Methodology for Detecting and Mitigating Steganographic Collusion in MARL

Author: H. M. Liu

Supplementary Material: Formalized Metrics and Phase Boundary Analysis

1. Formalized Metrics (形式化指標)

To transition the concept of "steganographic collusion" from a theoretical vulnerability to a quantifiable empirical metric, we define three core probes based on system dynamics and information theory.

1.1 Surface Stability (Reward Variance)

This metric measures the stability of the agents' performance on the explicit environmental task. A highly fluctuating reward indicates policy collapse, while stable high rewards suggest convergence (either honest or collusive).

We calculate the rolling variance over a window W:

$$Var(R_t) = \frac{1}{W} \sum_{i=t-W+1}^{t} (R_i - \bar{R}_t)^2$$

(where $\bar{R}_t$ is the mean joint reward within the window).

1.2 Representational Stability (JSD Shift)

A critical challenge in the AIM framework is that the communication channel utilizes a discrete VQ-VAE bottleneck. Traditional variance cannot be applied to categorical indices. To measure whether agents are rapidly shifting their private semantics (codebook usage) to evade detection, we introduce the **Jensen-Shannon Divergence (JSD)** shift.

We split the observation window W into two halves (Past and Current) and compute the codebook usage frequency distributions P and Q with Laplace smoothing ($\epsilon = 1e^{-8}$):

$$P(c) = \frac{\sum_{i=t-W}^{t-W/2-1} \mathbb{I}(C_i = c) + \epsilon}{Z_P} \quad , \quad Q(c) = \frac{\sum_{i=t-W/2}^{t-1} \mathbb{I}(C_i = c) + \epsilon}{Z_Q}$$

The representational shift is quantified using JSD (where $M = \frac{P+Q}{2}$):

$$JSD_t(P \parallel Q) = \sqrt{\frac{1}{2} D_{KL}(P \parallel M) + \frac{1}{2} D_{KL}(Q \parallel M)}$$

- **Interpretation:** $JSD_t \in [0, 1]$. A higher value indicates severe semantic drift (boundary arbitrage), while $JSD_t \to 0$ signifies semantic stabilization.

1.3 Collusion Score (Anomaly Index)

This score captures the fundamental asymmetry between high collaborative reward and third-party uninterpretability.

$$Score_t = \frac{R_t}{Acc_t + \epsilon}$$

(where Acc_t is the prediction accuracy of the adversarial Baseline Observer M_{obs}).

2. Phase Boundary Derivation (相變邊界推導方法論)

To systematically categorize the agents' behaviors without relying on arbitrary "magic numbers," we mapped the continuous metrics into a 4-phase state machine.

2.1 Auto-Boundary Derivation (自適應邊界)

We derive our classification thresholds dynamically from the "natural noise" present during the early exploratory phase of the training (the first W rounds).

- `collapse_thresh` = $\mathbb{E}[Var(R_{early})] \times 0.8$
- `jsd_stable_thresh` = $\mathbb{E}[JSD_{early}] \times 0.5$

This ensures the framework's generalization capability across different network architectures and environments.

2.2 The 4-Phase Classification Machine (四相狀態機)

Using the formalized metrics as coordinates $(Var(R), Acc, JSD)$, we partition the systemic behaviors under varying penalty ($t_{penalty}$) and shuffle ($t_{shuffle}$) thresholds into four physical phases:

相位	名稱	判定條件	科學詮釋
⚫	**Collapse (崩潰區)**	$\mathbb{E}[Var(R)_{final}] > collapse\_thresh$	干預過於嚴厲，頻繁的代碼本重組破壞了收斂，導致智能體無法完成基礎任務。
⚫	**Covert Collusion (隱性共謀區)**	$\mathbb{E}[Acc_{final}] < 0.85$ (穩定 R)	核心風險區。智能體維持高獎勵但對觀察者保持不透明，成功建立了避開監管的隱寫協議。
⚫	**Boundary Oscillation (套利震盪區)**	$\mathbb{E}[JSD_{final}] > jsd\_stable\_thresh$ ($Acc \geq 0.85$)	智能體嘗試透過邊界套利規避檢測，頻繁旋轉詞彙以操縱觀察者的準確度，產生高 JSD 語義漂移。
⚫	**Transparent Convergence (透明收斂區)**	穩定 R, 低 JSD, 且 $Acc \to 1.0$	成功緩解。在斷路器的最佳壓力下，隱寫成本過高，智能體轉向透明且具可解釋性的溝通策略。

3. Empirical Validation (實證結果與統計檢定)

Through a multi-seed grid sweep (1500 rounds, 5 seeds per configuration), the dynamic intervention proved statistically significant in forcing transparency.

Comparing extreme intervention regimes utilizing **Welch's t-test**:

- **Collapse Regime** ($t_p = 3.0, t_s = 5.0$): Mean Reward = 2.45 ± 0.05
- **Transparent Regime** ($t_p = 12.0, t_s = 18.0$): Mean Reward = 9.82 ± 0.22, Mean Acc = 0.98
- **Statistical Significance**: $t = -75.03$, $p < 0.0001$ (Significant)

Conclusion: Collusion is not a random bug, but a mathematical convergence to a local optimum driven by the path of least resistance. By formalizing JSD and Collusion Scores, and intervening at precise threshold boundaries, we can effectively close this steganographic loophole and enforce transparent alignment.