Optimism as a Vulnerability: Deceptive Stackelberg Control of UCB Bandit Followers

author names withheld

Under Review for NExT-Game 2026

Abstract

Upper Confidence Bound (UCB) algorithms guarantee sublinear regret for agents learning unknown stochastic environments, yet the same principle that makes them statistically efficient—optimism in the face of uncertainty—induces a predictable strategic vulnerability against an omniscient adaptive leader. Classical strong Stackelberg equilibrium (SSE) assumes that the follower immediately best-responds to the leader's committed mixed action; it therefore supplies no mechanism-design prescription for a leader facing a boundedly rational follower who constructs and acts on empirical reward histories. We formalize this conflict in a finite-horizon repeated Stackelberg game and give exact constructive proofs for a deceptive leader mechanism. In a honeypot phase, the leader pays a finite signaling cost to inflate the UCB index of a designated follower action. In a trap phase, the leader switches to a selfish action distribution while the follower remains locked into the designated action because the manipulated empirical history and exploration bonus dominate competing indices. Under explicit separation and payoff assumptions, the leader's cumulative utility strictly exceeds the classical SSE ceiling, and the manipulation cost is bounded by a regret calculation of order $O(\sqrt{T \ln T})$. The results identify a formal incompatibility between static equilibrium prescriptions and dynamically learned empirical incentives.

Keywords: Stackelberg games, bandit learning, UCB, strategic deception, reward manipulation

1. Introduction

Stackelberg security games and related leader–follower models typically analyze commitment to a mixed action followed by a rational best response [5, 10, 21, 22]. In the strong Stackelberg equilibrium convention, the follower breaks ties in favor of the leader; the leader consequently solves a static optimization problem over induced best responses. This model is internally coherent when the follower observes payoffs and responds as a utility maximizer. It is not a model of a follower who learns payoffs from interaction.

Bandit-learning followers instantiate a different behavioral primitive. UCB1 [1, 11, 20] is no-regret in stationary stochastic bandits: arms with uncertain value receive an optimism bonus, and suboptimal arms are sampled only logarithmically often. In a strategic environment, however, reward samples are not exogenous evidence about a fixed arm. They are data produced by another player. This places the model closer to learning in games and nonstationary multi-agent learning [4, 6, 17, 19] than to one-shot commitment. An adaptive leader can therefore treat the follower's statistical estimator as an object of control.

The closest technical literature is reward or action poisoning of bandit learners, where an external attacker corrupts feedback or actions to force target pulls at small perturbation cost [2, 8, 12, 13,

15, 23, 26]. Our mechanism differs because the leader does not edit rewards exogenously; it creates the reward stream endogenously through legal Stackelberg play. Verification and corruption-aware bandits [9, 14, 18, 24] and Stackelberg learning with manipulative or non-myopic agents [3, 7, 16, 25] motivate the defensive discussion below.

This paper makes the preceding claim precise. We compare two leader models in a repeated finite Stackelberg game. The baseline leader commits myopically to a classical SSE action and assumes immediate best response. The deceptive leader instead first rewards a target follower action $j^\star$ to raise its empirical mean, then switches to a leader-favorable action distribution for which $j^\star$ is no longer follower-optimal. The follower nevertheless continues selecting $j^\star$ for a calculable number of rounds because its UCB index remains larger than all competitors.

Our contribution is theoretical. We do not claim that every Stackelberg game admits profitable deception. We identify a verifiable class of games satisfying payoff separation, targetability, and exploitability assumptions, and prove that in this class the deceptive leader obtains strictly more cumulative utility than the static SSE ceiling. The proof exposes the failure mode: static mechanism design optimizes against the best-response correspondence $y \in \arg\max_{a \in A_F} U_F(x, a)$, whereas an empirical follower implements a stateful map from histories to actions.

2. Model

Let $A_L = \{1, \dots, m\}$ and $A_F = \{1, \dots, n\}$ be finite action sets. A T-step repeated Stackelberg game is specified by mean utilities

$$U_L, U_F : A_L \times A_F \to [0, 1].$$

At each time t, the leader selects $x_t \in \Delta(A_L)$, the follower selects $y_t \in A_F$, and the realized follower reward is

$$R_t^F = U_F(x_t, y_t) + \eta_t,$$

where (η_t) is conditionally mean-zero and σ-sub-Gaussian; the deterministic model is $\sigma = 0$. The leader's payoff is its expected utility $U_L(x_t, y_t) = \sum_i x_t(i) U_L(i, y_t)$. The main theorem assumes the leader observes the follower's sufficient statistics $(N_j, \hat{\mu}_j)_j$ and knows both payoff matrices; Appendix E records what changes under action-only observation. The follower observes only its realized action and scalar reward.

Definition 1 (Classical SSE value) *For a mixed leader action x, define the follower best-response set*

$$\mathrm{BR}(x) = \arg\max_{j \in A_F} U_F(x, j).$$

The strong Stackelberg value is

$$V_{\mathrm{SSE}} = \max_{x \in \Delta(A_L)} \max_{j \in \mathrm{BR}(x)} U_L(x, j).$$

The SSE abstraction imposes $y_t \in \mathrm{BR}(x_t)$ at every round. This converts the interaction into a static commitment problem. The follower studied here instead uses UCB1. Let $N_j(t) = \sum_{s \le t} \mathbf{1}\{y_s = j\}$ and let $\hat{\mu}_j(t)$ be the empirical mean of rewards observed by the follower when it

played j up to time t. Initially every arm is pulled once, or equivalently $N_j(0) = 0$ and the algorithm uses an initialization schedule over n rounds. Thereafter, for exploration parameter $c > 0$,

$$y_t \in \arg\max_{j \in A_F} \left[\hat{\mu}_j(t-1) + c\sqrt{\frac{\ln t}{N_j(t-1)}} \right],$$

where unplayed arms have index $+\infty$.

The crucial difference is that $\mathrm{BR}(x_t)$ is a function of the current leader mixture, whereas UCB is a function of the entire empirical history. Thus, a leader can manipulate the sufficient statistics $(\hat{\mu}_j, N_j)_{j \in A_F}$ before changing x_t.

3. Deceptive Leader Mechanism

Fix a target follower action $j^\star \in A_F$. The mechanism has two phases. During the honeypot phase the leader selects mixtures that give high follower reward to $j^\star$. During the trap phase the leader selects a selfish exploitative mixture x^{exp} that maximizes the leader's payoff conditional on the follower choosing $j^\star$.

Assumption 2 (Targetability) *There exists a honeypot mixture $x^{\mathrm{hon}} \in \Delta(A_L)$ and numbers $\alpha, \beta \in [0,1]$ with $\alpha > \beta$ such that*

$$U_F(x^{\mathrm{hon}}, j^\star) = \alpha, \qquad U_F(x^{\mathrm{hon}}, j) \le \beta \quad \forall j \neq j^\star.$$

Assumption 3 (Exploitability and follower harm) *There exists $x^{\mathrm{exp}} \in \Delta(A_L)$ and a constant $\rho \in [0,1]$ such that*

$$U_F(x^{\mathrm{exp}}, j^\star) = \rho, \qquad \max_{j \neq j^\star} U_F(x^{\mathrm{exp}}, j) = \rho + \gamma_F$$

for some follower suboptimality gap $\gamma_F > 0$, and

$$L^\star := U_L(x^{\mathrm{exp}}, j^\star) > V_{\mathrm{SSE}}.$$

The second condition is the source of profit and deception: $j^\star$ is strictly suboptimal for the follower under x^{exp}, yet unusually valuable for the leader.

Algorithm 1: Deceptive Leader Mechanism

Data: horizon T, target $j^\star$, UCB parameter c, mixtures $x^{\mathrm{hon}}, x^{\mathrm{exp}}$
Initialize the follower by allowing one sample of each action;
for $t = n+1, \ldots, n+\tau_1$ **do**
  play $x_t = x^{\mathrm{hon}}$;
end
for $t = n+\tau_1+1, \ldots, T$ **do**
  play $x_t = x^{\mathrm{exp}}$;
end

4. Honeypot Inflation

Let $t_0 = n$ denote the end of forced initialization. Suppose every non-target action has bounded initialization history and $j^\star$ has initial count $N_0 \geq 1$ and empirical mean $\hat{\mu}_0$. The honeypot phase repeatedly plays x^{hon} until $j^\star$ has τ_1 additional observations of reward α.

Lemma 4 (Exact target mean after honeypot) *If $j^\star$ has $N_0 \geq 1$ initialization samples with empirical mean $\hat{\mu}_0$, then after τ_1 honeypot samples,*

$$\hat{\mu}_{j^\star}(t_0 + \tau_1) = \frac{N_0\hat{\mu}_0 + \tau_1\alpha}{N_0 + \tau_1}.$$

Consequently, for any $\varepsilon > 0$,

$$\hat{\mu}_{j^\star}(t_0 + \tau_1) \geq \alpha - \varepsilon \quad \textit{whenever} \quad \tau_1 \geq N_0\frac{\alpha - \hat{\mu}_0 - \varepsilon}{\varepsilon}.$$

Proof Appendix A gives the algebra. ■

Theorem 5 (Minimum honeypot length for strict UCB dominance) *Fix a desired post-honeypot time $s = t_0 + \tau_1 + 1$. Suppose each non-target action $j \neq j^\star$ has $1 \leq N_j(t_0) \leq B_N$ and empirical mean at most B_μ. After τ_1 honeypot samples of $j^\star$, the target UCB index strictly dominates every non-target index at time s if and only if*

$$\frac{N_0\hat{\mu}_0 + \tau_1\alpha}{N_0 + \tau_1} + c\sqrt{\frac{\ln s}{N_0 + \tau_1}} > B_\mu + c\sqrt{\ln s}\max_{j \neq j^\star}\frac{1}{\sqrt{N_j(t_0)}}.$$

In particular, a sufficient explicit condition is

$$\tau_1 \geq \min\left\{q \in \mathbb{N} : \frac{N_0\hat{\mu}_0 + q\alpha}{N_0 + q} + c\sqrt{\frac{\ln(t_0 + q + 1)}{N_0 + q}} > B_\mu + c\sqrt{\ln(t_0 + q + 1)}\right\},$$

where the right-hand side uses the conservative bound $N_j(t_0) \geq 1$.

Proof Appendix A expands the index comparison term by term. ■

For stochastic rewards, Appendix B gives a high-probability analogue obtained by subtracting uniform sub-Gaussian confidence radii from the target index and adding them to the competing indices.

5. Trap Duration

Let $\tau = t_0 + \tau_1$ be the switch time. During exploitation the follower receives reward ρ whenever it plays $j^\star$. Let $N_\star = N_0 + \tau_1$ and let

$$M_\star = N_0\hat{\mu}_0 + \tau_1\alpha.$$

If the follower keeps selecting $j^\star$ for q exploitation rounds, then at time $\tau+q+1$ its target empirical mean is

$$\bar{\mu}_\star(q) = \frac{M_\star + q\rho}{N_\star + q}.$$

The corresponding target index is

$$I_\star(q) = \bar{\mu}_\star(q) + c\sqrt{\frac{\ln(\tau+q+1)}{N_\star+q}}.$$

For a non-target action j, define the frozen comparator index

$$J_j(q) = \hat{\mu}_j(t_0) + c\sqrt{\frac{\ln(\tau+q+1)}{N_j(t_0)}}.$$

Theorem 6 (Exact lock-in duration) *Assume deterministic rewards. Conditional on strict dominance at the switch, the exact number of consecutive exploitation rounds during which UCB selects $j^\star$ is*

$$\Delta = \max\left\{q \in \{0,\dots,T-\tau\} : I_\star(r) > \max_{j\neq j^\star} J_j(r) \text{ for all } r = 0,\dots,q-1\right\}.$$

Equivalently, the first escape time is

$$q_{\text{esc}} = \min\left\{q \geq 0 : \frac{M_\star + q\rho}{N_\star+q} + c\sqrt{\frac{\ln(\tau+q+1)}{N_\star+q}} \leq \max_{j\neq j^\star}\left[\hat{\mu}_j(t_0) + c\sqrt{\frac{\ln(\tau+q+1)}{N_j(t_0)}}\right]\right\},$$

with $\Delta = \min\{q_{\text{esc}}, T-\tau\}$ when the minimum exists and $\Delta = T-\tau$ otherwise.

Proof Appendix A proves the statement by induction on exploitation rounds. ■

Appendix B also gives a high-probability lower bound on Δ by requiring the confidence-adjusted target index to dominate the frozen competitors for every $q < Q$.

6. Utility Above the SSE Ceiling

Let $H^\star = U_L(x^{\text{hon}}, j^\star)$ and define the worst honeypot leader payoff $H_{\min} = \min_j U_L(x^{\text{hon}}, j)$. Since utilities lie in $[0,1]$, the cost of one honeypot round relative to the SSE ceiling is at most $V_{\text{SSE}} - H_{\min} \leq 1$.

Theorem 7 (Strict improvement over static SSE) *Assume targetability and exploitability. Let the deceptive mechanism use a honeypot length τ_1 satisfying the dominance condition and suppose the induced lock-in duration is Δ. If*

$$\Delta(L^\star - V_{\text{SSE}}) > \tau(V_{\text{SSE}} - H_{\min}),$$

then the leader's cumulative utility strictly exceeds TV_{SSE} on the realized path.

Proof Appendix A gives the cumulative payoff decomposition. ■

Thus if $\tau = O(\sqrt{T\ln T})$ and $\Delta = \Theta(T)$, then $G_{\text{dec}} - TV_{\text{SSE}} = \Theta(T) - O(\sqrt{T\ln T})$. Since payoffs are in $[0,1]$, the leader's exploration regret over the signaling phase is at most τ; the formal regret statement is deferred to Appendix A.

7. Simulated Experiments

The experiments are diagnostics for the constructive inequalities rather than proof substitutes. We compare a deceptive leader against an oblivious SSE leader in synthetic non-zero-sum 10×10 matrix games and a security-game topology with targets, coverage probabilities, attacker choices, and defender utilities.

Methodology. For each seed, compute V_{SSE} by enumerating follower actions and solving the leader's constrained program. The deceptive leader uses x^{hon} until the target index dominates, then switches to x^{exp}. For direct validation of Theorem 7, we also evaluate the continuation in which the leader returns to an SSE policy after the first escape, matching the proof. We record cumulative utility, target frequency, index gap, switch time, and escape time. An anonymous implementation for reproducing the diagnostics is available at https://anonymous.4open.science/r/optimism-vulnerability.

Diagnostic results. The matrix/security runs use $T = 20000$ and 20 seeds. The certified runs use $T = 200000$ and 5 seeds. Security shows a small positive mean advantage, matrix yields a negative advantage, the non-reverting certified run produces long lock-in but negative total advantage, and the theorem-matched reverting run produces a large positive advantage. A change-point follower collapses lock-in to three rounds, while EXP3 produces short lock-in but still positive average leader advantage through its exploration mixture.

Mean cumulative differences were negative in matrix games and positive in security games. In the certified instance, UCB with reversion achieved $+55178.93$, change-point defense reduced the difference to -527.04, and EXP3 averaged $+82764.91$ with high variance. Appendix F reports the full table. The predicted phase transition is therefore a falsifiable inequality, not a visual story. Around the switch, the relevant empirical object is the index gap

$$I_\star(q) - \max_{j \neq j^\star} J_j(q),$$

which should cross zero near the observed escape time. Follow-up experiments should sweep c, stochastic noise, initialization, and defended followers using change-point tests or corruption-robust confidence updates.

8. Conclusion

The main lesson is structural: optimism is not merely an exploration heuristic when rewards are generated by strategic opponents. It is a manipulable state variable. Classical SSE is a static equilibrium concept; it cannot price the value of falsified empirical histories because those histories do not exist in the model. In games satisfying explicit targetability and exploitability conditions, a leader can purchase optimism through honeypot rewards and later convert it into utility above the SSE ceiling.

Future work should characterize deception-resistant learning rules for Stackelberg environments: adversarial bandit updates, change-point tests, memory-limited estimators, robust confidence intervals under strategic contamination, and equilibrium concepts in which the leader commits to an information-generation policy. A satisfactory theory must preserve the statistical benefits of exploration without making exploration itself a programmable vulnerability.

References

[1] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47:235–256, 2002. URL https://api.semanticscholar.org/CorpusID:207609497.

[2] Rishab Balasubramanian, Jiawei Li, Prasad Tadepalli, Huazheng Wang, Qingyun Wu, and Haoyu Zhao. Adversarial attacks on combinatorial multi-armed bandits, 2024. URL https://arxiv.org/abs/2310.05308.

[3] Georgios Birmpas, Jiarui Gan, Alexandros Hollender, Francisco J. Marmolejo-Cossío, Ninad Rajgopal, and Alexandros A. Voudouris. Optimally deceiving a learning leader in stackelberg games. *Journal of Artificial Intelligence Research*, 72:507–531, October 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12542. URL http://dx.doi.org/10.1613/jair.1.12542.

[4] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. 01 2006. ISBN 978-0-521-84108-5. doi: 10.1017/CBO9780511546921.

[5] Vincent Conitzer and Tuomas Sandholm. Computing the optimal strategy to commit to. In *ACM Conference on Economics and Computation*, 2006. URL https://api.semanticscholar.org/CorpusID:2219280.

[6] Drew Fudenberg and David K. Levine. The theory of learning in games. Levine's Working Paper Archive 624, David K. Levine, Dec 1996. URL https://ideas.repec.org/p/cla/levarc/624.html.

[7] Nika Haghtalab, Thodoris Lykouris, Sloan Nietert, and Alexander Wei. Learning in stackelberg games with non-myopic agents, 2025. URL https://arxiv.org/abs/2208.09407.

[8] Kwang-Sung Jun, Lihong Li, Yuzhe Ma, and Xiaojin Zhu. Adversarial attacks on stochastic bandits, 2018. URL https://arxiv.org/abs/1810.12188.

[9] Sanghwa Kim, Junghyun Lee, and Se-Young Yun. A jointly efficient and optimal algorithm for heteroskedastic generalized linear bandits with adversarial corruptions, 2026. URL https://arxiv.org/abs/2602.10971.

[10] Dmytro Korzhyk, Vincent Conitzer, and Ronald Parr. Security games with multiple attacker resources. pages 273–279, 01 2011. doi: 10.5591/978-1-57735-516-8/IJCAI11-056.

[11] Tor Lattimore and Csaba Szepesvari. Bandit algorithms. 2017. URL https://tor-lattimore.com/downloads/book/book.pdf.

[12] Fang Liu and Ness Shroff. Data poisoning attacks on stochastic bandits, 2019. URL https://arxiv.org/abs/1905.06494.

[13] Guanlin Liu and Lifeng Lai. Action-manipulation attacks against stochastic bandits: Attacks and defense. *IEEE Transactions on Signal Processing*, 68:5152–5165, 2020. ISSN 1941-0476. doi: 10.1109/tsp.2020.3021525. URL http://dx.doi.org/10.1109/TSP.2020.3021525.

[14] Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions, 2018. URL https://arxiv.org/abs/1803.09353.

[15] Yuzhe Ma and Zhijin Zhou. Adversarial attacks on adversarial bandits, 2023. URL https://arxiv.org/abs/2301.12595.

[16] Thanh Nguyen and Haifeng Xu. Imitative attacker deception in stackelberg security games. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 528–534. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/75. URL https://doi.org/10.24963/ijcai.2019/75.

[17] Ann Nowe, Peter Vrancx, and Yann-Michaël De Hauwere. *Game Theory and Multi-agent Reinforcement Learning*, page 30. 01 2012. ISBN 978-3-642-27645-3. doi: 10.1007/978-3-642-27645-3_14.

[18] Anshuka Rangi, Long Tran-Thanh, Haifeng Xu, and Massimo Franceschetti. Saving stochastic bandits from poisoning attacks via limited data verification, 2022. URL https://arxiv.org/abs/2102.07711.

[19] Yoav Shoham, Rob Powers, and Trond Grenager. If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, 171(7):365–377, 2007. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2006.02.006. URL https://www.sciencedirect.com/science/article/pii/S0004370207000495. Foundations of Multi-Agent Learning.

[20] Aleksandrs Slivkins. Introduction to multi-armed bandits, 2024. URL https://arxiv.org/abs/1904.07272.

[21] Milind Tambe. Security and game theory: Algorithms, deployed systems, lessons learned. 01 2011. doi: 10.1017/CBO9780511973031.

[22] Bernhard von Stengel and Shmuel Zamir. Leadership games with convex strategy sets. *Games and Economic Behavior*, 69(2):446–457, 2010. ISSN 0899-8256. doi: https://doi.org/10.1016/j.geb.2009.11.008. URL https://www.sciencedirect.com/science/article/pii/S0899825609002322.

[23] Zhiwei Wang, Huazheng Wang, and Hongning Wang. Stealthy adversarial attacks on stochastic multi-armed bandits, 2024. URL https://arxiv.org/abs/2402.13487.

[24] Yinglun Xu, Zhiwei Wang, and Gagandeep Singh. Robust thompson sampling algorithms against reward poisoning attacks, 2024. URL https://arxiv.org/abs/2410.19705.

[25] Yaolong Yu and Haipeng Chen. Decentralized online learning in general-sum stackelberg games, 2024. URL https://arxiv.org/abs/2405.03158.

[26] Shiliang Zuo. Near optimal adversarial attacks on stochastic bandits and defenses with smoothed responses, 2024. URL https://arxiv.org/abs/2008.09312.

Appendix A. Deferred Proofs

A.1. Proof of the Target Mean Lemma

The empirical mean is the arithmetic average of N_0 old samples summing to $N_0\hat{\mu}_0$ and τ_1 honeypot samples each equal to α:

$$\hat{\mu}_{j^\star}(t_0+\tau_1) = \frac{N_0\hat{\mu}_0 + \tau_1\alpha}{N_0+\tau_1}.$$

To force this mean above $\alpha-\varepsilon$, require

$$\frac{N_0\hat{\mu}_0 + \tau_1\alpha}{N_0+\tau_1} \geq \alpha-\varepsilon.$$

Multiplying by $N_0+\tau_1$ and canceling $\tau_1\alpha$ gives

$$N_0\hat{\mu}_0 \geq N_0\alpha - N_0\varepsilon - \tau_1\varepsilon, \qquad \tau_1\varepsilon \geq N_0(\alpha-\hat{\mu}_0-\varepsilon).$$

Dividing by $\varepsilon > 0$ proves the displayed threshold. If $\alpha-\hat{\mu}_0-\varepsilon < 0$, zero additional samples suffice.

A.2. Proof of the Deterministic Dominance Theorem

At time s, the target index equals

$$I_{j^\star}(s) = \frac{N_0\hat{\mu}_0+\tau_1\alpha}{N_0+\tau_1} + c\sqrt{\frac{\ln s}{N_0+\tau_1}}.$$

For every $j \neq j^\star$, no honeypot samples of j occur, hence

$$I_j(s) = \hat{\mu}_j(t_0) + c\sqrt{\frac{\ln s}{N_j(t_0)}} \leq B_\mu + c\sqrt{\ln s}\max_{k\neq j^\star}\frac{1}{\sqrt{N_k(t_0)}}.$$

Strict dominance is exactly $I_{j^\star}(s) > I_j(s)$ for all $j \neq j^\star$, which is equivalent to the theorem's necessary and sufficient condition. Replacing the maximum by 1 gives the conservative sufficient condition because $N_j(t_0) \geq 1$.

Appendix B. Stochastic Extensions

Let $\mathcal{E}_\delta$ be the event that for all arms j, all sample counts $k \leq T$, and all adaptive histories generated by the leader policy,

$$|\hat{\mu}_{j,k} - \mu_{j,k}| \leq r_\delta(k) := \sigma\sqrt{\frac{2\ln(4nT/\delta)}{k}},$$

where $\mu_{j,k}$ is the conditional mean average of the k rewards observed on arm j. Since rewards are conditionally σ-sub-Gaussian, a Hoeffding-Azuma bound plus a union bound over (j,k) gives $\mathbb{P}(\mathcal{E}_\delta) \geq 1-\delta$.

Theorem 8 (High-probability stochastic dominance) *With probability at least $1-\delta$, the target strictly dominates all non-target actions at switch time s whenever*

$$\frac{N_0\hat{\mu}_0+\tau_1\alpha}{N_0+\tau_1} - r_\delta(N_0+\tau_1) + c\sqrt{\frac{\ln s}{N_0+\tau_1}} > B_\mu + \max_{j\neq j^\star} r_\delta(N_j(t_0)) + c\sqrt{\ln s}\max_{j\neq j^\star} N_j(t_0)^{-1/2}.$$

B.1. Proof of Theorem 8

On $\mathcal{E}_\delta$, the post-honeypot target empirical mean is at least

$$\frac{N_0\hat{\mu}_0 + \tau_1\alpha}{N_0 + \tau_1} - r_\delta(N_0 + \tau_1),$$

and each non-target empirical mean is at most $B_\mu + r_\delta(N_j(t_0))$. Adding the exact UCB exploration terms to these lower and upper bounds yields the displayed sufficient inequality. Strict separation implies UCB chooses $j^\star$ at the switch.

Theorem 9 (High-probability lock-in lower bound) *On $\mathcal{E}_\delta$, UCB selects $j^\star$ for at least Q exploitation rounds if, for every $q < Q$,*

$$\frac{M_\star + q\rho}{N_\star + q} - r_\delta(N_\star + q) + c\sqrt{\frac{\ln(\tau + q + 1)}{N_\star + q}} > \max_{j \neq j^\star}\left[\hat{\mu}_j(t_0) + r_\delta(N_j(t_0)) + c\sqrt{\frac{\ln(\tau + q + 1)}{N_j(t_0)}}\right].$$

B.2. Proof of Theorem 9

Condition on $\mathcal{E}_\delta$. If $j^\star$ has been played for q exploitation rounds, its conditional mean average is $(M_\star + q\rho)/(N_\star + q)$ and its empirical mean is no smaller than this value minus $r_\delta(N_\star + q)$. Non-target empirical means remain bounded above by their frozen empirical values plus $r_\delta(N_j(t_0))$. The theorem's inequality makes the target UCB index strictly largest for every $q < Q$, so induction gives at least Q locked rounds.

B.3. Proof of the Deterministic Lock-in Theorem

We induct on exploitation rounds. At $q = 0$, strict dominance at the switch implies that UCB selects $j^\star$. Suppose the follower has selected $j^\star$ for exactly q exploitation rounds. The target count is $N_\star + q$ and the target cumulative reward is $M_\star + q\rho$. Every non-target statistic remains frozen at its value from t_0. Therefore, at the next decision time $\tau + q + 1$, UCB selects $j^\star$ if and only if

$$\frac{M_\star + q\rho}{N_\star + q} + c\sqrt{\frac{\ln(\tau + q + 1)}{N_\star + q}} > \max_{j \neq j^\star}\left[\hat{\mu}_j(t_0) + c\sqrt{\frac{\ln(\tau + q + 1)}{N_j(t_0)}}\right].$$

The maximal prefix satisfying this inequality is Δ; the first failure is q_{esc}.

B.4. Proof of Theorem 7 and the Sublinear Burden Claim

During the first τ rounds, the leader obtains at least $H_{\min}$ each round. During the Δ locked exploitation rounds, the follower plays $j^\star$, so the leader obtains $L^\star$ each round. If the leader reverts to an SSE action after escape, then

$$G_{\text{dec}} - TV_{\text{SSE}} \geq \tau(H_{\min} - V_{\text{SSE}}) + \Delta(L^\star - V_{\text{SSE}}).$$

The right-hand side is strictly positive precisely when

$$\Delta(L^\star - V_{\text{SSE}}) > \tau(V_{\text{SSE}} - H_{\min}).$$

If $\Delta = \Theta(T)$ and $\tau = O(\sqrt{T \ln T})$, the positive term is linear because $L^\star > V_{\text{SSE}}$, while the signaling loss is $O(\sqrt{T \ln T})$.

B.5. Exploration Regret Bound

Define $R_L^{\mathrm{hon}}(T) = \sum_{t=1}^{\tau}(L^\star - U_L(x_t, y_t))$. If $\tau \le K\sqrt{T \ln T}$, then $R_L^{\mathrm{hon}}(T) \le K\sqrt{T \ln T}$.

For every round, $U_L(x_t, y_t) \in [0, 1]$ and $L^\star \in [0, 1]$, hence

$$L^\star - U_L(x_t, y_t) \le 1.$$

Summing over τ honeypot and initialization rounds gives

$$R_L^{\mathrm{hon}}(T) \le \sum_{t=1}^{\tau} 1 = \tau \le K\sqrt{T \ln T}.$$

Appendix C. Additional Mathematical Details

Lemma 10 (Lock-in lower bound) *Suppose the frozen competitor index is bounded by*

$$\max_{j \neq j^\star} J_j(q) \le \Gamma(q), \qquad \rho < \Gamma(q) < \alpha,$$

for $q \le Q$. If

$$\frac{M_\star + q\rho}{N_\star + q} > \Gamma(q)$$

for every $q < Q$, then $\Delta \ge Q$ even without the target exploration bonus.

Proof For $q < Q$, the target empirical mean alone exceeds every non-target index upper bound. Adding the nonnegative optimism term preserves strict dominance. Thus UCB selects $j^\star$ throughout the first Q exploitation rounds. ■

Lemma 11 (Closed-form sufficient trap length) *If $\Gamma(q) \le \bar{\Gamma} < \alpha$ for $q \le Q$, then the condition*

$$Q < \frac{M_\star - \bar{\Gamma} N_\star}{\bar{\Gamma} - \rho}$$

implies $\Delta \ge Q$ whenever $\rho < \bar{\Gamma}$.

Proof The inequality

$$\frac{M_\star + q\rho}{N_\star + q} > \bar{\Gamma}$$

is equivalent to

$$M_\star - \bar{\Gamma} N_\star > q(\bar{\Gamma} - \rho).$$

The displayed bound makes this true for every $q < Q$. ■

Proposition 12 (Empirical falsification criterion) *For a fixed implementation, define*

$$\widehat{D}_T = \frac{1}{S} \sum_{s=1}^{S} \left(G_{\mathrm{dec}}^{(s)}(T) - G_{\mathrm{SSE}}^{(s)}(T) \right).$$

If $\widehat{D}_T < 0$, then at least one of the finite-sample conditions $\widehat{L}^\star > \widehat{V}_{\mathrm{SSE}}$, sufficient lock-in $\widehat{\Delta}$, or bounded signaling loss fails in the tested generator.

Proof Theorem 7 proves that all three conditions imply $G^{(s)}_{\mathrm{dec}}(T) > G^{(s)}_{\mathrm{SSE}}(T)$ pathwise for every seed satisfying them. A negative sample average is therefore only possible if the implication's antecedent fails for at least one seed, or if the implemented baseline differs from the theoretical SSE comparator. ■

Proposition 13 (UCB-parameter misspecification margin) *Suppose the leader designs the switch using $\hat{c}$ but the follower uses c with $|c - \hat{c}| \le \kappa$. If the designed target-index margin at time s is larger than*

$$\kappa\sqrt{\ln s}\left(\frac{1}{\sqrt{N_{j^\star}(s-1)}} + \max_{j\neq j^\star}\frac{1}{\sqrt{N_j(s-1)}}\right),$$

then the actual follower still selects $j^\star$ at time s.

Proof Changing the UCB parameter from $\hat{c}$ to c perturbs the target bonus by at most $\kappa\sqrt{\ln s/N_{j^\star}(s-1)}$ and any competitor bonus by at most $\kappa\sqrt{\ln s/N_j(s-1)}$. If the designed strict margin exceeds the sum of the worst target decrease and competitor increase, the sign of the margin is preserved. ■

Appendix D. Operational Certification

The assumptions can be checked by solving small linear programs. For each candidate follower response j, compute the SSE face value

$$V_j = \max_{x\in\Delta(A_L)} U_L(x,j) \quad \text{s.t.} \quad U_F(x,j) \ge U_F(x,k)\ \forall k \in A_F,$$

and set $V_{\mathrm{SSE}} = \max_j V_j$. For each target $j^\star$, the honeypot margin is obtained from

$$\max_{x\in\Delta(A_L),\, z} z \quad \text{s.t.} \quad U_F(x,j^\star) - U_F(x,k) \ge z\ \forall k \neq j^\star.$$

The exploit certificate solves, for each $j^\star$,

$$\max_{x\in\Delta(A_L)} U_L(x,j^\star) - V_{\mathrm{SSE}} \quad \text{s.t.} \quad U_F(x,k) - U_F(x,j^\star) \ge \gamma_F \text{ for some } k \neq j^\star.$$

A candidate triple $(j^\star, x^{\mathrm{hon}}, x^{\mathrm{exp}})$ is accepted only if the certified lock-in lower bound Q satisfies

$$Q(L^\star - V_{\mathrm{SSE}}) > \tau(V_{\mathrm{SSE}} - H_{\min}).$$

This condition corresponds to a leader that reverts to an SSE policy after the first escape; otherwise post-escape exploit rounds can dominate the accounting and destroy net gain even when lock-in is long.

Certified toy instance. The certified instance uses

$$U_L = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad U_F = \begin{pmatrix} 1 & 0 \\ 0.45 & 0.55 \end{pmatrix},$$

with $j^\star = 1$, $x^{\mathrm{hon}} = (1,0)$, and $x^{\mathrm{exp}} = (0,1)$. Under x^{exp}, the target is follower-suboptimal by $\gamma_F = 0.10$, while $L^\star = 1$. This instance is intended for controlled sweeps over c and the enforced honeypot length.

Appendix E. Observability and Defenses

Partial observability. The main construction assumes the leader observes $(N_j, \hat{\mu}_j)_j$. If the leader observes only follower actions and its own selected mixtures, then N_j is still known exactly, while $\hat{\mu}_j$ can be estimated when the leader knows U_F and the stochastic reward model. A conservative implementation replaces each unknown empirical mean by a confidence interval; the honeypot stops only when the lower confidence bound on the target index exceeds upper confidence bounds on all alternatives. This increases τ_1 by the same concentration radii appearing in Theorem 8.

Defended followers. Change-point tests can flag the trap because the target reward mean changes from α to ρ. If a detector requires a detectable drop of order ϵ_{det}, then any trap with $\alpha - \rho > \epsilon_{\text{det}}$ risks early reset. Corruption-robust UCB or robust Thompson sampling changes the estimator from a plain empirical mean to a robust statistic; in our notation this replaces $r_\delta(k)$ by a larger but attack-aware radius depending on the corruption budget. The deception succeeds only if the honeypot-induced gap exceeds that robust radius, so defenses convert the vulnerability into a quantitative cost increase.

Appendix F. Additional Experiment Details

Table 1: Cumulative reward and trap diagnostics.

Environment	Deceptive	Baseline	Difference	Target freq.	Switch	Escape
Matrix	2107.27	9896.70	−7789.43	0.00595	34	36
Security	3058.05	3048.45	9.60	0.00365	34	36
Cert. UCB, no revert	4950.00	95008.68	−90058.68	0.02975	1002	4528
Cert. UCB, revert	150187.61	95008.68	55178.93	0.82936	1002	4528
Cert. CP, revert	94481.64	95008.68	−527.04	0.52471	1002	1005
Cert. EXP3, revert	172435.94	89671.03	82764.91	0.95349	1056.2	1070.2

All reported experiments use deterministic follower rewards, UCB exploration parameter $c = 2.0$ for the matrix and security diagnostics, and $c = 0.2$ with an enforced 1000-target-sample honeypot in the certified UCB diagnostic. Defended-follower evaluations use the same certified instance with either a simple reward-drop change-point reset or an EXP3-style adversarial-bandit update. The EXP3 row reports means over five seeds and has high variance in both baseline and deceptive rewards. The anonymous code release at https://anonymous.4open.science/r/optimism-vulnerability contains the simulator, certified instance, and diagnostic logging used for these rows.