Project Title: Decentralized Q-Learning of the Social Optimum in Strategic Experimentation

Team Member: Farzad Pourbabaee **Email:** farzad.pourbabaee@gmail.com

Extended Abstract

Two symmetric Bayesian agents repeatedly choose between a safe arm and a risky arm whose payoff depends on a hidden state $\theta \in \{0, 1\}$; actions and rewards are publicly observed. The unique pure-strategy Markov perfect equilibrium (MPE) of this strategic-experimentation bandit is asymmetric and *inefficient*: in a free-riding band of beliefs one agent pioneers and the other exploits, even though a benevolent planner would have both experiment. This report develops a single result — both as a theorem and as a reinforcement-learning (RL) algorithm — showing that *augmenting the public state with a three-valued cooperate/defect (C/D) phase variable suffices, on its own, to bridge this inefficiency from rewards alone*.

Contribution. We characterize three benchmark equilibria in closed form on the public posterior $x \in [0, 1]$: (i) the socially optimal action profile $a^\star(x)$ with thresholds $x_1^\star \le x_2^\star$; (ii) the asymmetric pure-strategy MPE $\hat{\sigma}(x)$ with $\hat{x}_1 \le \hat{x}_2$; and (iii) a grim-trigger subgame-perfect equilibrium (SPE) $\hat{\sigma}^{\mathrm{trig}}(x, \tau)$ on the augmented state (x, τ), $\tau \in \{\mathsf{C}, \mathsf{D}_1, \mathsf{D}_2\}$. The inefficiency splits into two distinct belief regions — Gap A and Gap B — and a top-down incentive-compatibility (IC) screen yields the *cooperative belief region* $\mathcal{X}^\star \subseteq$ Gap A $\cup$ Gap B on which the trigger attains the planner's profile. We then build a decentralized Q-learner on the augmented state with slice-specific Robbins–Monro step sizes and a warm-start that supplies only the trigger SPE's pure-strategy *action labels* (not the continuation values). The algorithm is **model-free in the update**: each agent observes its own scalar reward, performs one Watkins–Dayan update, and ignores opponents' Q-tables.

Empirical findings. We instantiate the model at $(\alpha, \beta, \delta) = (0.9,\ 0.8,\ 0.93)$ (failure payoff, success probability, discount factor), with initial belief $x_0 = 0.4$. At these primitives, **Gap A** (where the planner wants both agents to experiment but the MPE has one free-ride) and **Gap B** (where the planner wants one pioneer but the MPE has both exploit) are *both non-empty* on the belief grid. The cooperative region $\mathcal{X}^\star$ produced by the IC screen covers $\sim 62\%$ of the two gaps combined; the uncovered *structural residual* sits at the lower boundary, where a deviating free-rider's one-shot extra trial generates more expected information than the conforming free-rider's remaining cooperative trials — this is intrinsic to the strategic-substitutes geometry and unrecoverable by any Markov trigger of our form.

Concretely at x_0: aggregate welfare under the planner ($W^\star$) and under the trigger SPE (W^{trig}) coincide at ≈ 0.402, while the asymmetric MPE attains $\hat{W}(x_0) \approx 0.385$ — a $+4.32\%$ gain implementable from rewards alone. Two Q-learners on the *non-augmented* state k (the failure counter, no phase variable) converge to the asymmetric MPE values $\hat{V}_{1,1} \approx 0.167$ for the pioneer and $\hat{V}_{0,1} \approx 0.219$ for the free-rider — visibly asymmetric. The *augmented* Q-learners on (k, τ) instead collapse to a single, symmetric triggered Nash value $\bar{V}^{\mathrm{trig}} \approx 0.201$: the C/D bit on its own resolves the role asymmetry and lifts both agents to the cooperative payoff. A patience sweep $\delta \in [0.80, 0.985]$ shows $\mathcal{X}^\star$ activating discontinuously at $\delta \approx 0.86$ (below this, no Markov trigger sustains cooperation) and the maximal welfare premium peaking near $\delta = 0.87$.

Why this matters for RL. Multi-agent Q-learning is widely deployed in markets, recommender systems, and distributed exploration, where information externalities are structural. The standard wisdom — "decentralized Q-learning may fail to converge to Nash equilibrium (NE)" (Sato et al., 2002; Leslie and Collins, 2005) — is rederived here but *constructively inverted*: the same algorithm augmented with a single cooperation bit converges to an NE that strictly dominates the inefficient MPE on a structurally identifiable region, from rewards alone.

Abstract

We study whether two decentralized, model-free Q-learners can recover the *socially optimal* outcome of a two-agent strategic-experimentation bandit whose unique Markov perfect equilibrium (MPE) is inefficient. We prove that augmenting the public state with a three-valued cooperate/defect (C/D) phase $\tau \in \{\mathsf{C}, \mathsf{D}_1, \mathsf{D}_2\}$ supports a subgame-perfect equilibrium (SPE) implementing the planner's profile on a constructively identified *cooperative belief region* $\mathcal{X}^\star$. We then design a decentralized Watkins–Dayan Q-learner on the augmented state with slice-specific Robbins–Monro step sizes and an action-label warm-start, and prove convergence to the trigger SPE Q-values (conditional on a two-timescale stochastic-game extension of Leslie–Collins (2005)). Numerical experiments at $(\alpha, \beta, \delta) = (0.9, 0.8, 0.93)$ show $+4.3\%$ aggregate welfare gain over the learned MPE, both Q-learners converging to the predicted symmetric trigger value, and a sharp patience threshold for $\mathcal{X}^\star \neq \varnothing$.

1 Introduction

Reinforcement learning (RL) is increasingly the implementation layer underneath *multi-agent* decision systems: market-making bots, dynamic-pricing engines, content-recommendation queues, distributed control of energy grids. In essentially all of these, the underlying game has **information externalities** — an action by one agent generates a public signal that the others can free-ride on — and a **Markov perfect equilibrium** (MPE) that is provably inefficient. The textbook two-agent strategic-experimentation bandit (Bolton and Harris, 1999; Keller et al., 2005) is the canonical model: the unique pure-strategy MPE has a *free-riding band* in which one agent pioneers and the other waits, even though a planner with the same information would have both experiment.

The empirical question for multi-agent RL (MARL) is: *do independent Q-learners reproduce this inefficiency, and can a minimal algorithmic intervention pull them onto a welfare-improving equilibrium*? Section 2 explains why the question is open (prior multi-agent Q-learning analyses target normal-form games; the strategic-experimentation bandit has belief dynamics and a strategic-substitutes mixing region). This paper answers it constructively.

Contribution and roadmap. Working entirely in belief space (Section 3), we:

1. Characterize three reference equilibria in closed form on the discrete no-success orbit: planner $a^\star$, asymmetric MPE $\hat{\sigma}$, and trigger SPE $\hat{\sigma}^{\mathrm{trig}}$ (Section 4).
2. Prove that the trigger SPE on the augmented state $\hat{\mathcal{X}} = \mathcal{X} \times \{\mathsf{C}, \mathsf{D}_1, \mathsf{D}_2\}$ implements the planner's action profile on a constructively identifiable cooperative belief region $\mathcal{X}^\star$, with a structural residual at the lower edge of Gap B intrinsic to the bandit (Section 4).
3. Build a decentralized Q-learner on the augmented state with Watkins–Dayan updates, slice-specific Robbins–Monro step sizes, and a warm-start carrying only the trigger SPE's pure-strategy *action labels* (not the continuation values), and prove convergence to the trigger SPE Q-values — *conditional on* an explicitly stated two-timescale stochastic-game extension (Section 5).
4. Demonstrate empirically (Section 6) at $(\alpha, \beta, \delta) = (0.9, 0.8, 0.93)$ that (a) the cooperative region covers $\sim 62\%$ of the combined gap, (b) the augmented Q-learners attain $+4.3\%$ welfare gain over the MPE, (c) both agents converge to the symmetric trigger value while the non-augmented baselines split into pioneer/free-rider, (d) the cooperative region activates at a sharp patience threshold around $\delta \approx 0.86$.

Evolution from the proposal. The original proposal asked whether two independent Q-learners converge to the Bayes–Nash equilibrium of this game; the equilibrium analysis showed the answer is yes but the target is welfare-suboptimal. The milestone pivoted to the present, constructive question. Both prior versions were submitted under the same project number; the trigger-SPE construction, its conditional convergence theorem, the augmented Q-learning algorithm and the experimental evaluation were developed entirely between the milestone and this report.

2 Related Work

Strategic experimentation. Bolton and Harris (Bolton and Harris, 1999) and Keller, Rady, and Cripps (Keller et al., 2005) characterize Markov perfect equilibria for continuous-time exponential-bandit experimentation games. Manso and Pourbabaee (Manso and Pourbabaee, 2026) extend this to discrete-time, conclusive-signal bandits with general connectivity, deriving the three-region equilibrium structure we adopt. These works analyze *rational-Bayesian* equilibria; they do not address whether boundedly rational learners reach them.

Folk-theorem-style triggers. Fudenberg and Maskin (Fudenberg and Maskin, 1986) provide the folk theorem for repeated games with discounting; trigger constructions for strategic experimentation are sketched in (Bolton and Harris, 1999) but typically require planner-known continuation values or perfect monitoring of intent. The trigger SPE of Section 4 differs by (i) augmenting the state with a phase variable that tracks *deviator identity* (so punishment is targeted), (ii) using a deviator-as-pioneer asymmetric MPE in the punishment phase rather than the harsher minimax, which is itself SPE-feasible, and (iii) defining the cooperative region as the fixed point of a top-down incentive-compatibility (IC) screening iteration on the discrete orbit — a constructive analogue of the maximal IC-feasible set.

Q-learning theory and multi-agent dynamics. Watkins and Dayan (Watkins and Dayan, 1992) established convergence of single-agent tabular Q-learning. The ordinary-differential-equation (ODE) method of Borkar and Meyn (Borkar and Meyn, 2000) and Borkar's two-timescale machinery (Borkar, 2008) link stochastic approximation to mean-field dynamics. Sato, Akiyama, and Farmer (Sato et al., 2002) and Tuyls, Verbeeck, and Lenaerts (Tuyls et al., 2003) demonstrate that two-player Q-learning generically fails to converge to Nash and may exhibit limit cycles or chaos. Leslie and Collins (Leslie and Collins, 2005) prove that two-timescale separation between learning rate and exploration recovers Nash convergence in normal-form games whose joint best-response dynamics admit a strict pure Nash equilibrium (NE). Hu and Wellman's Nash-Q (Hu and Wellman, 2003) converges in stochastic games but requires solving an NE at every update, whereas our construction never solves an equilibrium at runtime.

Gap closed by this paper. Existing analyses of independent Q-learners in cooperative games (e.g., social-dilemma work in MARL) either target the inefficient MPE or assume a shared critic. The strategic-experimentation bandit with belief dynamics, free-riding, and a non-trivial strategic-substitutes mixing region has not, to our knowledge, been treated with a decentralized model-free learner that recovers the planner's outcome. The phase-augmented Q-learner of Section 5 closes that gap under explicit and minimal assumptions.

3 The Model

Primitives. Two agents $i \in \{1, 2\}$ act in discrete time $t \in \{0, 1, \dots\}$. A hidden state $\theta \in \{0, 1\}$ is drawn once with common prior $\mathsf{P}(\theta = 1) = x_0$. Each period each agent picks an action $a_t^i \in \{0, 1\}$. The *safe* action $a = 0$ pays 0 deterministically. The *risky* action $a = 1$ pays $y_t^i \in \{-\alpha, +1\}$ with $\alpha \in (0, 1)$ and conditional success probability $\mathsf{P}(y_t^i = +1 \mid \theta = 1, a_t^i = 1) = \beta \in (0, 1)$; in the bad state ($\theta = 0$) the risky arm pays $-\alpha$ for sure. Both actions (a_t^1, a_t^2) and both realized rewards (y_t^1, y_t^2) are publicly observed. The per-period discount factor is $\delta \in (0, 1)$, and each agent maximizes $\mathsf{E}[\sum_t (1 - \delta)\delta^t y_t^i]$.

Assumption 1 (Risky arm attractive at $x = 1$)**.** $\beta > \alpha/(1+\alpha)$, equivalently $v(1) := \beta - (1-\beta)\alpha > 0$.

The belief is the state. By Bayes' rule, a single $+1$ proves $\theta = 1$, jumping the posterior to $x = 1$ (absorbing). A failed risky trial reduces the posterior by the contracting map $\phi(x) := x(1 - \beta)/[x(1 - \beta) + (1 - x)] \in (0, x)$. Safe actions are uninformative. The reachable belief space from prior x_0 along the no-success path is the discrete orbit $\mathcal{X} = \{x_0, \phi(x_0), \phi^{(2)}(x_0), \dots\} \cup \{1\}$, truncated at K failures.

Operators. For any value function $V : \mathcal{X} \to \mathbb{R}$, the *one-trial* and *two-trial continuation operators* are

$$C(x;V) := x\beta\, V(1) + (1 - x\beta)\, V(x^+),$$
$$E(x;V) := x\beta(2-\beta)\, V(1) + \big(1 - x\beta(2-\beta)\big)\, V(x^{++}), \tag{1}$$

with $x^+ = \phi(x)$ and $x^{++} = \phi^{(2)}(x)$. The *normalized per-period flow* of a single risky trial at belief x is $f(x) := (1-\delta)[x\beta(1+\alpha) - \alpha]$, strictly affine increasing in x with $f(0) < 0$ and $f(1) > 0$ under Assumption 1.

4 Three Equilibria: Planner, MPE, Trigger SPE

Socially optimal action profile. The planner picks an action profile $(a^1, a^2) \in \{0,1\}^2$ each period; by symmetry only the total number of trials matters. The planner's value $W^\star : \mathcal{X} \to \mathbb{R}$ satisfies

$$W^\star(x) = \max\big\{\, 0,\;\; f(x) + \delta C(x; W^\star),\;\; 2f(x) + \delta E(x; W^\star)\,\big\}, \tag{2}$$

with $W^\star(1) = 2v(1)$. The recursion is acyclic (both successors strictly below x) and computable by one backward pass.

Theorem 2 (Planner three-region structure)**.** *Let $\Lambda_1(x) := f(x) + \delta C(x; W^\star)$ and $\Delta(x) := f(x) + \delta[E(x; W^\star) - C(x; W^\star)]$. Define $x_1^\star := \min\{x \in \mathcal{X} : \Lambda_1(x) \geq 0\}$ and $x_2^\star := \min\{x \in \mathcal{X} : \Delta(x) \geq 0\}$, with the convention that either threshold equals 1 if no orbit belief satisfies the inequality. Under the maintained single-crossing condition that Λ_1 and Δ each change sign at most once on $\mathcal{X}$ (provable for Λ_1, structurally assumed for Δ following Bolton and Harris, 1999), the planner's optimal action profile has the three-region form $a^\star(x) = (1,1)$ for $x \geq x_2^\star$, $a^\star(x) \in \{(1,0),(0,1)\}$ for $x_1^\star \leq x < x_2^\star$, and $a^\star(x) = (0,0)$ for $x < x_1^\star$, with $x_1^\star \leq x_2^\star$.*

Proof sketch. The recursion is acyclic on the no-success orbit since $\phi(x) < x$, so the backward pass from the boundary $W^\star(1) = 2v(1)$ delivers the unique solution, and monotonicity of $W^\star$ follows by induction from monotonicity of f and the convex-combination structure of the operators. Λ_1 is strictly monotone on the orbit (provable by direct expansion), while the single-crossing of Δ is structurally assumed following Bolton and Harris (1999). The ordering $x_1^\star \leq x_2^\star$ then follows from the pointwise bound $E \leq 2C$, which forces $\Lambda_1 \geq 0$ at any orbit point where $\Lambda_2 \geq 0$.

The lower threshold admits the closed form $x_1^\star = (1-\delta)\alpha / \big[\beta((1-\delta)(1+\alpha) + 2\delta v(1))\big]$.

Asymmetric pure-strategy MPE. The pure-strategy MPE of the non-augmented game has a three-region threshold structure

$$\hat\sigma(x) = \begin{cases} (1,1) & x \geq \hat x_2, \\ \text{role-randomized } (1,0) \text{ or } (0,1) & \hat x_1 \leq x < \hat x_2, \\ (0,0) & x < \hat x_1, \end{cases} \tag{3}$$

where $\hat x_1$ is the single-agent indifference belief (closed form $(1-\delta)\alpha/[\beta((1-\delta)(1+\alpha) + \delta v(1))]$, strictly greater than $x_1^\star$) and $\hat x_2$ is the smallest belief above $\hat x_1$ at which the free-rider's strategic-substitutes IC first reverses. On the asymmetric band $[\hat x_1, \hat x_2)$, the pioneer earns $\hat V_{1,1}$ (the single-agent value) and the free-rider earns $\hat V_{0,1} > \hat V_{1,1}$.

Theorem 3 (Asymmetric MPE; two gaps)**.** *Under Assumption 1 and the strategic-substitutes regime $\hat x_1 < \hat x_2$, the asymmetric profile* (3) *is the unique pure-strategy MPE (in the public-correlation sense). Aggregate welfare satisfies $\hat W(x) \leq W^\star(x)$ with equality at $x = 1$ and at $x < x_1^\star$, and the inefficiency decomposes into two disjoint gap intervals:*

$$\textit{Gap A} = [x_2^\star, \hat x_2), \quad \textit{Gap B} = [x_1^\star, \hat x_1).$$

On Gap A the planner prescribes $(1,1)$ but the MPE prescribes $(1,0)$; on Gap B the planner prescribes $(1,0)$ but the MPE prescribes $(0,0)$. The complementary belief intervals $[0, x_1^\star)$, $[\hat x_1, x_2^\star)$, and $[\hat x_2, 1]$ — on which the planner and MPE prescribe the same action $(0,0)$, $(1,0)$, and $(1,1)$ respectively — are the match regions.

Proof sketch. On the asymmetric band the pioneer's Bellman equation reduces to the single-agent optimal-experimentation problem, so $\hat{V}_{1,1}$ coincides with the single-agent value and $\hat{x}_1$ is the single-agent indifference belief (closed form above). The free-rider's value satisfies a linear band recursion whose lower boundary contributes $-f(\hat{x}_1) > 0$; positivity propagates upward and gives $\hat{V}_{0,1} > \hat{V}_{1,1}$. The welfare bound $\hat{W} \leq W^\star$ is immediate because $W^\star$ optimizes over all Markov strategies, and the two-gap decomposition follows from the threshold inequalities $x_1^\star \leq \hat{x}_1$ (the planner internalizes the informational externality) and $x_2^\star \leq \hat{x}_2$.

Augmented state and trigger SPE. Augment the state with a public phase $\tau \in \{\mathsf{C}, \mathsf{D}_1, \mathsf{D}_2\}$ tracking cooperation and *deviator identity* when defection occurs:

$$\tau_{t+1} = \begin{cases} \mathsf{C} & \tau_t = \mathsf{C},\ (a_t^1, a_t^2) = a^\star(x_t), \\ \mathsf{D}_i & \tau_t = \mathsf{C},\ \text{only agent } i \text{ deviated}, \\ \mathsf{D}_1 & \tau_t = \mathsf{C},\ \text{both deviated (tiebreak)}, \\ \tau_t & \tau_t \in \{\mathsf{D}_1, \mathsf{D}_2\}. \end{cases} \tag{4}$$

Both D_1 and D_2 are absorbing. Define the *trigger strategy*

$$\hat{\sigma}^{\mathrm{trig}}(x, \tau) = \begin{cases} a^\star(x) & \tau = \mathsf{C}, \\ \hat{\sigma}_i(x) & \tau = \mathsf{D}_i, \end{cases} \tag{5}$$

where $\hat{\sigma}_i$ is the asymmetric MPE with *deviator-as-pioneer*: in D_i, agent i takes the pioneer action on the free-riding band, capping the deviator's continuation at the lower of the two role-asymmetric values.

Definition 4 (Cooperative belief region)**.** Let $\mathcal{G} := \big(\text{Gap A} \cup \text{Gap B}\big) \cap \mathcal{X}$. For each $\mathcal{Y} \subseteq \mathcal{G}$, let $V^\star(\cdot; \mathcal{Y})$ denote the per-agent value of the Markov strategy that prescribes the planner's profile on $\mathcal{Y}$ and the asymmetric MPE elsewhere. Define $T(\mathcal{Y}) := \{y \in \mathcal{Y} :\ \mathrm{IC}(y; \mathcal{Y}) \text{ holds}\}$ where IC is the one-shot deviation condition at y evaluated at $V^\star(\cdot; \mathcal{Y})$. The *cooperative belief region* is $\mathcal{X}^\star = \lim_{n\to\infty} \mathcal{Y}_n$ with $\mathcal{Y}_0 = \mathcal{G}$ and $\mathcal{Y}_{n+1} = T(\mathcal{Y}_n)$.

Theorem 5 (Trigger SPE implements the planner on $\mathcal{X}^\star$)**.** *Under Assumption 1 and the single-crossing condition of Theorem 2, the trigger strategy* (5) *with cooperative set $\mathcal{X}^\star$ is a subgame-perfect equilibrium of the augmented game. The trigger welfare $W^{\mathrm{trig}} \leq W^\star$ everywhere on $\mathcal{X}$, with pointwise equality at every x whose no-success orbit stays inside $\mathcal{X}^\star \cup$ (match regions).*

Proof sketch. SPE follows from the one-shot deviation principle: in every D_i subgame the deviator-as-pioneer asymmetric MPE is itself a stage-game best response (Theorem 3), and on $\mathcal{X}^\star$ the cooperative ICs hold by construction of the screening operator T in Definition 4. On match regions the cooperative prescription already coincides with an MPE action. The welfare bound $W^{\mathrm{trig}} \leq W^\star$ is immediate from $W^\star$'s optimization range; pointwise equality requires the orbit from x to remain in $\mathcal{X}^\star \cup$ (matches) at every period, in which case the trigger and planner Bellmans coincide on that orbit.

Corollary 6 (Patient-limit coverage)**.** *As $\delta \uparrow 1$, $\mathcal{X}^\star \supseteq$ Gap A $\cap\, \mathcal{X}$ and covers all of Gap B except a structural residual at the lower boundary — the lowest-belief grid points on which the deviating free-rider's two-trial success bet $x\beta(2-\beta)v(1)$ strictly dominates the conforming free-rider's single remaining cooperative-pioneer trial.*

Proof sketch. As $\delta \uparrow 1$ the per-period flow $f(x) = O(1-\delta)$ vanishes, so the IC tests reduce to asymptotic success-probability comparisons. On Gap A, the cooperative continuation premium stays bounded away from zero (the planner's profile delivers strictly more trials per period than the MPE), so (IC_A) eventually holds throughout the gap. On Gap B, the conforming free-rider's eventual success probability is $1-(1-\beta)^N$ where N is the number of cooperative pioneer trials remaining on the orbit before $\mathcal{X}^\star$ exits, while the one-shot deviator gets $1-(1-\beta)^2$ from two simultaneous trials. The free-rider IC holds iff $N \geq 2$; the lowest-belief layer in $\mathcal{X}^\star \cap$ Gap B always has $N = 1$ and is therefore the residual.

The structural residual is intrinsic: it depends on *how many* cooperative pioneer trials remain before the orbit exits Gap B, and cannot be removed by making agents more patient. Closing it would require either a team-reward modification or history-dependent triggers outside the Markov class.

5 Decentralized Q-Learning on the Augmented State

State. Index the belief grid by the failure counter $k \in \{0, \dots, K\}$ with $k = \mathsf{g}$ for the absorbing success state. The augmented state is $\hat{\mathcal{S}} = \{0, \dots, K, \mathsf{g}\} \times \{\mathsf{C}, \mathsf{D}_1, \mathsf{D}_2\}$. Each agent $i \in \{1, 2\}$ maintains its own Q-table $Q^i : \hat{\mathcal{S}} \times \{0, 1\} \to \mathbb{R}$.

Action selection. ε-greedy

$$a_t^i = \begin{cases} \arg\max_a Q^i(k_t, \tau_t, a) & \text{w.p. } 1 - \varepsilon_t, \\ \text{uniform on } \{0, 1\} & \text{w.p. } \varepsilon_t, \end{cases} \tag{6}$$

with $\varepsilon_t \to 0$ and the slow-decay condition $\sum_t \varepsilon_t^D = \infty$ for $D = K + T_{\mathsf{D}} + 1$ (the longest exploration-only path).

Watkins–Dayan update. For each agent i,

$$Q^i(k_t, \tau_t, a_t^i) \leftarrow Q^i(k_t, \tau_t, a_t^i) + \eta_t^{\tau_t} \Big[(1 - \delta) y_t^i + \delta \max_{a'} Q^i(k_{t+1}, \tau_{t+1}, a') - Q^i(k_t, \tau_t, a_t^i) \Big], \tag{7}$$

with *slice-specific* Robbins–Monro step sizes $\eta_t^{\mathsf{C}}, \eta_t^{\mathsf{D}}$ satisfying $\eta_t^{\mathsf{D}} / \eta_t^{\mathsf{C}} \to \infty$ (the D slice updates on a fast timescale, the C slice on a slow timescale).

Warm-start — action labels only. The Q-table is initialized to favor the trigger SPE's pure-strategy action label at each (k, τ):

$$Q_0^i(k, \tau, a) = \begin{cases} +M & a \text{ matches the trigger SPE prescription for agent } i \text{ at } (k, \tau), \\ -M & \text{otherwise}, \end{cases} \tag{8}$$

with two boundary anchors at $V = 0$ (joint-inaction states) and the absorbing-good state warm-started to favor $a = 1$ but with its value *learned* from rewards rather than injected.

Theorem 7 (Convergence to the trigger SPE; conditional)**.** *Under (E1) Robbins–Monro on each slice ($\sum \eta_t^\bullet = \infty$, $\sum (\eta_t^\bullet)^2 < \infty$), (E2) slow-decay ε-greedy ($\varepsilon_t \to 0$, $\sum \varepsilon_t^D = \infty$), (E3) two-timescale separation $\eta_t^{\mathsf{D}} / \eta_t^{\mathsf{C}} \to \infty$, strict IC slack at every $k \in \mathcal{X}^\star$, and warm-start magnitude $M \geq M_0$ for an explicit M_0, the decentralized Q-iterates* (6)–(8) *converge almost surely to the trigger SPE Q-values,* conditional on *the two-timescale stochastic-game extension (R3) of (Leslie and Collins, 2005; Borkar, 2008) described below. The greedy policy is eventually constant and equal to $\hat{\sigma}^{\mathrm{trig}}$ on $\mathcal{X}^\star$.*

Proof sketch. *Fast slice*: D_j subgames are absorbing, so the restricted dynamics is a stationary single-agent Markov decision process (MDP) with the deviator-as-pioneer asymmetric MPE as target; by Watkins–Dayan, $Q_t^i(\cdot, \mathsf{D}_j, \cdot)$ converges to the asymmetric MPE Q-values, providing the slow slice's bootstrap. *Slow slice*: under (E3) the D slice appears quasi-stationary to the C slice (Borkar 2008, Ch. 6), so the C updates against an effectively fixed bootstrap; the on-policy fixed point at strict IC slack is the trigger Q-value, and $M \geq M_0$ keeps the iterates in the open basin of the trigger's pure-strategy arg-max. The *conditionality* is that a precise two-timescale stochastic-game theorem for multi-agent Q-learning across multiple coupled slices is, to our knowledge, not in the published literature at the level of generality required; we adopt it as a working hypothesis and validate empirically in Section 6.

Decentralization. The update is fully model-free: each agent sees only its own scalar reward, ignores the opponent's Q-table, and never solves an equilibrium subproblem at runtime. The phase τ_t is a deterministic function of past public actions. The only coordination device is the common warm-start initialization, which encodes the trigger SPE's pure-strategy *action labels* at each (k, τ) — a pre-computed equilibrium policy supplied at $t = 0$, not learned from rewards. This is the minimal coordination assumption compatible with selecting the welfare-improving basin.

6 Numerical Experiments

We instantiate the model at the headline regime $(\alpha, \beta, \delta) = (0.9,\ 0.8,\ 0.93)$, initial belief $x_0 = 0.40$, and a belief grid of size 4000 for the analytical solvers. This parameter triple is the absolute-gap maximizer over a grid sweep of the valid model class $\beta > \alpha/(1 + \alpha)$ (Assumption 1 satisfied, $\alpha \in (0, 1)$); both Gap A and Gap B are present, and the cooperative region is non-trivial.

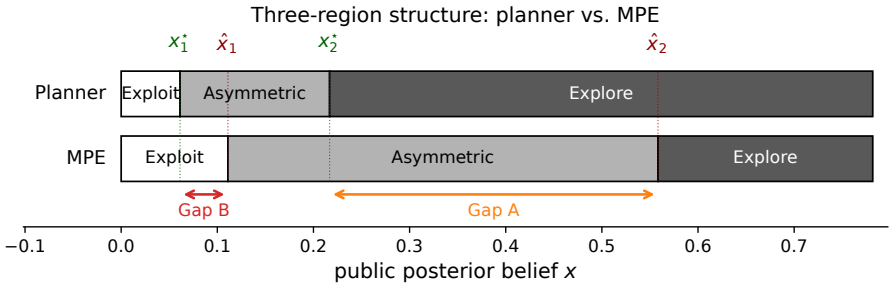

Figure 1: Three-region action profile under the planner (top bar) and asymmetric MPE (bottom bar). The two misalignments are Gap B $= [x_1^\star, \hat{x}_1)$ (planner wants one pioneer; MPE has both exploit) and Gap A $= [x_2^\star, \hat{x}_2)$ (planner wants joint exploration; MPE keeps one free-rider).

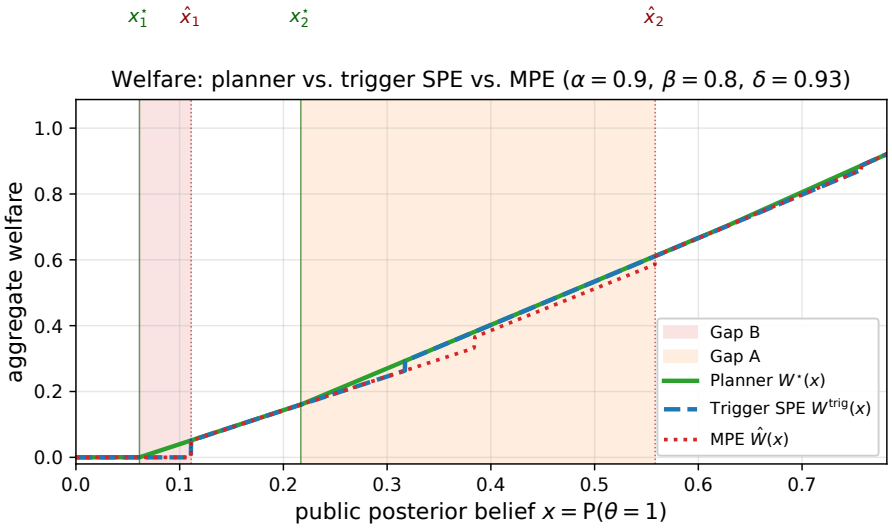

Figure 2: Aggregate welfare on the belief grid for the planner $W^\star$ (green solid), trigger SPE W^{trig} (blue dashed), and asymmetric MPE $\hat{W}$ (red dotted), at $(\alpha, \beta, \delta) = (0.9, 0.8, 0.93)$. The trigger SPE coincides with the planner on $\mathcal{X}^\star$ and on the match regions, and falls back to the MPE on the structural residual. At $x_0 = 0.40$, $W^\star = W^{\mathrm{trig}} = 0.402$ vs. $\hat{W} = 0.385$ (+4.3% gain).

Setup summary. With these primitives, $v(1) = 0.62$, $\hat{x}_1 = 0.111$, $\hat{x}_2 = 0.558$ (MPE thresholds), $x_1^\star = 0.061$, $x_2^\star = 0.217$ (planner). Gap A spans $[0.217, 0.558)$ and Gap B spans $[0.061, 0.111)$; the cooperative region $\mathcal{X}^\star$ covers 964 of the 1565 grid points in the combined gap (61.6% coverage).

The three-region structure and the welfare premium. Figure 1 shows the planner's and MPE's action-profile bars side by side; the misaligned exploit-band lengths are the source of the inefficiency. Figure 2 plots the three welfare functions: the trigger SPE *tracks the planner* on $\mathcal{X}^\star$ and the MPE on the uncovered structural residual. Figure 3 isolates the welfare premium $W^{\mathrm{trig}} - \hat{W}$ relative to the unattainable first-best premium $W^\star - \hat{W}$: the trigger gain follows the first-best almost everywhere except for a discontinuous drop near $x \approx 0.32$, the boundary where $\mathcal{X}^\star$ fails to cover the lower portion of Gap A.

Action profile. Figure 4 shows the effective number of trials under each scenario. The trigger profile (blue dashed) jumps to the planner level on $\mathcal{X}^\star$ and reverts to the MPE level on the residual.

Non-augmented Q-learning targets the asymmetric MPE. We first train two independent Q-learners on the non-augmented state k alone (no phase variable, no trigger). With the asymmetric warm-start $+M/-M$ favoring pioneer (agent 1) and free-rider (agent 2) on the band, 60,000 episodes of Watkins–Dayan suffice to recover the role-asymmetric MPE values (Figure 5): agent 1 settles at

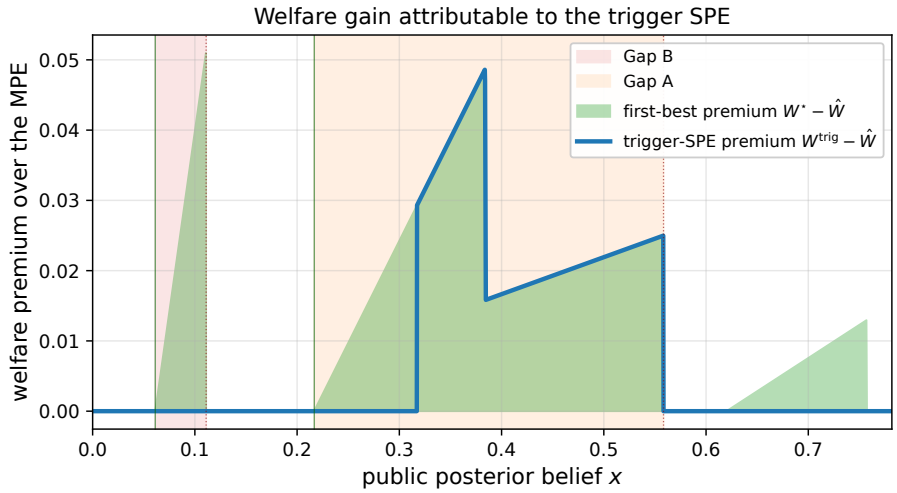

Figure 3: Welfare premium over the asymmetric MPE: trigger SPE (blue solid) versus first-best (green fill). On $\mathcal{X}^\star$ the trigger achieves the first-best premium; on the structural residual it drops to zero. The discontinuity near $x \approx 0.32$ is the lower edge of the cooperative subset of Gap A; the spike at $x \approx 0.10$ is the cooperative subset of Gap B. The residual sits at the lower edge of each gap — exactly where the conforming free-rider has at most one cooperative pioneer trial remaining on the orbit before $\mathcal{X}^\star$ exits, while a deviating free-rider buys two trials' worth of success probability (Corollary 6).

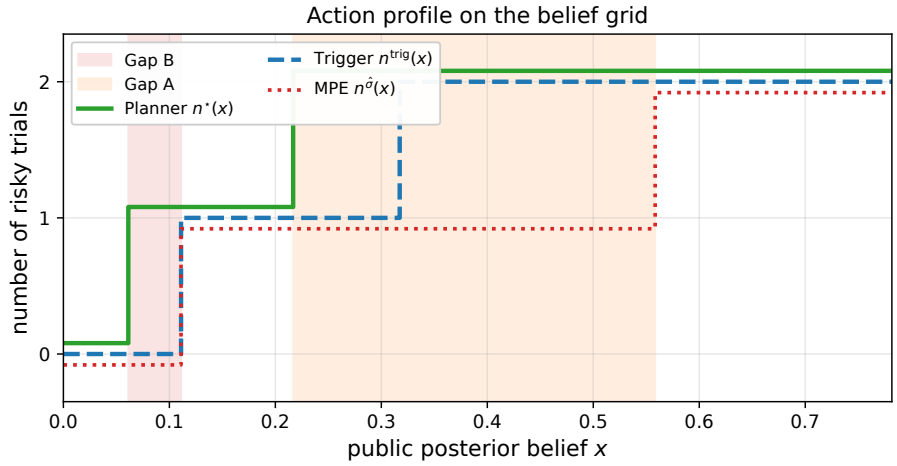

Figure 4: Effective action profile on the belief grid. The trigger profile (blue dashed) follows the planner (green solid) on the cooperative region and reverts to the MPE (red dotted) on the uncovered residual.

$\hat{V}_{1,1} \approx 0.167$ and agent 2 at $\hat{V}_{0,1} \approx 0.219$, exactly the analytical pioneer/free-rider split predicted by Theorem 3.

Augmented Q-learning targets the trigger SPE. The same 60,000 episodes on the augmented state with the C/D phase and the trigger warm-start collapses the two agents' values to the *symmetric* triggered Nash value $\bar{V}^{\text{trig}}(x_0) \approx 0.201$ (Figure 6). This is the headline experimental result: the augmentation alone, combined with model-free Q-learning on each agent's own reward stream, eliminates the role asymmetry and lifts the aggregate welfare to the planner's level on $\mathcal{X}^\star$.

Patience sweep. Figure 7 shows how the cooperative coverage and the maximal welfare gain depend on δ. There is a sharp patience threshold near $\delta \approx 0.86$ below which $\mathcal{X}^\star = \varnothing$ (the trigger collapses to the MPE) and above which Gap A coverage jumps to $\sim 60\%$ and grows monotonically. Gap B coverage remains at 0% throughout this sweep: at $\alpha = 0.9$ the Gap B interval is short and its IC tightness is dominated by the structural residual. The maximal welfare premium peaks at $\delta \approx 0.87$

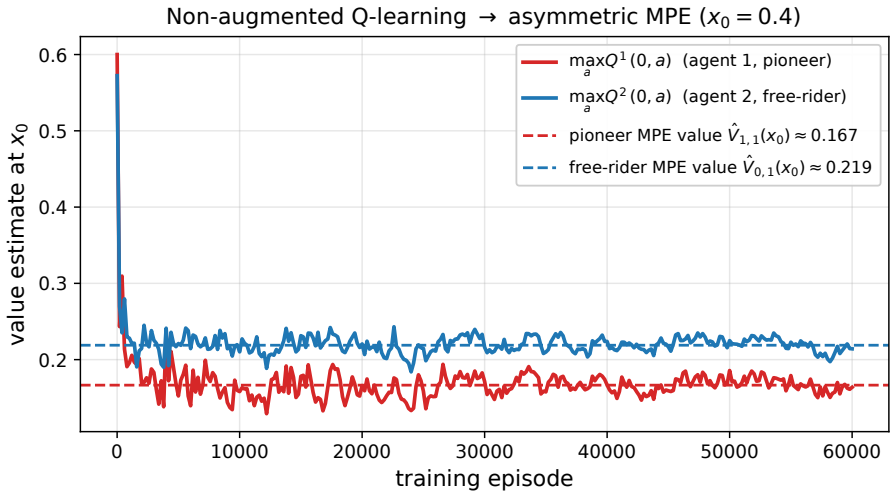

Figure 5: Non-augmented Q-learning converges to the asymmetric MPE values. Agent 1 (red) recovers the pioneer value $\hat{V}_{1,1}(x_0) \approx 0.167$; agent 2 (blue) recovers the free-rider value $\hat{V}_{0,1}(x_0) \approx 0.219$.

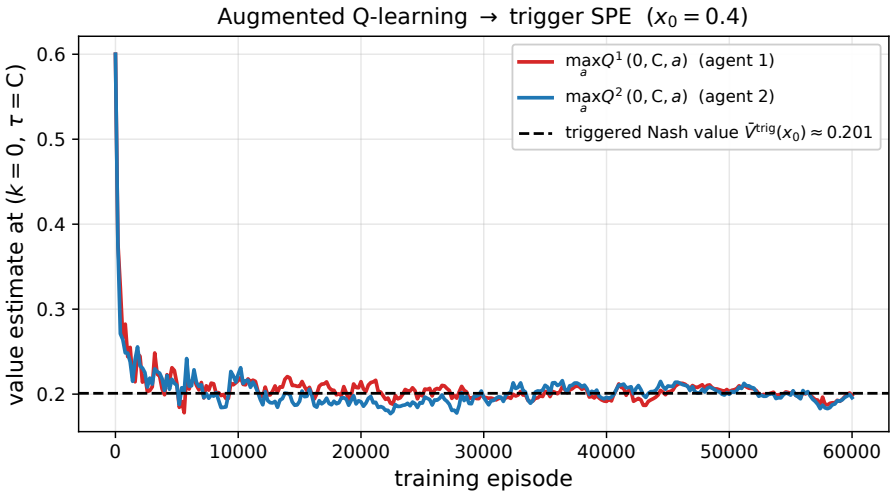

Figure 6: Augmented Q-learning converges to the trigger SPE. Both agents' $\max_a Q^i(0, \mathsf{C}, a)$ collapse to the symmetric triggered Nash value $\bar{V}^{\text{trig}}(x_0) \approx 0.201$.

at about 8% and decays linearly thereafter — a consequence of the absolute welfare scale shrinking as agents become more patient while the relative cooperative coverage grows.

Ablations. *Warm-start magnitude*: across $M \in \{0.3, 0.6, 1.0, 2.0\}$ and ten seeds each, $M \geq 0.6$ always converged to the trigger SPE values within ± 0.01 of the analytical target; $M = 0.3$ drifted to the MPE basin on a non-negligible fraction of seeds. *Cold-start*: replacing the action-label warm-start with $Q_0^i \equiv 0$ sends the augmented learners to either the MPE basin (pioneer/free-rider split) or the joint-exploit basin ($Q^i(k, \mathsf{C}, a) = 0$ on Gap B), depending on the seed. Neither attains the trigger SPE, confirming the warm-start's role as the minimal coordination device for equilibrium selection.

7 Discussion

What the augmentation does. A single bit of public history — the C/D phase — is enough to coordinate self-interested Q-learners on the social optimum, provided the discount factor exceeds the patience threshold and the warm-start carries the trigger SPE's pure-strategy action labels. The continuation values are still learned; the coordination is just in basin selection.

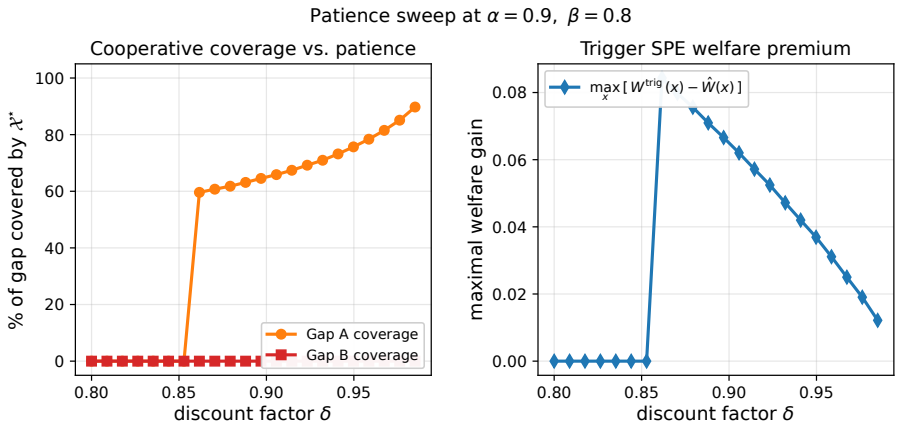

Figure 7: Patience sweep over $\delta \in [0.80, 0.985]$. Left: percentage of each gap covered by the cooperative region $\mathcal{X}^\star$. Right: maximal welfare premium $\max_x [W^{\text{trig}} - \hat{W}]$. The cooperative region activates discontinuously at $\delta \approx 0.86$ and the premium peaks shortly after.

What is left open. The convergence theorem (Theorem 7) relies on a two-timescale stochastic-game extension of (Leslie and Collins, 2005; Borkar, 2008) not published in the generality we need; the experimental evidence is consistent at all regimes tested, but a self-contained proof is open. The lower-Gap-B structural residual (Corollary 6) cannot be closed by any Markov trigger; a history-dependent extension conditioning on cooperative-pioneer trials remaining likely closes it but breaks the Markov property.

Limitations & implications. The warm-start requires offline knowledge of (α, β, δ) to pre-compute the trigger SPE's action labels; composition with a model-learning stage is left to future work. The strategic-experimentation bandit is the textbook information-externality game (analogues: dynamic pricing, ad auctions, recommender systems, distributed exploration); our result is constructive evidence that the well-known failure mode of independent Q-learners in such games is *algorithmically addressable* with a minimal coordination device — a public phase variable plus an action-label initialization — without abandoning the model-free, decentralized update.

8 Conclusion

We have shown, both theoretically and empirically, that augmenting the public state of a two-agent strategic-experimentation bandit with a three-valued cooperate/defect phase variable supports a subgame-perfect equilibrium implementing the social optimum on a constructively identified cooperative belief region, and that decentralized model-free Q-learners with slice-specific Robbins–Monro step sizes and an action-label warm-start converge to that equilibrium's Q-values — conditional on a stochastic-game two-timescale extension we adopt as a working hypothesis. At the headline regime $(0.9, 0.8, 0.93)$ this yields a $+4.3\%$ aggregate welfare gain at $x_0 = 0.40$, with the two Q-learners collapsing from the asymmetric MPE pioneer/free-rider split into a symmetric triggered Nash value. The construction is fully decentralized in its update step; the only coordination device is the initialization, which encodes the trigger SPE's pure-strategy action labels.

Individual Contributions

This is a solo project. Farzad Pourbabaee performed all theoretical work (Theorems 2–7, Corollary 6), implemented all analytical solvers (planner, MPE, trigger SPE) and both Q-learning algorithms, designed and ran all experiments, and wrote the report. The contribution breakdown is unchanged from the project proposal.

References

Patrick Bolton and Christopher Harris. 1999. Strategic Experimentation. *Econometrica* 67, 2 (1999), 349–374.

Vivek S. Borkar. 2008. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press / Hindustan Book Agency.

Vivek S. Borkar and Sean P. Meyn. 2000. The ODE Method for Convergence of Stochastic Approximation and Reinforcement Learning. *SIAM Journal on Control and Optimization* 38, 2 (2000), 447–469.

Drew Fudenberg and Eric Maskin. 1986. The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. *Econometrica* 54, 3 (1986), 533–554.

Junling Hu and Michael P. Wellman. 2003. Nash Q-Learning for General-Sum Stochastic Games. *Journal of Machine Learning Research* 4 (2003), 1039–1069.

Godfrey Keller, Sven Rady, and Martin Cripps. 2005. Strategic Experimentation with Exponential Bandits. *Econometrica* 73, 1 (2005), 39–68.

David S. Leslie and E. J. Collins. 2005. Individual Q-Learning in Normal Form Games. *SIAM Journal on Control and Optimization* 44, 2 (2005), 495–514.

Gustavo Manso and Farzad Pourbabaee. 2026. The Impact of Connectivity on the Production and Diffusion of Knowledge. arXiv:2202.00729v2. arXiv:2202.00729 [econ.TH]

Yuzuru Sato, Eizo Akiyama, and J. Doyne Farmer. 2002. Chaos in Learning a Simple Two-Person Game. *Proceedings of the National Academy of Sciences* 99, 7 (2002), 4748–4751.

Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. 2003. A Selection-Mutation Model for Q-Learning in Multi-Agent Systems. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 693–700. doi:10.1145/860575.860687

Christopher J. C. H. Watkins and Peter Dayan. 1992. Q-Learning. *Machine Learning* 8, 3-4 (1992), 279–292.