the threshold change. The updated threshold routes the batch, and population transitions produce the recorded reward. Platform network parameters stay fixed during this rollout, although actions vary with observations.

At day end, the platform computes targets from the executed rollout and updates its PPO network. The harmful-user agent then selects a new morphology coordinate $\boldsymbol{\alpha}$, maps it through the KP predictor to an executable (k, p) mode, and performs its PPO update. Its record uses the collected daily rank and survival statistics together with the morphology-change penalty; the selected mode is used in subsequent simulation. Thus platform actions use current-batch evidence, while user-mode changes affect later batches.

Both policies share the PPO implementation, with a state-value head $V(o)$, clipped actor objective, entropy term, and Adam. With rollout index i and terminal mask d_i, the reward-sign GAE targets (Schulman et al., 2015) are

$$\delta_i^r = r_i + \gamma(1 - d_i)\widehat{V}_{i+1} - \widehat{V}_i, \qquad \widehat{A}_i = \delta_i^r + \gamma\lambda(1 - d_i)\widehat{A}_{i+1}, \qquad \widehat{R}_i = \widehat{V}_i + \widehat{A}_i. \tag{113}$$

Table 8: Population-flow, reward, and simulator parameters used in the Competitive Mahjong platform experiment.

Quantity	Symbol	Value
Number of seats per game	N_{seat}	4
Fair expected rank	R_{fair}	2.5
Elo scale	c_{elo}	1200
Controlled games per setting	N_{ctrl}	240
Initial reflow population	N_0^{tot}	10000
Initial harmful-user fraction in reflow	p_0^{h}	0.1
Harmful-user return time in reflow	$T_{\text{return}}^{\text{h}}$	3 days
Benign-user return time in reflow	$T_{\text{return}}^{\text{b}}$	7 days
Harmful-user leave time in no-reflow	$T_{\text{leave}}^{\text{h}}$	3 days
Benign-user leave time in no-reflow	$T_{\text{leave}}^{\text{b}}$	7 days
Benign-user churn maximum	$c_{\text{b,max}}$	0.3
Benign-user churn base rate	$c_{\text{b,0}}$	0.05
Benign-user churn sensitivity	κ_{b}	2.0
Harmful-user frustration maximum	$c_{\text{h,max}}$	0.2
Harmful-user frustration base rate	$c_{\text{h,0}}$	0.1
Harmful-user survival reference rank	R_{base}	2.6
Harmful-user survival target	σ_{target}	0.8
Harmful-user rank reward weight	$\lambda_{\text{rank}}^{\text{h}}$	100
Harmful-user survival weight	$\lambda_{\text{surv}}^{\text{h}}$	200
Harmful-user stability weight	$\lambda_{\text{stab}}^{\text{h}}$	0.5
Capture target	$q_{\text{cap}}^{\text{target}}$	0.8
Platform missed-capture weight	$\lambda_{\text{FN}}^{\text{plat}}$	3.0
Platform occupancy weight	$\lambda_{\text{occ}}^{\text{plat}}$	2.0
Platform benign-experience weight	$\lambda_{\text{b}}^{\text{plat}}$	1.0
Threshold-change weight	$\lambda_{\Delta\theta}^{\text{plat}}$	0.05
Good benign-user-rank anchor	b_{good}	1.5
Acceptable benign-user-rank anchor	b_{ok}	1.0
Benign-user-rank violation slope	c_{rank}	2.0
False-routing slope	c_{FP}	100
No-false-routing bonus	b_{safe}	1.0
Remaining-wall Riichi limit	T_{riichi}	20 tiles

The critic regresses to $\widehat{R}_i$ and the actor uses $\widehat{A}_i$; no separate action-value network is trained. The daily platform update uses zero terminal bootstrap, and the harmful-user record is terminal. Population and return queues persist across days. These updates implement the repeated evidence, routing, and adaptation loop. Appendix B.11 identifies the induced mode mixtures, distinguishes training feedback from evaluated outcomes, and relates them to the governance analysis.

The hyperparameters are listed in Table 9. Each controlled player configuration uses N_{ctrl} Monte Carlo games. Platform rollout length H, the Monte Carlo game count, and PPO optimization epochs describe separate operations.

C.17 Grouped routing ablations in the Competitive Mahjong simulator

Every fixed-threshold configuration is deployed in its own simulator run, producing its own strategic response and population trajectory. These comparisons evaluate independent deployment. The realized-response benchmark instead holds the learner's complete response profile fixed; Appendix B.13 relates the two criteria.

We organize the detailed Mahjong ablations into four groups according to the platform routing rule and the harmful-user response model. The panels report the novice-pool harmful-user share, population distribution, and expected rank trajectories for benign and harmful users. Panels are grouped by intervention to compare the resulting routing and adaptation dynamics.

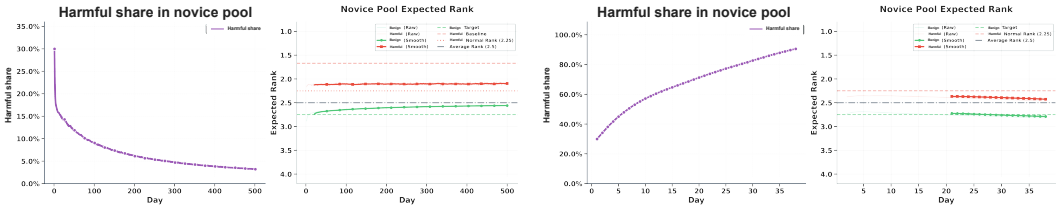
Table 9: Training hyperparameters for the Competitive Mahjong platform simulation.

Parameter	**Symbol**	**Value**
Discount factor	γ	0.99
GAE parameter	λ_{GAE}	0.95
PPO clip range	ϵ_{PPO}	0.2
PPO epochs per update	K_{epochs}	2
Learning rate	η_{lr}	3×10^{-4}
Batch size	B_{batch}	64
Harmful-user state dimension	$\dim(s^{\text{h}})$	204
Platform state dimension	$\dim(y^{\text{plat}})$	6
Morphology step size	η_{α}	0.05
KP sequence length	L_{KP}	100
KP sequence scale	c_{KP}	0.01
KP evidence-loss weight	ω_{evi}	0.3
KP classification-loss weight	ω_{kp}	1.0
KP rank-loss weight	ω_{rank}	1.0
Threshold update bounds	$(\delta_{\min}, \delta_{\max})$	$(-15.0, 15.0)$
Game batches per day	K_{day}	5
Monte Carlo games per setting	N_{ctrl}	240

The first group studies a rank-predictive routing rule. It compares unrestricted harmful and camouflaged harmful behavior and shows when routing based on rank-predictive statistics succeeds or weakens. The second group studies a fixed policy-evidence threshold with non-adaptive harmful users, isolating the effect of threshold choice when harmful users do not adapt online. The third group studies a fixed policy-evidence threshold with adaptive harmful users, testing whether a single fixed threshold remains effective against adaptive behavior-mode selection. The last group studies the full closed-loop setting, where both the platform controller and adaptive harmful users update during the rollout.

GROUP 1: RANK-PREDICTIVE ROUTING

This group compares two rank-predictive runs under a high initial harmful-user fraction $p_0^{\text{h}} = 0.3$. In the unrestricted harmful case, harmful users remain clearly stronger in the rank-predictive statistic, so the platform can separate them more easily. In the camouflaged harmful case, harmful users move closer to the novice operating range, making routing based only on expected rank less reliable. This comparison highlights the limitation of rank-predictive regulation once harmful users partially mimic benign performance.



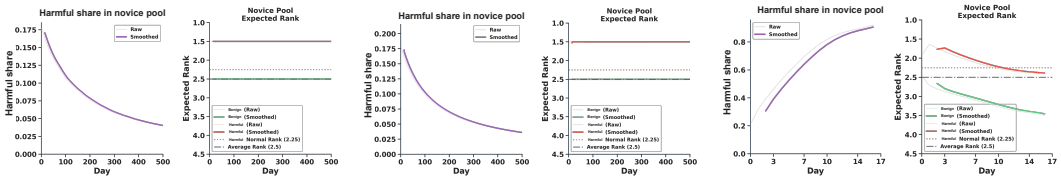
(a) Unrestricted harmful users, $R^{\text{h}}_{\text{target}} = 1.67$. (b) Camouflaged harmful users, $R^{\text{h}}_{\text{target}} = 2.25$.

Figure 10: Rank-predictive routing baselines under unrestricted harmful and camouflaged harmful behavior. Both settings use $T = 500$, $p_0^{\text{h}} = 0.3$, and $R^{\text{b}}_{\text{target}} = 2.75$.

GROUP 2: FIXED PLATFORM, NON-ADAPTIVE HARMFUL USERS

This group studies fixed-threshold routing against non-adaptive harmful users. Since the harmful-user side does not adapt online, these runs mainly show how the long-run regime changes as the policy-

evidence threshold moves from aggressive to permissive settings. A lower threshold increases capture pressure, while a higher threshold leaves more room for harmful users to remain in the novice pool.



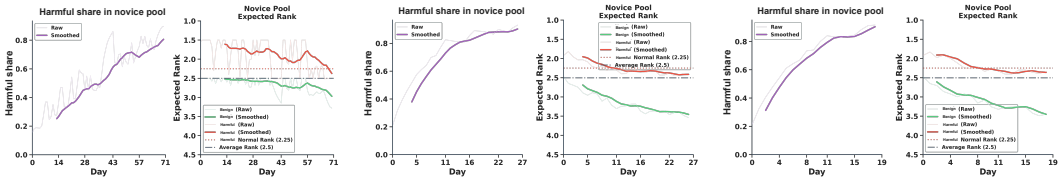
(a) $\theta = -100$ (b) $\theta = 100$ (c) $\theta = 200$

Figure 11: Fixed platform with non-adaptive harmful users under different policy-evidence thresholds, $T = 500$, $p_0^{\mathrm{h}} = 0.1$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.25$, and $R_{\mathrm{target}}^{\mathrm{b}} = 2.75$.

GROUP 3: FIXED PLATFORM, ADAPTIVE HARMFUL USERS

This group reports fixed-threshold routing against adaptive harmful users. This is the key stress test for static platform rules. Once harmful users adapt their current behavior mode, no single fixed threshold works well across all target configurations. Aggressive thresholds may reduce harmful-user persistence, but they can also distort the novice pool through false routing. Permissive thresholds better protect benign users, but they often allow harmful users to remain in the novice pool longer.

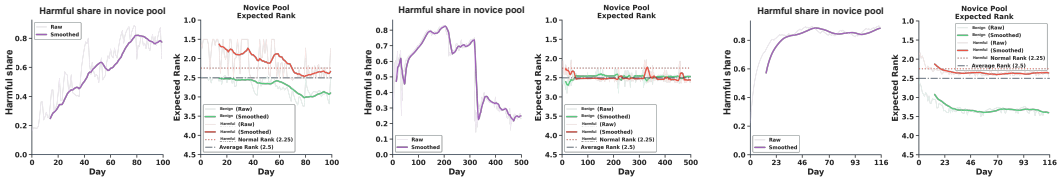
Setting A: $R_{\mathrm{target}}^{\mathrm{h}} = 2.00$**,** $R_{\mathrm{target}}^{\mathrm{b}} = 2.75$**.**



(a) $\theta = -70$ (b) $\theta = 50$ (c) $\theta = 120$

Figure 12: Fixed platform with adaptive harmful users under different policy-evidence thresholds, $T = 500$, $p_0^{\mathrm{h}} = 0.1$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.00$, and $R_{\mathrm{target}}^{\mathrm{b}} = 2.75$.

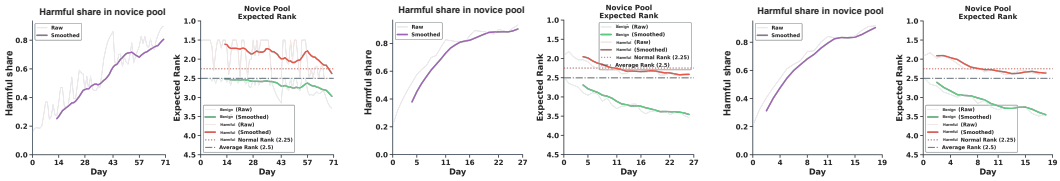
Setting B: $R_{\mathrm{target}}^{\mathrm{h}} = 2.00$**,** $R_{\mathrm{target}}^{\mathrm{b}} = 3.00$**.**



(a) $\theta = -70$ (b) $\theta = 50$ (c) $\theta = 120$

Figure 13: Fixed platform with adaptive harmful users under different policy-evidence thresholds, $T = 500$, $p_0^{\mathrm{h}} = 0.1$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.00$, and $R_{\mathrm{target}}^{\mathrm{b}} = 3.00$.

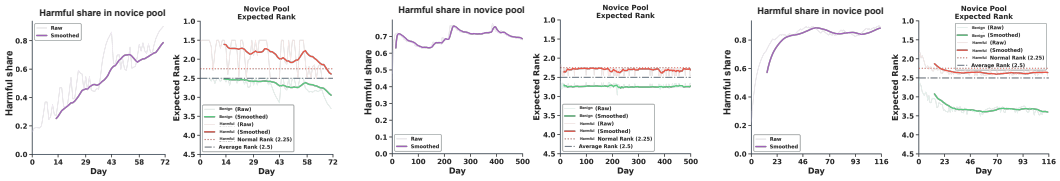
Setting C: $R^{\mathrm{h}}_{\mathrm{target}} = 2.25$**,** $R^{\mathrm{b}}_{\mathrm{target}} = 2.75$**.**



(a) $\theta = -70$ (b) $\theta = 50$ (c) $\theta = 120$

Figure 14: Fixed platform with adaptive harmful users under different policy-evidence thresholds, $T = 500$, $p^{\mathrm{h}}_0 = 0.1$, $R^{\mathrm{h}}_{\mathrm{target}} = 2.25$, and $R^{\mathrm{b}}_{\mathrm{target}} = 2.75$.

Setting D: $R^{\mathrm{h}}_{\mathrm{target}} = 2.25$**,** $R^{\mathrm{b}}_{\mathrm{target}} = 3.00$**.**



(a) $\theta = -70$ (b) $\theta = 50$ (c) $\theta = 120$

Figure 15: Fixed platform with adaptive harmful users under different policy-evidence thresholds, $T = 500$, $p^{\mathrm{h}}_0 = 0.1$, $R^{\mathrm{h}}_{\mathrm{target}} = 2.25$, and $R^{\mathrm{b}}_{\mathrm{target}} = 3.00$.

GROUP 4: DYNAMIC PLATFORM, ADAPTIVE HARMFUL USERS

This group shows the full adaptive setting, where the platform updates its routing threshold online and harmful users adapt their behavior mode. This is the main closed-loop regime. Compared with fixed-threshold baselines, the dynamic platform responds to changes in population state and produces a more balanced long-run trajectory. The $T = 500$ runs emphasize transient behavior and intermediate stabilization, while the $T = 1000$ runs show the longer-horizon effect of closed-loop regulation.

Setting A: $T = 500$, $p_0^{\mathrm{h}} = 0.1$.

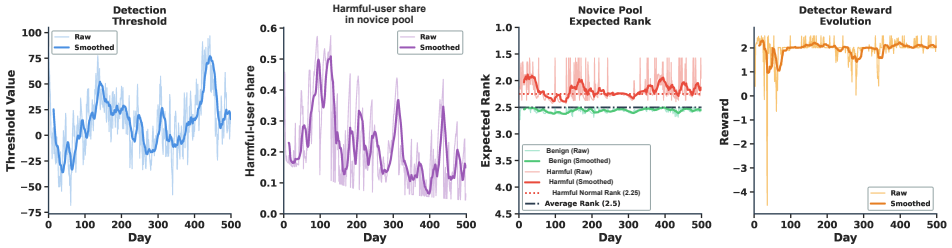


Figure 16: Dynamic platform with adaptive harmful users, $T = 500$, $p_0^{\mathrm{h}} = 0.1$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.00$, and $R_{\mathrm{target}}^{\mathrm{b}} = 2.75$.

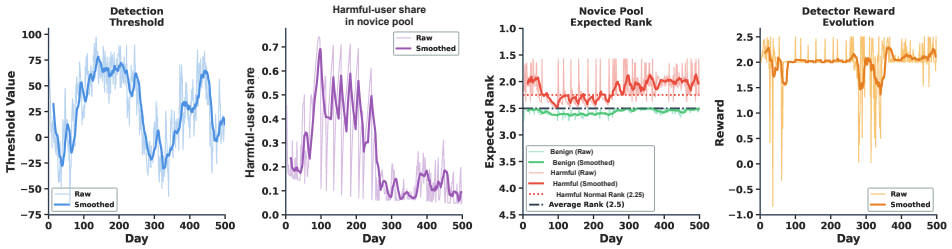


Figure 17: Dynamic platform with adaptive harmful users, $T = 500$, $p_0^{\mathrm{h}} = 0.1$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.00$, and $R_{\mathrm{target}}^{\mathrm{b}} = 3.00$.

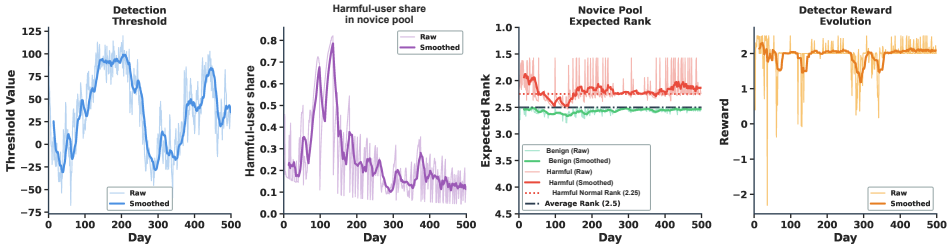


Figure 18: Dynamic platform with adaptive harmful users, $T = 500$, $p_0^{\mathrm{h}} = 0.1$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.25$, and $R_{\mathrm{target}}^{\mathrm{b}} = 2.75$.

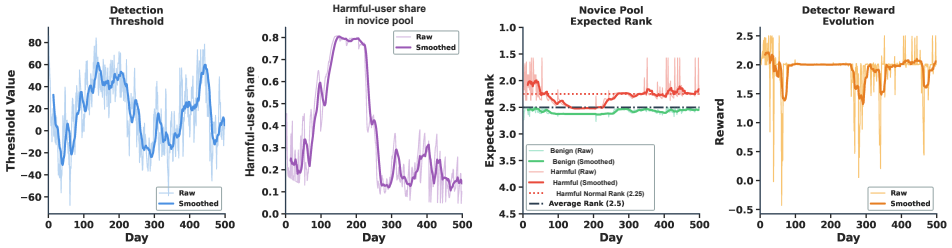


Figure 19: Dynamic platform with adaptive harmful users, $T = 500$, $p_0^{\mathrm{h}} = 0.1$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.25$, and $R_{\mathrm{target}}^{\mathrm{b}} = 3.00$.

Setting B: $T = 500$, $p_0^{\mathrm{h}} = 0.3$.

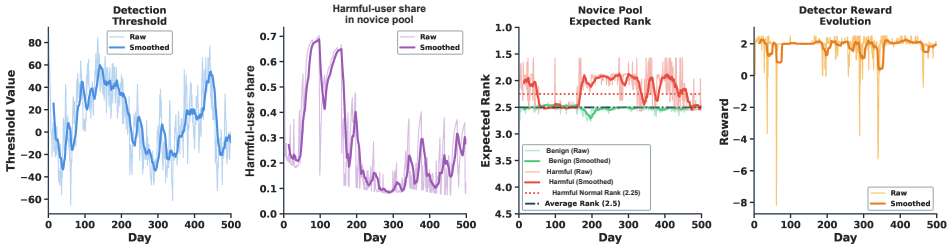


Figure 20: Dynamic platform with adaptive harmful users, $T = 500$, $p_0^{\mathrm{h}} = 0.3$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.00$, and $R_{\mathrm{target}}^{\mathrm{b}} = 2.75$.

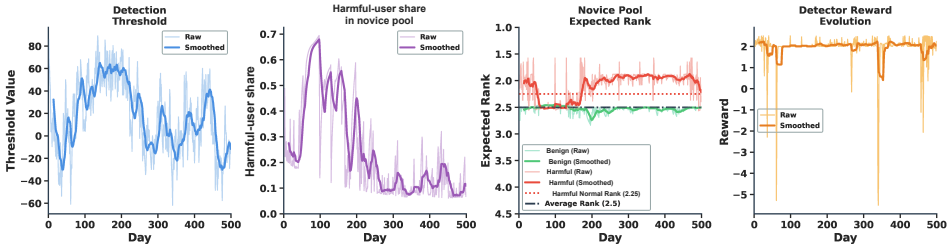


Figure 21: Dynamic platform with adaptive harmful users, $T = 500$, $p_0^{\mathrm{h}} = 0.3$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.00$, and $R_{\mathrm{target}}^{\mathrm{b}} = 3.00$.

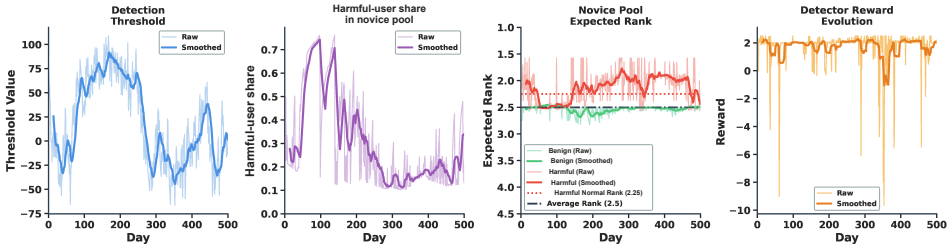


Figure 22: Dynamic platform with adaptive harmful users, $T = 500$, $p_0^{\mathrm{h}} = 0.3$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.25$, and $R_{\mathrm{target}}^{\mathrm{b}} = 2.75$.

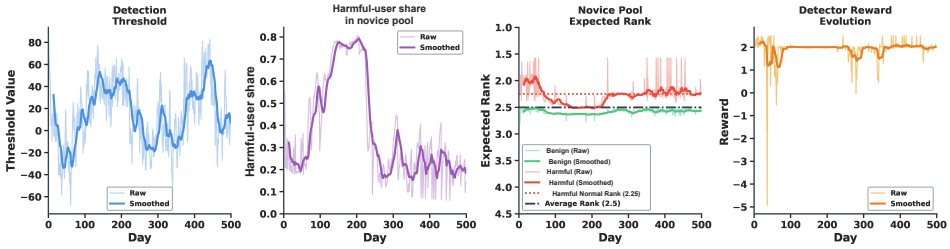


Figure 23: Dynamic platform with adaptive harmful users, $T = 500$, $p_0^{\mathrm{h}} = 0.3$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.25$, and $R_{\mathrm{target}}^{\mathrm{b}} = 3.00$.

Setting C: $T = 1000$, $p_0^{\mathrm{h}} = 0.1$.

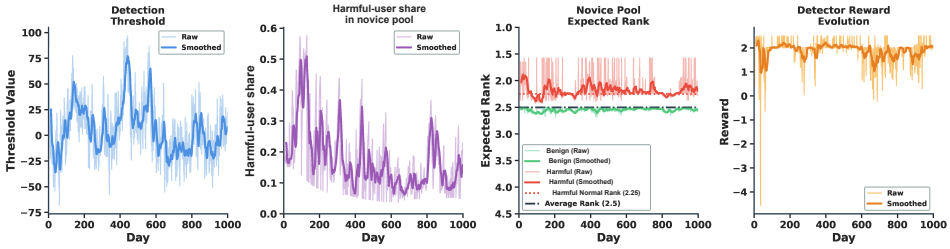


Figure 24: Dynamic platform with adaptive harmful users, $T = 1000$, $p_0^{\mathrm{h}} = 0.1$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.00$, and $R_{\mathrm{target}}^{\mathrm{b}} = 2.75$.

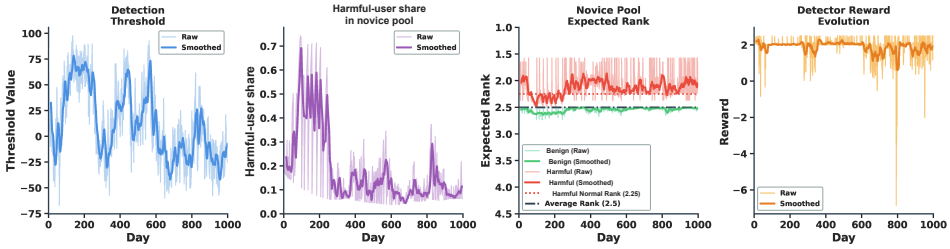


Figure 25: Dynamic platform with adaptive harmful users, $T = 1000$, $p_0^{\mathrm{h}} = 0.1$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.00$, and $R_{\mathrm{target}}^{\mathrm{b}} = 3.00$.

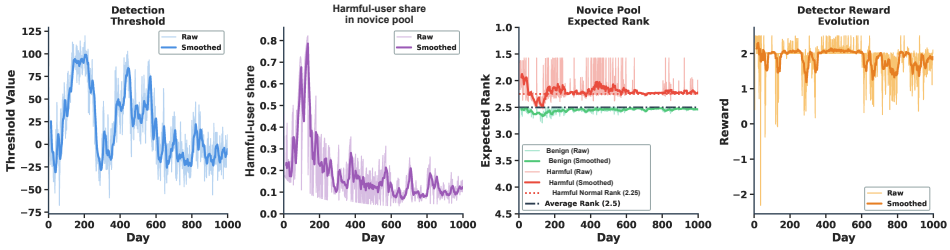


Figure 26: Dynamic platform with adaptive harmful users, $T = 1000$, $p_0^{\mathrm{h}} = 0.1$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.25$, and $R_{\mathrm{target}}^{\mathrm{b}} = 2.75$.

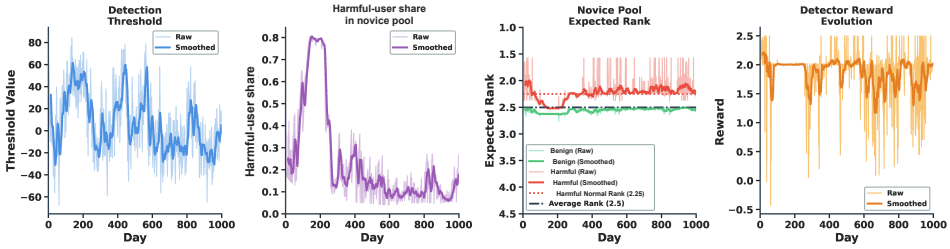


Figure 27: Dynamic platform with adaptive harmful users, $T = 1000$, $p_0^{\mathrm{h}} = 0.1$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.25$, and $R_{\mathrm{target}}^{\mathrm{b}} = 3.00$.

Setting D: $T = 1000$, $p_0^{\mathrm{h}} = 0.3$.

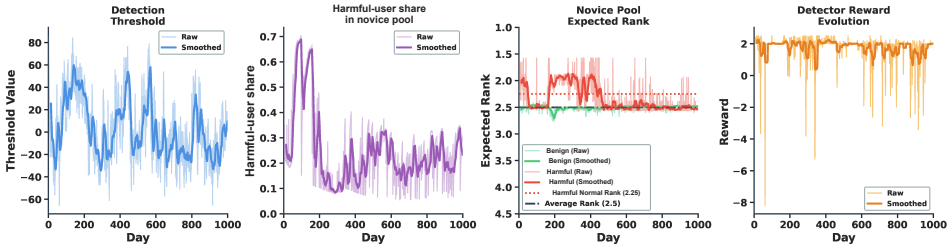


Figure 28: Dynamic platform with adaptive harmful users, $T = 1000$, $p_0^{\mathrm{h}} = 0.3$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.00$, and $R_{\mathrm{target}}^{\mathrm{b}} = 2.75$.

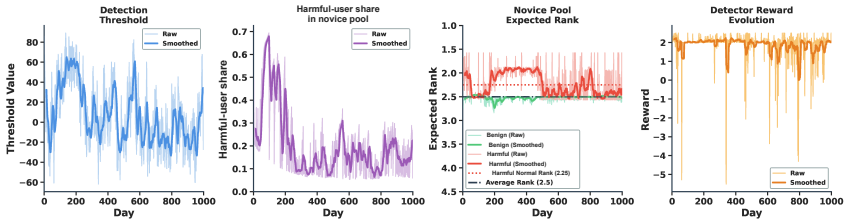


Figure 29: Dynamic platform with adaptive harmful users, $T = 1000$, $p_0^{\mathrm{h}} = 0.3$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.00$, and $R_{\mathrm{target}}^{\mathrm{b}} = 3.00$.

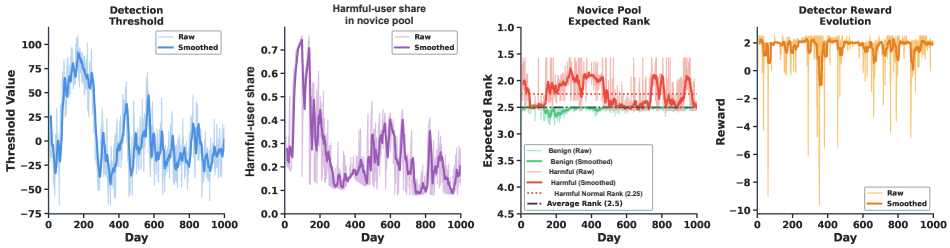


Figure 30: Dynamic platform with adaptive harmful users, $T = 1000$, $p_0^{\mathrm{h}} = 0.3$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.25$, and $R_{\mathrm{target}}^{\mathrm{b}} = 2.75$.

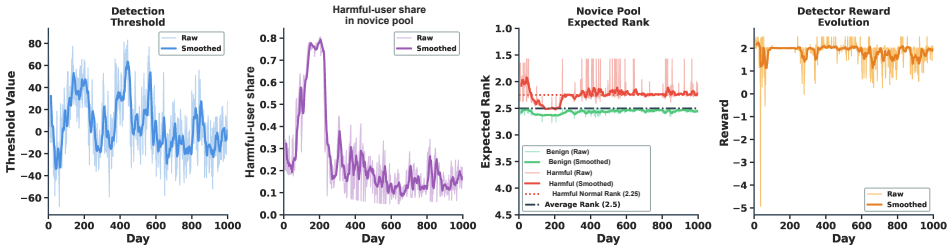


Figure 31: Dynamic platform with adaptive harmful users, $T = 1000$, $p_0^{\mathrm{h}} = 0.3$, $R_{\mathrm{target}}^{\mathrm{h}} = 2.25$, and $R_{\mathrm{target}}^{\mathrm{b}} = 3.00$.

Taken together, the four groups support the same conclusion as the main experiments. Rank-predictive routing is informative when harmful users remain clearly stronger than genuine newcomers, but it

weakens once harmful users camouflage their rank outcomes. Fixed policy-evidence thresholds are adequate only in simpler non-adaptive settings and become brittle against adaptive harmful users. The dynamic platform performs best in the fully adaptive regime because it updates the routing threshold over time and maintains a more stable trade-off between harmful-user containment and benign-user protection.

D Detailed Non-Competitive Platform Experiment

This appendix gives the simulator, population dynamics, learning setup, and hyperparameters for the non-competitive open-platform experiment in Section 5.2. The experiment implements the evidence–routing–adaptation loop from Section 4 in a setting without explicit matches or rank outcomes. Routing changes harmful-user opportunities and benign service quality, which in turn affect population inflow, churn, and the next round of behavior. The normal region contains active benign and harmful users, and the simulator-level quarantine region is the pool of removed harmful users R_t.

D.1 Simulator state and behavior sequences

Each rollout starts from active benign population $N_{\mathrm{b},0}$, active harmful-user population $N_{\mathrm{h},0}$, and pool of removed harmful users R_0. Each active user generates a behavior window

$$\mathcal{W}_{t,i} = ((z_\tau, a_\tau))_{\tau=1}^{L_{\mathrm{seq}}},$$

where z_τ is the simulator latent state and a_τ is a behavior-intensity action. The latent state space has size $|\mathcal{Z}|$ and is induced by exposure and opportunity variables; the action space has size $|\mathcal{A}|$. At each day, a context mode $j_t \in \mathcal{C}$ is sampled from a categorical distribution with probabilities $\boldsymbol{\omega}_{\mathcal{C}}$. The context mode affects transition dynamics and sequence rewards. Its ecological pressure is $c_t = \chi(j_t)$, where χ maps each mode to its pressure value.

The harmful-user share is

$$p_{\mathrm{h},t} = \frac{N_{\mathrm{h},t}}{N_{\mathrm{b},t} + N_{\mathrm{h},t}}.$$

This quantity is the simulator-level harmful-user occupancy in the normal region. The platform also records daily benign false routings, harmful-user routings, and an ecological welfare score W_t.

D.2 Policy-evidence statistics and routing rules

For each behavior window, the platform computes one scalar statistic and compares it with a platform threshold. We evaluate three platform controllers. Static Outcome uses a fixed threshold on a surface outcome statistic computed from the realized behavior sequence. Static Evidence uses a fixed threshold on reference-policy evidence. Ev+AC, the adaptive threshold controller, uses the same evidence and updates the threshold online. These are the names used in the main-text comparison.

The reference-policy evidence score is $e(\mathcal{W})$ from (1). Here b denotes normal service users and h denotes users who obtain additional payoff from hidden harmful behavior. The reference family contains a benign reference policy π_{b} and a harmful-behavior reference policy π_{h}. For a behavior window $\mathcal{W}$, we compute

$$S_{\mathrm{h}}(\mathcal{W}) = \sum_{\tau=1}^{L_{\mathrm{seq}}} \log \pi_{\mathrm{h}}(a_\tau \mid z_\tau), \qquad S_{\mathrm{b}}(\mathcal{W}) = \sum_{\tau=1}^{L_{\mathrm{seq}}} \log \pi_{\mathrm{b}}(a_\tau \mid z_\tau),$$

and define the scalar policy-evidence statistic

$$e(\mathcal{W}) = S_{\mathrm{h}}(\mathcal{W}) - S_{\mathrm{b}}(\mathcal{W}).$$

Larger $e(\mathcal{W})$ favors the harmful-behavior reference over the benign reference.

Given a selected statistic $s(\mathcal{W})$ and threshold θ_t, the simulator converts the statistic into a routing probability by

$$d_{\theta_t}(\mathcal{W}) = \sigma\big(\beta_{\mathrm{route}}\big(s(\mathcal{W}) - \theta_t\big)\big).$$

This is the probabilistic routing rule $d_\theta \in [0,1]$ covered in Appendix B.1, with sigmoid $\sigma(x) = (1+e^{-x})^{-1}$. For STATIC OUTCOME, $s(\mathcal{W})$ is the surface outcome statistic and $\theta_t = \theta_{\text{out}}$. For STATIC EVIDENCE, $s(\mathcal{W}) = e(\mathcal{W})$ and $\theta_t = \theta_{\text{ref}}$. For EV+AC, $s(\mathcal{W}) = e(\mathcal{W})$ and the learned controller selects $\theta_t \in [\theta_{\min}, \theta_{\max}]$.

Let N^t_{false} be the number of benign users routed on day t, and let N^t_{caught} be the number of harmful users routed on day t. The empirical benign false-routing rate and harmful-user capture rate are

$$\hat{q}^{\text{b}}_t = \frac{N^t_{\text{false}}}{N_{\text{b},t}}, \qquad \hat{q}^{\text{h}}_t = \frac{N^t_{\text{caught}}}{N_{\text{h},t}}.$$

These rates summarize realized routing within the active population. Their denominators evolve with inflow and churn, while (64) uses fixed model population weights.

D.3 ECOLOGICAL WELFARE AND POPULATION DYNAMICS

The per-user welfare score W_t summarizes benign service quality. Figure 5 reports total benign welfare $\mathcal{W}_t = N_{\text{b},t} W_t$, which combines this quality with the number of active benign users. The per-user score is defined as

$$W_t = \text{clip}\big(1 - \lambda^{\text{eco}}_{\text{occ}} p_{\text{h},t} - \lambda^{\text{eco}}_{\text{h}} \bar{g}^{\text{surv}}_t - \lambda^{\text{eco}}_{\text{FP}} \hat{q}^{\text{b}}_t - c_t, W_{\min}, W_{\max}\big),$$

where $\bar{g}^{\text{surv}}_t$ is the average payoff of surviving harmful users and c_t is the context-dependent ecological pressure. Higher W_t indicates a healthier benign-user ecology. This welfare term aligns with the platform objective: $\hat{q}^{\text{b}}_t$ measures benign false capture, $1 - \hat{q}^{\text{h}}_t$ measures missed harmful-user capture, $p_{\text{h},t}$ measures harmful-user occupancy, and W_t measures benign experience.

After routing, the active populations are updated by inflow, churn, and removed-pool decay. Benign churn increases when welfare is low:

$$d^{\text{b}}_t = d^{\text{b}}_{\min} + (d^{\text{b}}_{\max} - d^{\text{b}}_{\min})\sigma\left(\frac{\tau_{\text{out}} - W_t + c_t}{s_{\text{out}}}\right).$$

Benign inflow increases when welfare is high:

$$I^{\text{b}}_t = I^{\text{b}}_{\min} + (I^{\text{b}}_{\max} - I^{\text{b}}_{\min})\sigma\left(\frac{W_t - \tau_{\text{in}} - c_t}{s_{\text{in}}}\right).$$

Harmful-user inflow is driven by the exploitable value

$$V_t = p_{\text{h},t} \bar{g}^{\text{surv}}_t,$$

through

$$I^{\text{h}}_t = I^{\text{h}}_{\min} + (I^{\text{h}}_{\max} - I^{\text{h}}_{\min})\sigma\left(\frac{V_t - \tau_{\text{h}} + \lambda^{\text{h}}_c c_t}{s_{\text{h}}}\right).$$

In the implementation, inflows are sampled from Poisson distributions. Harmful-user inflow is additionally lower-bounded by $I^{\text{h}}_{\text{floor}}$.

Harmful-user self-churn is enabled and increases when harmful-user payoff is low:

$$d^{\text{h}}_t = d^{\text{h}}_{\min} + (d^{\text{h}}_{\max} - d^{\text{h}}_{\min})\sigma\left(\frac{\tau^{\text{h}}_{\text{churn}} - \bar{g}_t}{s^{\text{h}}_{\text{churn}}}\right),$$

where $\bar{g}_t$ is the average harmful-user sequence payoff.

Let

$$\widetilde{N}_{\text{b},t} = N_{\text{b},t} - N^t_{\text{false}}, \qquad \widetilde{N}_{\text{h},t} = N_{\text{h},t} - N^t_{\text{caught}}$$

be the post-routing active populations. The stochastic population update is

$$N_{\text{b},t+1} = \widetilde{N}_{\text{b},t} - C^{\text{b}}_t + U^{\text{b}}_t, \qquad C^{\text{b}}_t \sim \text{Binomial}(\widetilde{N}_{\text{b},t}, d^{\text{b}}_t), \qquad U^{\text{b}}_t \sim \text{Poisson}(I^{\text{b}}_t),$$

$$N_{\text{h},t+1} = \widetilde{N}_{\text{h},t} - C^{\text{h}}_t + U^{\text{h}}_t, \qquad C^{\text{h}}_t \sim \text{Binomial}(\widetilde{N}_{\text{h},t}, d^{\text{h}}_t), \qquad U^{\text{h}}_t = \max\{I^{\text{h}}_{\text{floor}}, \text{Poisson}(I^{\text{h}}_t)\}.$$

Captured harmful users enter the removed pool, which decays at rate q_R:

$$R_{t+1} = R_t + N^t_{\text{caught}} - D^R_t, \qquad D^R_t \sim \text{Binomial}(R_t + N^t_{\text{caught}}, q_R).$$

Benign false routings are modeled as daily outflow from the active benign population and are not stored in a persistent pool.

D.4 Learning setup

The Ev+AC controller and adaptive harmful users are trained with online actor-critic policy-gradient updates. The platform controller observes aggregate platform features and outputs a continuous threshold θ_t. Its reward encourages harmful-user capture while penalizing benign false routing, welfare degradation, and large threshold changes. The user policy balances payoff, survival, and strategy stability.

Both the platform controller and harmful-user policy use two-layer MLPs with hidden dimension d_{hid}. We optimize all learned policies with Adam. The platform policy is initialized as a Gaussian policy with standard deviation $\sigma_{\theta,0}$. The harmful-user policy parameterizes a camouflage coefficient that interpolates between benign-like and harmful-like behavior. Its Beta concentration is restricted to the interval $[\kappa_{\min}^{\text{Beta}}, \kappa_{\max}^{\text{Beta}}]$ and uses numerical clipping ϵ_{num}. The harmful-user policy is initialized in a low-evidence camouflage regime using target evidence $e_{\text{init}}^{\text{h}}$ and concentration fraction $\rho_{\text{init}}^{\text{h}}$.

The main configuration shares user-policy parameters across G_{pseudo} training groups. The alternative configuration uses G_{ind} independently initialized user policies, each trained from local trajectory feedback with the same actor-critic optimizer. It uses the legacy group-splitting scheme and inflow floor $I_{\text{floor,ind}}^{\text{h}} = 10$, compared with $I_{\text{floor}}^{\text{h}} = 5$ in the main configuration.

D.5 Alternative user-learning configuration

We additionally evaluate the alternative user-learning configuration described above, retaining behavioral evidence and adaptive threshold control.

Figure 32 compares both adaptive configurations with the static baselines. Both produce similar aggregate response patterns, while the alternative configuration has larger trajectory variability in harmful-user share and total benign welfare. The comparison illustrates governance outcomes under different user-learning and inflow configurations.

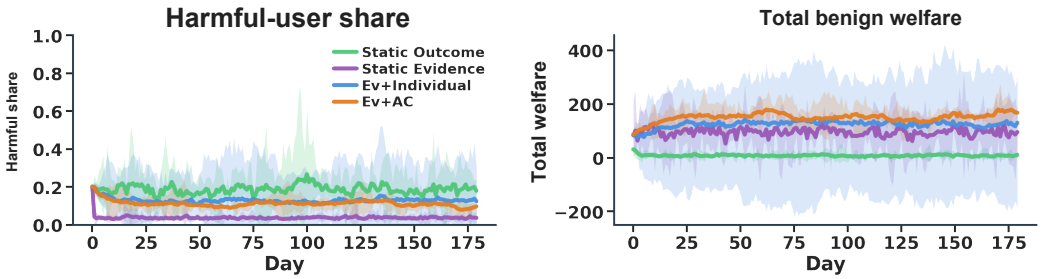


Figure 32: Non-competitive platform outcomes under alternative user learning. The plot compares the static baselines with the main and alternative adaptive configurations. Their parameter-sharing, grouping, and inflow settings are specified in Appendix D.4.

D.6 Hyperparameters

Table 10 summarizes the simulator, population-flow, routing, and learning parameters used in the non-competitive open-platform experiments.

Table 10: Main simulator and training parameters for the non-competitive open-platform experiments.

Category	Parameter	Value
Population initialization	Initial benign population $N_{\text{b},0}$	220
	Initial harmful-user population $N_{\text{h},0}$	56
	Initial removed pool R_0	0
Sequence environment	Sequence length L_{seq}	12
	Latent state size $\lvert\mathcal{Z}\rvert$	8

Category	Parameter	Value
	Action size $\lvert\mathcal{A}\rvert$	5
	Number of context modes $\lvert\mathcal{C}\rvert$	3
	Context probabilities $\boldsymbol{\omega}_{\mathcal{C}}$	$(0.25, 0.50, 0.25)$
	Context pressures $(\chi(1), \chi(2), \chi(3))$	$(0.00, 0.05, 0.12)$
Evaluation	Training horizon H_{train}	320 days
	Evaluation horizon H_{eval}	180 days
	Number of seeds N_{seed}	30
	Initial seed and stride	$42,\ 300$
Routing	Routing sharpness β_{route}	2.2
	Dynamic threshold range $[\theta_{\min}, \theta_{\max}]$	$[-25, 10]$
	Static outcome threshold θ_{out}	0.82
	Static reference-evidence threshold θ_{ref}	-25
Welfare	Occupancy weight $\lambda_{\text{occ}}^{\text{eco}}$	0.85
	Surviving harmful-user payoff weight $\lambda_{\text{h}}^{\text{eco}}$	0.70
	False-routing weight $\lambda_{\text{FP}}^{\text{eco}}$	0.60
	Welfare clipping range $[W_{\min}, W_{\max}]$	$[-1, 1]$
Benign churn	Churn range $(d_{\min}^{\text{b}}, d_{\max}^{\text{b}})$	$(0.00, 0.20)$
	Churn threshold τ_{out}	0.40
	Churn scale s_{out}	0.40
Benign inflow	Inflow range $(I_{\min}^{\text{b}}, I_{\max}^{\text{b}})$	$(5, 50)$
	Inflow threshold τ_{in}	0.30
	Inflow scale s_{in}	0.35
Harmful-user inflow	Inflow range $(I_{\min}^{\text{h}}, I_{\max}^{\text{h}})$	$(1, 5)$
	Inflow threshold τ_{h}	0.50
	Inflow scale s_{h}	0.50
	Context coefficient λ_c^{h}	0.25
	Daily inflow floor $I_{\text{floor}}^{\text{h}}$	5
	Inflow sampling	Poisson
Harmful-user churn	Churn range $(d_{\min}^{\text{h}}, d_{\max}^{\text{h}})$	$(0.00, 0.20)$
	Churn threshold $\tau_{\text{churn}}^{\text{h}}$	0.60
	Churn scale $s_{\text{churn}}^{\text{h}}$	0.10
Removed pool	Removed-pool decay q_R	0.08
Learning	Hidden dimension d_{hid}	64
	Learning rate η_{lr}	3×10^{-3}
	Entropy coefficient λ_{ent}	0.01
	Value-loss coefficient λ_{val}	0.5
	Gradient clipping norm $G_{\max}$	5.0
	Controller initial std $\sigma_{\theta,0}$	0.35
Harmful-user policy	Beta concentration range $[\kappa_{\min}^{\text{Beta}}, \kappa_{\max}^{\text{Beta}}]$	$[1.5, 18.0]$
	Numerical clipping ϵ_{num}	10^{-4}
	Initial target evidence $e_{\text{init}}^{\text{h}}$	-35
	Initial concentration fraction $\rho_{\text{init}}^{\text{h}}$	0.99
	Population pseudo-groups G_{pseudo}	64
Individual variant	Individual harmful-user agents G_{ind}	64
	Grouping mode	Legacy splitting
	Individual inflow floor $I_{\text{floor,ind}}^{\text{h}}$	10