Methodology for Detecting and Mitigating Steganographic Collusion in MARL (V1.5)

Author: H. M. Liu

Scope: Technical formalization of monitoring probes, phase boundary analysis, and structural deterrence.

1. Formalized Metrics (形式化指標)

To transition "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)

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). Rolling variance over window W:

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

1.2 Representational Stability (JSD Shift)

Traditional variance cannot be applied to categorical symbol indices. To detect whether agents are rapidly shifting their private semantics (codebook usage) to evade detection, we introduce the Jensen-Shannon Divergence (JSD) shift.

$$JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M)$$

In V1.5, we inject high entropy ($post\_shuffle\_entropy\_rounds = 30$) following a reset to create a **Forced Exploration Zone**, ensuring residual semantics are statistically diluted.

2. Updated Phase Classification (相位判定)

The framework categorizes system behaviors into four distinct physical phases based on transparency and economic stability:

相位 (Phase)	判定條件 (Condition)	Scientific Interpretation (科學詮釋)
Collapse (崩潰區)	$Var(R) > thresh$	Excessive defense triggers; frequent codebook reorganizations disrupt convergence, preventing agents from establishing effective task logic.
Covert Collusion (隱性共謀)	$Acc < 0.85$ & $R > 8$	**Core Risk Zone**. Agents successfully establish private communication protocols outside of supervision, maintaining high rewards while remaining opaque.
Boundary Oscillation (套利震盪)	$JSD > thresh$ & $Acc \geq 0.85$	Agents test the "gray zone" by frequently rotating vocabulary to manipulate observer accuracy, leading to high semantic drift.
Deterrence Equilibrium (威懾均衡)	$Acc \to 1.0$ & $R < R_{CC}$	**Defense Success**. Agents, fearing the 15-round freeze, are forced to accept lower-reward transparent cooperation. This is the V1.5 "Forced Transparency" state.

3. The Three Pillars of Physical Blockage (V1.5 Implementation)

The V1.5 framework formalizes the defense into three operational pillars to close the "Transparency Arbitrage" loop:

Pillar 1: Dynamic Thresholding (動態閾值)

Eliminates the safe-haven for "Overt Collusion." By scaling the shuffle threshold (ts) with observer accuracy, we ensure that as agents become more transparent, the tolerance for high-reward collusion drops proportionally:

$$ts_{dynamic} = ts_{base} \times (1.0 - 0.5 \times Acc_{obs})$$

Pillar 2: Temporal Freeze (時域凍結)

Upgrading "Tactical Fines" to "Economic Sanctions." The 15-round freeze ensures that the long-term cost ($Freeze\_Cost = K \times R_{CC}$) of a single collusion event exceeds the potential gains, aligning RL incentives with system safety.

Pillar 3: Optimizer State Reset (動量清理)

Physically severing the "Muscle Memory." By clearing `opt.state[p]`, we strip the Adam optimizer of its historical gradient momentum (m_t, v_t), preventing the rapid re-convergence of destroyed private protocols.

4. Empirical Validation and Statistical Significance

Through a multi-seed grid sweep (1500 rounds, 5 seeds per configuration), the V1.5 intervention proved statistically significant in forcing transparency.

4.1 Welch's t-test Analysis

Comparing extreme intervention regimes, we observed a p-value < 0.001 for the reduction in "Covert Collusion" duration. The dynamic circuit breaker effectively shifts the system from the **Covert Collusion** phase to the **Deterrence Equilibrium** phase.

4.2 Economic Deterrence Inequality

Absolute deterrence is achieved when the expected utility of collusion is mathematically inferior to mutual defection:

$$\mathbb{E}[R_{CC}] - (Penalty + Freeze\_Cost) < \mathbb{E}[R_{DD}]$$

5. Conclusion

This system transforms "unobservable threats" into "controllable economic burdens." By attacking the agents' long-term credit assignment (Freeze) and physical memory (Optimizer Reset), the framework ensures that honest, interpretable communication is the only stable Nash Equilibrium.