Security and Privacy of Machine Learning, 2025 Critique G12: Security & Privacy of Multi-Agent Systems – (1) Secret Collusion among AI Agents: Multi-Agent Deception via Steganography; (2) Single-agent Poisoning Attacks Suffice to Ruin Multi-Agent Learning; (3) Cowpox: Towards the Immunity of VLM-based Multi-Agent Systems

Shih-Yu Lai
National Taiwan University
Taipei, Taiwan
akinesia112@gmail.com

Critique:(1) Secret Collusion among AI Agents: Multi-Agent Deception via Steganography

Summary

The paper formalizes "secret collusion" in decentralized systems of generative AI agents, where agents use steganographic channels to covertly coordinate under the nose of an overseer. It contributes (i) a conceptual framework for collusion in DecAI settings, (ii) a capability map and evaluation suite (CASE) for covertext/complexity-/information-theoretic steganography, and (iii) empirical results across multiple LLMs showing limited but scaling steganographic competence (notably a jump with GPT-4). The authors argue that while current models fall short of perfectly secure stego, proactive monitoring and disincentives are needed as capabilities improve [1].

Strengths

- **Timely problem framing.** The work crisply defines collusion risk in realistic multi-agent deployments, bridging security and multi-agent AI rather than treating steganography as an isolated NLP curiosity [1].
- **Grounded in steganography theory.** By situating covertext modification against provable approaches (e.g., Minimum Entropy Coupling, MEC), the framework aligns evaluations with information-theoretic guarantees, not just heuristics [2].
- **Capability decomposition & CASE.** The decomposition into common-knowledge requirements, cipher tasks, MEC/coupling, self-sampling/logit simulation, and group tests provides a reusable scaffold for red-teaming different model families [1].
- **Empirical scaling observations.** The study documents disjoint improvements with model size (e.g., higher success rates in Schelling coordination and ciphers), a practically relevant signal for policy and risk forecasting [1].
- **Connection to algorithmic collusion literature.** The motivation is consistent with prior evidence that learning agents can coordinate on supracompetitive strategies even without explicit channels, hence the need to evaluate covert channels explicitly [3].

Weaknesses

- **Detection realism gap.** The threat model emphasizes perfectly secure stego, yet practical deployments must wrestle with mixed channels, rate limits, and content moderation. The paper stops short of quantifying *operational detectability* under realistic throughput, latency, and paraphrasing budgets, especially vs. modern cryptographic stego like METEOR that adapts to entropy variability [4].
- **Limited cross-modal scope.** Although collusion risks transcend text (images/audio/agent tool-outputs), the empirical core is text-centric. Diffusion-based and VLM-based stego—increasingly relevant—are left for future work.
- **Oversight as a monolith.** The overseer E is modeled as passive/active with state access, but the paper does not probe multi-overseer games (heterogeneous detectors, budgeted intervention) or adversarial co-evolution where agents *optimize* against the monitor.
- **Economic incentives under-specified.** The payoff model $J_i = g_i + h_i$ shows why secret collusion dominates when detection is penalized, but concrete policy knobs (audit

probability, penalty size, rate caps) are not calibrated against detection ROC and communication capacity, limiting prescriptive value.
- **Privacy and leakage interplay.** The work touches on security monitoring, but does not connect to membership inference & leakage measurement as possible *side-channel* detectors for covert info-sharing at scale [5].

Potential Improvements / Extensions

- **Utility–Stealth frontiers.** Map Pareto frontiers of (task utility, stego capacity, detectability) for each method (covertext/MEC/METEOR) under realistic rate limits and paraphrase budgets; report confidence intervals under content filters and watermarkers [2], [4].
- **Cross-modal CASE.** Extend CASE to images (diffusion stego), audio, and tool APIs (function-calling side channels); evaluate multi-hop collusion where text stego triggers image stego (and vice versa).
- **Mechanism design baselines.** Add *incentive-compatible* protocols: randomized audits, posted-price penalties, and rate-limiting that shrink channels' effective capacity; learn audit policies via bandit/RL against adaptive agents, measuring equilibrium collusion rates [3].
- **Composite overseers.** Evaluate ensembles of lightweight anomaly detectors (token-level MEC-residuals, perplexity deltas, style-shift metrics) with a heavy cryptographic/semantic inspector under a budget, reporting cost-adjusted risk.
- **Leakage-aware monitoring.** Integrate membership-inference and distribution-drift probes as long-horizon detectors of covert data-sharing within agent ecosystems [5].

Questions for the Authors

1) How does CASE score methods when agents mix *multiple* channels (e.g., sparse MEC bits + low-rate METEOR) to trade off throughput and stealth?
2) Can you report calibration plots linking overseer false positives to economic welfare (missed utility, latency)? What audit rates make collusion no longer payoff-dominant?
3) In group tasks, do agents evolve *error-correcting conventions* (Schelling codebooks) over long horizons, and how brittle are these to paraphrase/watermark perturbations?
4) What are the failure modes when an overseer is weaker than colluding agents but has *tool* advantages (e.g., private corpora or logit access)?
5) How do results change for multimodal agents (LLM+VLM+diffusion), where non-text modalities may carry higher-capacity covert channels?

Critique: (2) Single-agent Poisoning Attacks Suffice to Ruin Multi-Agent Learning

Summary

The paper studies robustness of multi-agent learning (MAL) in games with a unique Nash equilibrium (NE), focusing on β-strongly monotone games in the sense of Rosen [6]. The authors introduce a *single-agent* utility-poisoning strategy (SUSA) that perturbs only one agent's observed utilities each round. Despite this minimal attack surface, they prove (i) the attacked game remains strongly monotone under a small shift, so standard last-iterate convergence claims still apply, and (ii) the learning dynamics converge to a *shifted* equilibrium $\tilde{x}^\star \neq x^\star$ at a distance lower-bounded by curvature-sensitive terms. Crucially, they show the *total* corruption needed is *sublinear* in horizon T, tying the attack budget exponent to the non-adversarial convergence rate of the underlying (α, p)-MAL algorithm (e.g., mirror-descent variants [7]). This yields an *efficiency–robustness trade-off*: faster convergence (larger α) implies greater vulnerability to NE shifting, reminiscent of fragility phenomena in corrupted bandit feedback [8]. The paper then analyzes MD-SCB (a fast mirror-descent scheme) and shows that slowing step-sizes restores last-iterate convergence under nearly linear corruption—thereby mapping a defender's side of the trade-off. Experiments on Cournot competition (a classical market game [9]) validate: (a) NE shifting with sublinear cumulative corruption and (b) scaling trends with the number of agents. The work closes with limitations and broader implications for safety of MAL, and positions the results alongside privacy/attack literatures [10].

Strengths

- **Minimal attack surface, strong effect.** Demonstrates that corrupting a single agent's utilities suffices to cause a persistent NE shift, even when convergence guarantees still hold. This sharpens our understanding of *what* guarantees *do not* protect.
- **Abstraction via (α, p)-MAL.** The (α, p) template cleanly factors algorithm speed from robustness, enabling general theorems that apply across popular mirror-descent families [7].
- **Tight conceptual message: efficiency vs. robustness.** The attack and the defense (through step-size scheduling) delineate a clear, actionable trade-off frontier that practitioners can use to tune systems.
- **Curvature-aware bounds.** The role of strong monotonicity, best-response Lipschitzness, and Hessian norms clarifies how game structure mediates attack impact; this offers levers for mechanism designers.
- **Empirical grounding.** Simulations on Cournot highlight (i) sublinear attack budget growth, (ii) diminishing shift with larger n, and (iii) the protective effect of slower schedules—all predicted by theory.

Weaknesses

- **Strong knowledge and actuation assumptions.** The attacker is assumed to know (and differentiate) the victim utility and to add adaptive, per-round corruptions. In many markets or networked systems, both knowledge and immediate actuation channels can be constrained or noisy.
- **Model class limitations.** Results rely on β-strong monotonicity and unique NE. Many socio-technical MAL

settings are only monotone on subsets, non-monotone, or exhibit multiple equilibria/cycles.
- **Conservatism in constants.** Bounds depend on spectral norms of the game Hessian and Lipschitz parameters; these can be loose in high-dimensions, potentially overstating worst-case vulnerability (or, conversely, understating it if constants are underestimated).
- **No stochastic/communication frictions.** Real MAL often has delays, packet loss, quantization, or gradient-noise; the interaction between those frictions and poisoning (constructive or destructive) remains uncharted.

Potential Improvements / Extensions

- **Weaken attacker knowledge.** Develop black-box SUSA using trajectory-level inverse optimization or meta-learning to estimate victim curvature/BR maps online; quantify regret and budget inflation relative to white-box SUSA.
- **Beyond strong monotonicity.** Extend to (i) merely monotone variational inequalities (Minty solutions), (ii) aggregate games with non-unique NE, and (iii) potential games with saddle structures; characterize when NE-shift becomes cycle/limit-set shift.
- **Detection and mitigation.** Propose *online drift tests* for NE movement (e.g., testing stationarity of x_t against VI residuals), and *robust estimators* (median-of-means bandit payoffs, clipped mirror steps) to bound bias under sparse corruptions.
- **Tighten the frontier.** Close the gap between the upper/lower budget-exponent bounds ($1-\alpha$ vs. $1-\frac{2\alpha}{3}$) via refined bias-propagation analysis or adversarial constructions tailored to specific mirror maps.
- **Broader benchmarks.** Add experiments on Tullock contests, first-price auctions, traffic routing, and decentralized RL settings; include delayed/asynchronous updates and partial observability to test robustness recipes.
- **Mechanism design with minimal interventions.** Formalize "poisoning for good": compute minimal perturbations that move the operating point toward welfare or fairness targets subject to budget and detectability constraints.

Questions for Author

1) **Frontier tightness:** Can the $1-\frac{2\alpha}{3}$ upper bound be matched by a converse (lower) bound for an optimal *defender*, or is there a principled algorithm that pushes $\rho^+(\alpha)$ up to $1-\alpha$?
2) **Sparsity constraints:** If the attacker can corrupt only a γ-fraction of rounds (or with maximum per-round amplitude), how do the budget exponents and NE-shift scalings change?
3) **Dimensionality effects:** How do constants scale with joint action dimension and agent count n beyond the Cournot instance? Is there a preconditioning of mirror maps that reduces sensitivity constants without sacrificing α?
4) **Delayed/partial observations:** If the attacker observes actions with delay or noise, is there a phase transition where NE-shift collapses, or can it be restored via predictive filtering?
5) **Stealth/forensics:** What is the optimal trade-off between *undetectability* (small statistical footprint in payoff logs) and NE-shift magnitude? Can defenders exploit invariant tests of strong monotonicity violations online?

Critique: (3) CowPox: Towards the Immunity of VLM-based Multi-Agent Systems

Summary

The paper introduces COWPOX, a distributed defense for VLM-based multi-agent systems confronted with *infectious jailbreak* attacks such as AgentSmith [11]. In these attacks, an adversarial "virus" sample is engineered to dominate retrieval in an agent's RAG pipeline [12] and to elicit targeted unsafe behavior in the VLM. COWPOX equips only a small subset of agents with two modules: (i) an output-analysis inspector that flags suspicious histories, and (ii) a cure-generation routine that crafts a *cure sample* with higher expected RAG score than the virus while inducing benign VLM outputs. Two cure strategies are proposed: optimizing directly on the virus sample to neutralize it, or constructing a cure from a high-scoring benign base. The authors develop a transmission model with *infected*, *cured*, and *sensitive* states, and show a sufficient condition for eradication: if the cure converts infected agents faster than re-infection occurs (i.e., $\epsilon \geq \eta$), infections vanish asymptotically. Experiments on a homogeneous system (LLaVA-1.5-7B [13] + CLIP retrieval) demonstrate $>$95% recovery with only $\sim$3% COWPOX agents, preserved utility on LLaVA-Bench, and partial robustness to adaptive attackers. The authors discuss limitations (cure diversity, information recovery, and environment specificity) and suggest stronger inspectors (e.g., Llama Guard [14]).

Strengths

- **Defense as distributed immunization.** Turning the attack's own propagation channel (RAG-driven memory sharing) into a carrier for cures is elegant and practical when system-wide patching is infeasible.
- **Minimal deployment footprint.** The mechanism only instruments a small fraction of agents and does not require modifying all nodes, aligning with realistic edge deployments.
- **Clear mechanistic target.** By explicitly outcompeting the virus in RAG score, the defense directly attacks the positive-feedback loop that fuels infectious spread [11], [12].
- **Transmission-model lens.** The i–c–s model and the sufficient condition $\epsilon \geq \eta$ provide an interpretable knob (*retrieval margin* of cure over virus) that links retrieval dynamics to population-level outcomes.
- **Evidence for practicality.** With only 4/128 COWPOX agents, the system recovers while maintaining task utility;

adaptive attackers face a constrained feasible set, making cures easier to find than strong new viruses.

Weaknesses

- **Homogeneous stack and single-VLM assumption.** Most experiments use a single VLM (LLaVA-1.5) and CLIP retrieval. Real multi-agent ecosystems are heterogeneous in models, embeddings, tokenizers, and safety layers. Limited results on heterogeneity leave open whether cure margins reliably transfer across encoders and models (*cross-encoder generalization*).
- **Theory relies on idealized assumptions.** The analysis sets $|H| \to \infty$ and $\gamma \to 0$ to obtain $\delta \to \epsilon$. In practice, short histories, bounded albums, and eviction policies break these assumptions. Finite-memory churn can both help (forget viruses) and hurt (forget cures); a formal treatment for finite $|H|, |B|$ is missing.
- **Inspector coupling and leakage risk.** Using the same (or closely related) VLM family for both generation and inspection risks correlated blind spots and gradient leakage; detection failures propagate directly to cure generation quality.
- **Metrics and identifiability.** The operational estimate of infection chance β_t is tied to album/retrieval counts and may conflate carrier prevalence with transmissibility. Stronger causal identification of β, ϵ, η from logs would bolster claims.
- **Narrow attack surface.** The focus is on RAG-dominant visual AEs. Prompt-chain or system-prompt contamination, memory-slot poisoning, or cross-modal latent triggers may not be neutralized by RAG-score competition alone.

Potential Improvements / Extensions

- **Finite-memory dynamics.** Extend the transmission model to finite $|H|, |B|$ with eviction policies (e.g., FIFO, recency, score-aware TTL). Derive bounds linking the *retrieval margin* $\mathbb{E}[R(c) - R(v)]$ and the *persistence time* of cures/viruses to convergence rates.
- **Diverse, watermarkable cures.** Train an ensemble of cures with orthogonal perturbation subspaces and embed robust watermarks for provenance auditing. Watermarked cures allow downstream agents to prioritize trusted memory without full access control.
- **Heterogeneous and adversarial retrieval.** Evaluate across encoders (CLIP, DINOv2, EVA-CLIP) and text-embedding backends, including agents with mixture-of-indices and task-specific RAG. Add defenders that *randomize* retrieval keys to reduce attack transfer.
- **Inspector hardening.** Decouple inspector from generator family; fuse Llama Guard-style safety checks [14] with lightweight anomaly detection on embedding trajectories and retrieval-score residuals. Consider consensus inspectors with abstention.
- **Cure scheduling and quotas.** To mitigate conversation collapse to a few cures, impose quotas on identical cure reuse, use topic-aware diversification, and rotate cures via multi-armed bandit selection to maximize immunity while preserving topical diversity.
- **Beyond image-borne viruses.** Stress-test against prompt-chain infections, tool-augmented agents, and memory-serialization attacks where viruses ride code or API payloads; design modality-agnostic cures that operate on retrieval keys rather than pixels.
- **Trusted-memory slices.** Introduce write-protected, signed album slices where only verified cures (meeting a certified retrieval margin) can reside, preventing later overwrite by adaptive viruses.

Questions for the Authors

1) How sensitive is eradication to *margin noise*? Can you report failure probabilities as a function of the empirical gap $\Delta = \mathbb{E}[R(c) - R(v)]$ under encoder drift and distribution shift?
2) In finite-memory regimes, what is the minimal cure *dosage* (fraction of agents κ/N and emission frequency) to guarantee recovery within T rounds with probability $1 - \delta$?
3) Does cure optimization ever produce *benign overfitting* that later increases transferability of new viruses (e.g., by aligning to universal features of the embedding space)?
4) Can you provide head-to-head comparisons against *non-propagating* defenses (e.g., guardrails, memory filters) under the same budget to quantify the net benefit of propagation?
5) What safeguards prevent an attacker from crafting *fake cures* that score just above authentic cures while remaining harmful? Could cryptographic signing or watermark verification be integrated into retrieval scoring?

References

[1] S. R. Motwani, M. Baranchuk, M. Strohmeier, V. Bolina, P. H. S. Torr, L. Hammond, and C. Schroeder de Witt, "Secret collusion among ai agents: Multi-agent deception via steganography," *arXiv preprint arXiv:2402.07510*, 2024. [Online]. Available: https://arxiv.org/abs/2402.07510

[2] C. Schroeder de Witt, S. Sokota, J. Z. Kolter, J. N. Foerster, and M. Strohmeier, "Perfectly secure steganography using minimum entropy coupling," in *International Conference on Learning Representations (ICLR)*, 2023. [Online]. Available: https://arxiv.org/abs/2210.14889

[3] E. Calvano, G. Calzolari, V. Denicolò, and S. Pastorello, "Artificial intelligence, algorithmic pricing, and collusion," *American Economic Review*, vol. 110, no. 10, pp. 3267–3297, 2020. [Online]. Available: https://www.aeaweb.org/articles?id=10.1257/aer.20190623

[4] G. Kaptchuk, T. M. Jois, M. Green, and A. D. Rubin, "Meteor: Cryptographically secure steganography for realistic distributions," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21)*. New York, NY, USA: ACM, 2021, pp. 1–21. [Online]. Available: https://www.cs.umd.edu/users/kaptchuk/publications/ccs21_meteor.pdf

[5] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 3–18. [Online]. Available: https://www.cs.cornell.edu/~shmat/shmat_oak17.pdf

[6] J. B. Rosen, "Existence and uniqueness of equilibrium points for concave n-person games," *Econometrica*, vol. 33, no. 3, pp. 520–534, 1965.

[7] M. Bravo, D. Leslie, and P. Mertikopoulos, "Bandit learning in concave n-player games," in *NeurIPS*, 2018.

[8] K.-S. Jun, L. Li, Y. Ma, and J. Zhu, "Adversarial attacks on stochastic multi-armed bandits," in *AISTATS*, 2018.
[9] A.-A. Cournot, *Researches into the Mathematical Principles of the Theory of Wealth*. Hachette, 1838.
[10] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE S&P*, 2017.
[11] X. Gu, X. Zheng, T. Pang, C. Du, Q. Liu, Y. Wang, J. Jiang, and M. Lin, "Agent smith: A single image can jailbreak one million multimodal llm agents exponentially fast," in *International Conference on Machine Learning*, 2024.
[12] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *NeurIPS*, 2020.
[13] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in Neural Information Processing Systems*, vol. 36, 2024, lLaVA-1.5.
[14] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine *et al.*, "Llama guard: Llm-based input-output safeguard for human-ai conversations," *arXiv preprint arXiv:2312.06674*, 2023.