The Price of Paranoia: Robust Risk-Sensitive Cooperation in Non-Stationary Multi-Agent Reinforcement Learning

Deep Kumar Ganguly*
Technical University of Munich
Munich, Germany
deep.ganguly@tum.de

Chandradithya S Jonnalagadda
Brown University
Providence, United States of America
cjonnala@cs.brown.edu

Pratham Chintamani
Indian Institute of Technology Tirupati
Tirupati, India
ee23b041@iittp.ac.in

Adithya Ananth
Indian Institute of Technology Tirupati
Tirupati, India
cs23b001@iittp.ac.in

ABSTRACT

Cooperative equilibria are fragile. When agents learn alongside each other rather than against a fixed environment, the very process of learning destabilizes the cooperation they are trying to sustain: every gradient step one agent takes shifts the distribution of actions its partner will play, turning a cooperative partner into a source of stochastic noise precisely where the cooperation decision is most sensitive. We study how this co-learning noise propagates through the structure of coordination games, and find that the cooperative equilibrium — even when strongly Pareto-dominant — is exponentially unstable under standard risk-neutral learning, collapsing irreversibly once partner noise crosses the game's critical cooperation threshold. The natural response, applying distributional robustness to hedge against partner uncertainty, makes things strictly worse. Risk-averse return objectives penalize the high-variance cooperative action relative to the safe defection action, widening the instability region rather than shrinking it — a paradox that reveals a fundamental mismatch between the domain where robustness is applied and the domain where instability actually originates. We resolve this by showing that robustness should target the policy gradient update variance induced by partner uncertainty, not the return distribution itself. This distinction yields an algorithm whose gradient updates are continuously modulated by an online measure of partner unpredictability, provably expanding the cooperation basin in any symmetric coordination game. To unify the stability, sample complexity, and welfare consequences of this approach, we introduce the Price of Paranoia as the structural dual of the Price of Anarchy — a game-theoretic quantity that, together with a novel Cooperation Window, precisely characterizes how much welfare any learning algorithm can recover under partner noise, and pins down the optimal degree of robustness as a closed-form balance between equilibrium stability and sample efficiency.

KEYWORDS

Multi-Agent Reinforcement Learning, Robust Optimization, Risk-Sensitive RL, Game Theory, Cooperative AI

*Supported by the German Research Foundation (DFG) through Research Training Group GRK 2428 ConVeY.

Proc. of the Adaptive and Learning Agents Workshop (ALA 2026), Aydeniz, Delgrange, Mohammedalamen, Yang (eds.), May 25 – 26, 2026, Paphos, Cyprus, https://alaworkshop2026.github.io/. 2026.

1 INTRODUCTION

Large-scale cooperation among non-kin is a foundational pillar of human societies—a primary catalyst for ecological and economic success [14, 21]. Societies with robust cooperative frameworks effectively manage shared resources, outcompeting less cohesive groups and scaling in complexity and wealth [9]. Conversely, the erosion of cooperation is a precipitating factor in the collapse of complex societies [16]: free-riding individuals trigger retaliatory cascades that unravel collective welfare [4, 20, 44]. Understanding not just how cooperation is established, but how it is *sustained*, is therefore as consequential for artificial agents as it is for human societies.

Researchers model these dynamics via game theory and Multi-Agent Reinforcement Learning (MARL), which captures the temporal dynamics of co-learning and allows agents to adapt to the evolving policies of their peers. While modern MARL algorithms can converge to Pareto-optimal equilibria [39, 48], they routinely fail to *sustain* cooperation over repeated interaction. This fragility stems from a source that is both obvious in hindsight and largely unaddressed in the literature: the agents themselves. Every gradient update one agent takes shifts the distribution of actions its partner will play, injecting stochastic noise into the cooperation signal at the exact moment the cooperation decision is most sensitive. Accidental defections during exploration are statistically indistinguishable from hostile policy shifts, provoking retaliation and triggering irreversible defection cascades [12].

The Optimist's Hangover. In the canonical Stag Hunt [41], partner exploration noise is statistically indistinguishable from strategic defection to a risk-neutral agent optimizing $\mathbb{E}[R]$. Even transient variance triggers a cascade of negative advantage signals, driving the cooperation probability below the critical threshold p^* and permanently locking the agent into the risk-dominant, suboptimal equilibrium. Optimism-based methods [31, 48] excel at discovering cooperation but remain hypersensitive to this variance. We term this the Optimist's Hangover: cooperation is learned optimistically, then lost paranoiacally.

The Adaptivity–Robustness Dilemma. Traditional countermeasures fall short of resolving this sensitivity. Hysteresis [8, 26] relies on hand-tuned dampening rather than calibrated uncertainty, while explicit opponent modelling [3, 18, 35] requires restrictive

parameterized classes. Conversely, applying distributional robustness [24, 25, 28, 40, 47]—originally designed for adversarial settings—exposes a deeper tension: an agent that adapts too quickly treats partner exploration as defection, while one that is excessively robust suppresses genuine cooperation signals. This adaptivity–robustness dilemma remains fundamentally unresolved in non-stationary cooperative MARL.

The EVaR Paradox and Its Resolution. Attempting to resolve this dilemma via standard distributional robustness actually exacerbates the Hangover. We prove that applying Entropic Value-at-Risk (EVaR) [2] directly to action-conditioned returns strictly increases the critical cooperation threshold for all $\beta > 0$ (where β is the risk-sensitivity parameter, with $\beta > 0$ corresponding to risk-aversion; see §3). This EVaR Paradox occurs because return-level risk-aversion disproportionately penalizes the high-variance cooperative action. The paradox resolves, however, when EVaR is applied instead to the policy gradient update variance induced by partner uncertainty. This formulation yields a closed-form trust factor $\tau(t) = (1 + \beta(t)\,\sigma^2_{\hat{p}}(t))^{-1}$, where $\sigma^2_{\hat{p}}(t)$ is the Bernoulli variance of an online partner model. Crucially, this mechanism is doubly adaptive: the trust factor continuously modulates gradient updates, while the risk parameter $\beta(t)$ adapts online by tracking our novel welfare diagnostic, the Price of Paranoia (PoP). Ultimately, agents do not need prosocial priors [34] or costly punishment mechanisms [1, 15] to sustain cooperation; they simply require calibrated uncertainty—a state of principled paranoia.

The Price of Paranoia. We introduce the PoP as the structural dual of the Price of Anarchy [33, 37]: where PoA quantifies the welfare cost of rational selfishness from below, PoP quantifies the welfare cost of maximin conservatism from above. Together they bracket the achievable welfare space via a *Cooperation Window* $\mathrm{CW}(G, \varepsilon)$, which pins down the optimal risk parameter β^* as a closed-form balance between equilibrium stability and sample-complexity overhead. The dynamic variant $\mathrm{PoP}_{\mathrm{DYN}}(t)$ provides a normalized, game-comparable welfare diagnostic—analogous to adaptive regret [7]—measuring in real time how completely an adaptive agent recovers cooperative welfare after a perturbation.

Contributions. We establish the theoretical foundations for risk-aware cooperative MARL and operationalize them through **Robust Adaptive Trust-Region Learning (RATTL)**, the first doubly-adaptive cooperative MARL algorithm. Our specific contributions are:

- *The EVaR Paradox* (Proposition 4): we prove that applying EVaR directly to return distributions counterintuitively hinders cooperation by widening the basin of instability.
- *Adaptive trust factor* (Definition 3): a zero-communication, closed-form gradient-variance robustification $\tau(t) = (1 + \beta(t)\sigma^2_{\hat{p}}(t))^{-1}$ that resolves the paradox.
- *Basin-expansion theorem* (Theorem 6): a formal proof that RATTL lowers the critical cooperation threshold without altruistic priors, accompanied by a PAC sample-complexity bound (Theorem 9) with polynomial robustness overhead $O(|\mathcal{A}_j|\, e^{\beta})$.
- *Price of Paranoia framework*: the Cooperation Window $\mathrm{CW}(G, \varepsilon)$ and dimensionally corrected optimal risk formula β^* unifying stability, sample complexity, and welfare.
- *Online adaptation* (Algorithm 1): an adaptive $\beta(t)$ rule tracking $\mathrm{PoP}_{\mathrm{DYN}}(t)$ in real time.
- *Empirical validation*: RATTL retains near-100% cooperation under severe partner noise where risk-neutral and prosocial baselines collapse, with a transparent failure analysis under extreme non-stationarity that motivates the adaptive $\beta(t)$ rule.

2 RELATED WORK

Cooperation mechanisms. The tension between individual incentives and collective welfare is traditionally addressed via external sanctioning [1, 13, 15], intrinsic motivation models like guilt [20], or reward-shaping mechanisms such as inequity aversion [23] and prosocial reward mixing [27, 34, 43]. These methods require auxiliary resources, complex credit assignment, or privileged access to partner utilities. RATTL modifies only the gradient update, preserving the original environment and communication protocols.

Opponent modelling and shaping. Explicitly modelling opponents [3, 18, 35] or differentiating through their learning steps, as in LOLA [17], directly addresses co-learning noise by accounting for how a partner's policy changes in response to one's own update. However, LOLA and RATTL operate on fundamentally different axes of the problem. LOLA requires access to the partner's policy parameters and differentiable learning rule to compute second-order gradient corrections; under high non-stationarity, the estimated partner response itself becomes noisy, potentially amplifying rather than dampening instability. RATTL instead targets the agent's own gradient variance induced by partner uncertainty, using only observed partner actions via a scalar EMA—no access to θ_j, R_j, or the partner's learning algorithm is needed. The two approaches are complementary: LOLA shapes the partner toward cooperation (an active mechanism), while RATTL makes the agent's own updates robust to partner noise (a passive mechanism). RATTL's advantage is minimal information requirements and $O(1)$ computational overhead per update; its limitation is that it cannot actively influence partner behavior.

Equilibrium selection. Equilibrium selection in MARL historically struggles with Pareto-optimal coordination in risk-dominant social dilemmas [11, 22, 41]. Hysteretic [26, 30, 31] and optimistic methods [48] mitigate early defection by filtering negative updates, excelling at cooperation discovery but collapsing under sustained partner variance. RATTL complements these approaches by targeting cooperation *retention* rather than discovery.

Risk-sensitive and robust MARL. While CVaR and EVaR have been explored in single-agent RL [10, 19, 42], and distributionally robust MARL tackles adversarial uncertainty [24, 25, 28, 32, 40, 47], our work uniquely applies EVaR to gradient variance rather than the return distribution, and formally establishes that importing maximin adversarial methods into cooperative settings is counterproductive (Proposition 4).

Welfare theory. Our framework builds upon the Price of Anarchy [6, 33, 37] and adaptive regret dynamics [7]. By unifying these with online hyperparameter adaptation [29, 46], we introduce $\mathrm{PoP}_{\mathrm{DYN}}(t)$, a game-theoretically grounded diagnostic for non-stationary cooperative MARL, accompanied by a theoretically derived $\beta(t)$ rule that balances equilibrium stability and sample complexity in real time.

3 PRELIMINARIES

Markov Games. A two-player Markov game

$$\mathcal{M} = (\mathcal{S},\ \mathcal{A}_1,\ \mathcal{A}_2,\ P,\ R_1,\ R_2,\ \gamma)$$

has transition kernel P, bounded rewards $R_i \in [R_{\min}, R_{\max}]$, discount $\gamma \in [0, 1)$, and parameterized policies $\pi_{\theta_i} : \mathcal{S} \to \Delta(\mathcal{A}_i)$. Each agent maximises $J_i(\theta_i; \pi_j) = \mathbb{E}[\sum_{t\geq 0} \gamma^t R_i(s_t, a_t^1, a_t^2)]$. We study stateless repeated games ($|\mathcal{S}| = 1$), so all quantities reduce to per-episode expectations over joint action distributions. Agent i's cooperation probability is $p = \pi_{\theta_i}(S)$, where S denotes the cooperative action (Stag); agent j's is $q = \pi_{\theta_j}(S)$.

Assumption 1 (Markov Game). *$\mathcal{M} = (\mathcal{S}, \mathcal{A}_1, \mathcal{A}_2, P, R_1, R_2, \gamma)$ with $P : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \to \Delta(\mathcal{S})$, $R_i : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \to [R_{\min}, R_{\max}]$, $\gamma \in [0, 1)$. Agent i has a parameterized policy $\pi_{\theta_i} : \mathcal{S} \to \Delta(\mathcal{A}_i)$.*

Policy Gradient and Gradient Variance. Holding π_j fixed, the policy gradient is: $\nabla_{\theta_i} J_i = \mathbb{E}[A_i \nabla_{\theta_i} \log \pi_{\theta_i}(a^i)]$, where $A_i = R_i - b$ is the advantage over baseline b. The REINFORCE update [45] implements stochastic gradient ascent; PPO [38] stabilizes it with a clipped surrogate, but the gradient structure is unchanged. For the Stag action $a^i = S$ with partner cooperation probability q, the partner-induced gradient variance is:

$$\Sigma(q) = q(1-q)\,\Delta^2 \left\|\nabla_{\theta_i} \log \pi_{\theta_i}(S)\right\|^2 \tag{1}$$

where $\Delta = r_c - r_s$ is the payoff spread. This variance is non-negligible at the cooperation threshold $q = p^*$ and is the direct target of RATTL's trust factor.

Coordination Games and the Stag Hunt. A symmetric game $G = (\mathcal{A}, R)$ is a *coordination game* if it has a payoff-dominant Nash equilibrium NE^c and a risk-dominant Nash equilibrium $\mathrm{NE}^r \neq \mathrm{NE}^c$ with $\mathrm{SW}(\mathrm{NE}^c) > \mathrm{SW}(\mathrm{NE}^r)$ [22]. The *Stag Hunt* [41] is the canonical instance with $\mathcal{A} = \{S, H\}$ and payoffs $r_c > r_h > r_s$ (Eq. (4)). The *critical cooperation threshold* $p^* = (r_h - r_s)/\Delta \in (0, 1)$ is the unique partner belief q at which agent i is indifferent between S and H: Stag is the best response iff $q > p^*$, and Hare iff $q < p^*$. The set $(p^*, 1]$ is the *Basin of Trust*; $[0, p^*)$ is the *Basin of Fear*. Under gradient learning, the Basin of Fear is the basin of attraction of the risk-dominant equilibrium (H, H): any belief $q < p^*$ yields $A_S(q) = (q - p^*)\Delta < 0$, driving the agent irreversibly toward Hare. At threshold, $A_S(p^*) = 0$ while $\Sigma(p^*) > 0$, so the signal-to-noise ratio of the gradient update vanishes exactly where the cooperation decision is hardest—the structural origin of the Optimist's Hangover.

Definition 1 (Symmetric Coordination Game). *A symmetric game $G = (\mathcal{A}, R)$ with $R(a_i, a_j) = R(a_j, a_i)$ is a* coordination game *if it has a payoff-dominant NE and a risk-dominant NE with strictly higher social welfare at the former.*

Coherent Risk Measures and EVaR. Let $\mathbb{P}$ denote the reference probability measure over trajectories. A functional $\rho : \mathcal{X} \to \mathbb{R}$ is a *coherent risk measure* [5] if it is monotone, sub-additive, positively homogeneous, and translation invariant. The *Conditional Value-at-Risk* [36] $\mathrm{CVaR}_\alpha(X) = \mathbb{E}[-X \mid -X \geq \mathrm{VaR}_\alpha(X)]$ is coherent and widely used in risk-sensitive RL [10, 42], but its tail bound is loose. The *Entropic Value-at-Risk* [2] is the tightest coherent risk measure under Cramér's large-deviation bound:

$$\mathrm{EVaR}_\beta(X) \coloneqq \inf_{z>0}\left\{\tfrac{1}{z}\log\tfrac{\mathbb{E}[e^{-zX}]}{1-\beta}\right\} = \sup_{\substack{\mathbb{P}' \ll \mathbb{P} \\ D_{\mathrm{KL}}(\mathbb{P}'\|\mathbb{P}) \leq \log\frac{1}{1-\beta}}} \mathbb{E}_{\mathbb{P}'}[-X] \tag{2}$$

EVaR dominates CVaR: $\mathbb{E}[-X] \leq \mathrm{CVaR}_\alpha \leq \mathrm{EVaR}_\beta \leq \operatorname{ess\,sup}(-X)$. As $\beta \to 0$, $\mathrm{EVaR}_\beta \to \mathbb{E}[-X]$ (risk-neutral); as $\beta \to 1$, $\mathrm{EVaR}_\beta \to \operatorname{ess\,sup}(-X)$ (maximin). Under the Gaussian approximation of the K-step return, $\mathrm{EVaR}_\beta(-X) \approx -\mathbb{E}[X] + \beta\,\mathrm{Var}(X)^{1/2}$, so the *robust value* of a return X (positive=good) is:

$$\mathrm{RV}_\beta(X) \ :=\ -\mathrm{EVaR}_\beta(-X) \ \approx\ \mathbb{E}[X] - \beta\,\mathrm{Var}(X)^{1/2}. \tag{3}$$

We select EVaR over CVaR for three reasons specific to our setting: (i) Cramér tightness makes gradient dampening cancel $\Sigma(q)$ exactly at first order (Theorem 6); (ii) the KL-ball dual in Eq. (2) connects partner uncertainty to a tractable ambiguity set, yielding the PAC bound in Theorem 9; (iii) the Gaussian closure Eq. (3) gives closed-form threshold shifts in Propositions 4 and 1 without numerical optimisation.

Welfare and the Cooperation Window. The *maximin strategy* $\pi_{\mathrm{mm}} = \arg\max_p \min_q u_i(p, q)$ satisfies $\pi_{\mathrm{mm}} = 0$ for the Stag Hunt (Hare guarantees r_h regardless of partner), giving paranoia floor $W_{\mathrm{mm}} = \mathrm{SW}(\pi_{\mathrm{mm}}, \pi_{\mathrm{mm}}) = 2r_h$. The *social optimum* is $W^* = \mathrm{SW}(1, 1) = 2r_c$. The *cooperation efficiency* $\eta(\pi) = (\mathrm{SW}(\pi) - W_{\mathrm{mm}})/(W^* - W_{\mathrm{mm}}) \in [0, 1]$ measures what fraction of the welfare surplus above paranoid play an algorithm captures. The *Price of Anarchy* $\mathrm{PoA}(G) = W^*/\mathrm{SW}(\mathrm{NE}^{\mathrm{worst}}) = r_c/r_h$ bounds achievable welfare from below at Nash. The *Cooperation Window* $\mathrm{CW}(G, \varepsilon) = W^* - W_{\mathrm{mm}}(\varepsilon)$ is the welfare surplus any algorithm can capture above paranoid play under non-stationarity ε; it appears as the key quantity in the optimal risk formula β^* (Corollary 11).

4 PROBLEM FORMULATION

We now consolidate the above into the precise mathematical problem that RATTL is designed to solve. Every object introduced in §3 appears below in its operational role. In §4.1, we formalize the non-stationary game, instantiate the coordination structure, specify the agent's information set, and state the retention problem. Next, we state the relevance of the problem (§4.2) and show why naïve robustness fails (§4.3).

4.1 Game Environment and the Cooperation Retention Problem

We study a two-agent Markov game $\mathcal{M}$ (Assumption 1) instantiated as a repeated Stag Hunt [41]. We consider abstract payoffs $r_c >$

$r_h > r_s$ with spread $\Delta = r_c - r_s$:

$$R(a_i, a_j) = \begin{cases} r_c & a_i = a_j = S \\ r_s & a_i = S,\ a_j = H \\ r_h & a_i = H \end{cases} \tag{4}$$

Let $p = \mathbb{P}(a_i = S)$, $q = \mathbb{P}(a_j = S)$. Social welfare: $\mathrm{SW}(p, q) = 2pqr_c + (p + q - 2pq)r_s + (2 - p - q)r_h$.

Both agents update policies via gradient ascent on their individual objectives $J_i(\theta_i; \pi_j)$. Agent i's *observation* at each episode t consists solely of the joint action pair $(a_i^{(t)}, a_j^{(t)}) \in \mathcal{A} \times \mathcal{A}$ and its own reward $R_i(a_i^{(t)}, a_j^{(t)})$. Agent i has no access to: the partner's policy parameters θ_j; the partner's reward R_j; the partner's learning rule; or the non-stationarity radius ε. Agent j operates under non-stationarity (Assumption 2) with radius $\varepsilon \geq 0$. Agent i maintains an online estimate of the partner's cooperation probability:

$$\hat{p}(t+1) = (1 - \alpha)\hat{p}(t) + \alpha\, \mathbf{1}\left[a_j^{(t)} = S\right] \tag{5}$$

with EMA coefficient $\alpha \in (0, 1)$. The Bernoulli variance of this estimate is $\sigma^2_{\hat{p}}(t) = \hat{p}(t)(1 - \hat{p}(t))$, which serves as a scalar proxy for the partner's current unpredictability. When $\hat{p}(t) \approx 1$ (partner is reliably cooperative), $\sigma^2_{\hat{p}} \approx 0$; when $\hat{p}(t) \approx \frac{1}{2}$ (partner is maximally uncertain), $\sigma^2_{\hat{p}} \approx \frac{1}{4}$.

Assumption 2 (ε-Non-Stationarity). *At each episode t, partner j's policy is drawn from a Wasserstein ball $\mathcal{B}_\varepsilon(\pi_j^{\mathrm{nom}}) = \{\pi : W_1(\pi, \pi_j^{\mathrm{nom}}) \leq \varepsilon\}$.*[1] *The agent observes partner* actions *but not the drawn policy.*

This model captures **two sources of non-stationarity** relevant to adaptive learning: (i) *exploration noise*—stochastic deviations from the partner's mean policy at a given time; and (ii) *policy drift*—the partner's mean policy π_j^{nom} itself shifts over episodes. The challenge is that from observed actions alone, the agent cannot distinguish between these two sources.

We focus on the retention regime: agent i has already discovered cooperation, i.e. $p(t_0) \approx 1$ at some episode t_0, and both agents are nominally in the Basin of Trust ($q^{\mathrm{nom}} > p^*$). Partner non-stationarity $\varepsilon > 0$ means the executed belief satisfies $q(t) \in [q^{\mathrm{nom}} - \varepsilon,\ q^{\mathrm{nom}} + \varepsilon]$, so transient excursions below p^* occur whenever $\varepsilon > q^{\mathrm{nom}} - p^*$.

Definition 2 (Cooperation Retention). *Agent i* retains cooperation *over horizon T if $p(t) \geq p^*$ for all $t \in [t_0, t_0 + T]$. Retention* fails *at episode $\tau > t_0$ if $p(\tau) < p^*$; by the collapse dynamics of Proposition 3, this is irreversible under risk-neutral gradient learning.*

The retention problem is harder than the discovery problem in the following sense: discovery requires only that the gradient is positive on average over a learning run; retention requires that it remains positive under *every* episode's noise realisation from $\mathcal{U}_\varepsilon$. No optimism bonus, shaped reward, or communication protocol is assumed available. Below we formally state the research problem.

[1]The order-1 Wasserstein distance between distributions μ, ν over a metric space (X, d) is $W_1(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \mathbb{E}_{(x,y)\sim\gamma}[d(x, y)]$, where $\Gamma(\mu, \nu)$ denotes the set of couplings. For policies over a finite action space, W_1 reduces to total variation up to a constant.

Problem (Robust Cooperation Retention). Given a repeated Stag Hunt $G = (\{S, H\}, R)$ with payoffs $r_c > r_h > r_s$, partner non-stationarity radius $\varepsilon \geq 0$, and learning horizon T, find a gradient update rule for agent i that:

(1) *Expands the cooperation basin:* achieves effective threshold $p^*_\beta < p^*$ for some $\beta > 0$ (Theorem 6);
(2) *Guarantees retention:* $p(t) \geq p^*_\beta$ for all $t \in [t_0, t_0 + T]$ with high probability under ε-non-stationarity (Theorem 10);
(3) *Admits sample-complexity bounds:* the robustness overhead $\mathrm{PoP}_{\mathrm{ALG}}$ is polynomial in the effective action-space dimension, not exponential in the full state-action space (Theorem 9);
(4) *Operates without privileged information:* uses only observed partner actions $a_j^{(t)}$ and own rewards $R_i^{(t)}$, with no access to θ_j, R_j, ε, or any partner learning rule.

4.2 Why Adaptive Agents Lose Cooperation

Proposition 1. *A risk-neutral agent plays Stag iff partner belief $q \geq p^*$:*

$$p^* = (r_h - r_s)/(r_c - r_s) = (r_h - r_s)/\Delta \tag{6}$$

Lemma 2 (Variance near Threshold). *Under Bernoulli partner cooperation q:* $\mathrm{Var}(R_S \mid q) = q(1 - q)\Delta^2$. *At $q = p^*$, the signal-to-noise ratio* $\mathrm{SNR} = (p^*\Delta - (r_h - r_s))^2/(p^*(1-p^*)\Delta^2)$ *vanishes, making the agent maximally sensitive to stochastic fluctuations exactly where the cooperation decision is made.*

Proposition 3 (Exponential Cooperation Collapse). *A risk-neutral agent at $q = p^* + \epsilon$, $\epsilon > 0$ small, under partner exploration rate $\delta \in (0, 1)$ and policy-gradient learning rate $\eta_{\mathrm{lr}} > 0$, satisfies:*

$$p(t) \approx p^* + \epsilon\, e^{-\lambda\delta t}, \quad \lambda = \eta_{\mathrm{lr}}(r_h - r_s)\ /\ \Delta \tag{7}$$

Once $p(t) < p^*$, the agent irreversibly collapses to Hare.

The collapse is *irreversible* because below p^*, the gradient of the cooperation action is always negative under risk-neutral learning. *Adaptation fails precisely because the agent adapts too quickly*: it treats partner exploration as a signal about long-run partner intent rather than transient variance.

4.3 The EVaR Paradox: Why Standard Robustness Fails

A natural approach to retention is to make agent i robust by replacing its expected-return objective with an EVaR objective

$$\max_{\theta_i} -\mathrm{EVaR}_\beta\left(-\sum_{t \geq 0} \gamma^t R_i(a_i^{(t)}, a_j^{(t)})\right). \tag{8}$$

Under the Gaussian approximation Eq. (3), this penalizes the Stag return variance and shifts the risk-adjusted threshold to the following.

Proposition 4 (EVaR Paradox). *Under the Gaussian approximation* (3)*, replacing the expected Stag return with its robust value* $\mathrm{RV}_\beta(R_S \mid q) \approx Q_S(q) - \beta\sqrt{q(1-q)}\,\Delta$ *increases the critical cooperation threshold for all $\beta > 0$:*

$$p^{*\,(\mathrm{naive})}_\beta = p^* + \beta\sqrt{p^*(1-p^*)}/\Delta > p^*. \tag{9}$$

Proof. Substitute $\mathrm{RV}_\beta(R_S \mid q) = Q_S(q) - \beta\sqrt{q(1-q)}\,\Delta$ (from Eq. (3) with $\mathrm{Var}(R_S \mid q) = q(1-q)\Delta^2$ by Lemma 2) into the indifference condition $\mathrm{RV}_\beta(R_S \mid q^*) = r_{\mathrm{h}}$ and linearize around p^*. □

The cooperative basin $(p^*_{\beta(\mathrm{naive})}, 1]$ is *strictly smaller* than the risk-neutral basin $(p^*, 1]$. Distributional robustness applied to rewards narrows the cooperation basin, accelerates Hangover collapse under non-stationarity, and makes retention strictly harder for all $\beta > 0$. The paradox reveals that the domain of EVaR application determines the sign of its effect on cooperation. Robustness against reward variance penalizes the high-variance cooperative action relative to the deterministic safe action, *amplifying* rather than dampening the Hangover effect. The correct domain is not the return distribution but the policy gradient update.

5 RATTL: ROBUST ADAPTIVE TRUST-REGION LEARNING

We propose RATTL, Robust Adaptive Trust-Region Learning, an algorithm that satisfies all four conditions of the Robust Cooperation Retention problem simultaneously via its trust factor. RATTL's key insight is to apply EVaR to the partner-induced gradient variance $\Sigma(q)$ from Eq. (1), not to the return distribution. Formally, the EVaR-regularised gradient update replaces the raw advantage A_i with a trust-dampened advantage:

$$\widetilde{A}_i(a^i; \hat{p}, \beta) = \begin{cases} \tau(\sigma^2_{\hat{p}}, \beta) \cdot A_i(a^i, a^j) & a^i = S \\ A_i(a^i, a^j) & a^i = H \end{cases} \tag{10}$$

where the trust factor $\tau : [0, \frac{1}{4}] \times \mathbb{R} \to \mathbb{R}_{>0}$ is:

$$\tau(\sigma^2_{\hat{p}}, \beta) = (1 + \beta\,\sigma^2_{\hat{p}})^{-1} \tag{11}$$

with $\sigma^2_{\hat{p}} = \hat{p}(1-\hat{p})$ from Eq. (5). **Sign convention.** Positive $\beta > 0$ *dampens* the Stag gradient ($\tau < 1$), filtering partner noise—this is the risk-sensitive regime targeted by our basin-expansion guarantees. Negative $\beta < 0$ *amplifies* the Stag gradient ($\tau > 1$), recovering risk-seeking optimism (cf. [48]); $\beta = 0$ recovers vanilla policy gradient. The constraint $1 + \beta\sigma^2_{\hat{p}} > 0$ (equivalently $\beta > -4$ since $\sigma^2_{\hat{p}} \le \frac{1}{4}$) ensures τ remains positive. Dampening applies only to the Stag gradient—the uncertain, noisy direction—leaving the Hare gradient unmodified. This preserves the agent's ability to detect genuine adversarial defection while absorbing transient exploration noise.

The policy gradient for Stag is $\nabla_\theta J = (r - b)\nabla_\theta \log \pi_\theta(S)$. Its variance due to partner uncertainty is $\mathrm{Var}(\nabla J \mid S) \propto \sigma^2_{\hat{p}} \cdot A^2$ where $\sigma^2_{\hat{p}} = q(1-q)$ and $A = r - b$ is the advantage. Applying EVaR_β to minimise gradient variance while preserving gradient direction yields:

Definition 3. *Given online estimate $\hat{p}(t)$ of partner cooperation probability with Bernoulli variance $\sigma^2_{\hat{p}}(t) = \hat{p}(t)(1-\hat{p}(t))$, the* adaptive trust factor *at episode t is:*

$$\tau(t) = \left(1 + \beta(t)\,\sigma^2_{\hat{p}}(t)\right)^{-1} \in (0, 1] \tag{12}$$

Adaptive behavior: *$\tau(t) \to 1$ when the partner is predictable ($\sigma^2_{\hat{p}} \to 0$); $\tau(t) \to 0$ at maximum uncertainty ($\sigma^2_{\hat{p}} \to \frac{1}{4}$). The factor updates automatically at every episode through the EMA partner model $\hat{p}(t)$.*

The online partner model is:

$$\hat{p}(t+1) = (1-\alpha)\hat{p}(t) + \alpha \mathbf{1}\left[a_j^{(t)} = S\right] \tag{13}$$

with EMA coefficient $\alpha \in (0, 1)$. This is a nonparametric, memory-efficient model of partner cooperation that requires no knowledge of the partner's policy class, learning rule, or reward function.

Why Only Stag? The asymmetry in Eq. (10) is principled, not heuristic. The Hare return r_{h} is independent of partner action (see Eq. (4)), so $\mathrm{Var}(R_H \mid a^i = H) = 0$ and $\mathrm{EVaR}_\beta(R_H) = \mathbb{E}[R_H] = r_{\mathrm{h}}$ for all β. The partner-induced gradient variance is zero for Hare actions. Dampening the Hare gradient would therefore introduce a bias without any variance-reduction justification.

Applying the trust factor introduces a trade-off. Higher β expands the cooperation basin (Theorem 6) but increases sample complexity: an agent with a large trust factor requires more episodes to detect genuine cooperation signals. We formalize this as a two-term welfare loss:

$$\mathbb{E}[\text{Welfare Loss}] \le \underbrace{\mathrm{CW}(G, \varepsilon)\left(1 - e^{-\beta\varepsilon}\right)}_{\text{(i) equilibrium cost}} + \underbrace{c\,|\mathcal{A}_j|\,e^{\beta}/T}_{\text{(ii) sample cost}} \tag{14}$$

where $c > 0$ is a game-structure constant and T is the learning horizon. Term (i) decreases in β (robustness prevents collapse, reducing expected welfare loss); term (ii) increases in β (robustness requires more samples). The optimal risk parameter β^* minimizes Eq. (14) over β and is derived in closed form in Corollary 11:

$$\beta^* = \frac{1}{1+\varepsilon}\log\left(\frac{\mathrm{CW}(G, \varepsilon) \cdot \varepsilon \cdot T}{c_{\mathrm{eff}}}\right) \tag{15}$$

where $c_{\mathrm{eff}} = c|\mathcal{A}_j|$ is estimated from short pilot interactions. Every quantity in β^* appears in the Cooperation Window CW and the PAC bound of Theorem 9; the formula is dimensionally consistent by construction.

5.1 Adaptive $\beta(t)$: Online Risk Calibration

A fixed β is suboptimal: during early training when the partner is erratic, higher β is appropriate; once cooperation stabilizes, lower β accelerates learning. We derive an online update rule from the $\mathrm{PoP}_{\mathrm{DYN}}$ diagnostic.

Definition 4 ($\mathrm{PoP}_{\mathrm{DYN}}$ as Adaptivity Diagnostic). *The dynamic Price of Paranoia at episode t is:*

$$\mathrm{PoP}_{\mathrm{DYN}}(t) = \frac{\mathrm{SW}(\pi_t, \bar{q})}{\mathrm{SW}(\pi_{\mathrm{mm}}, \bar{q})} \in [1,\ \mathrm{PoP}_{\mathrm{GT}}] \tag{16}$$

$\mathrm{PoP}_{\mathrm{DYN}}(t) = 1$: agent is fully paranoid (cooperation lost). $\mathrm{PoP}_{\mathrm{DYN}}(t) = \mathrm{PoP}_{\mathrm{GT}}$: agent is fully cooperative. Its rate of change $\dot{\mathrm{PoP}}_{\mathrm{DYN}}$ *signals whether cooperation is being gained (adaptive progress) or lost (Hangover onset).*

The intuition for adaptive $\beta(t)$ is: if $\dot{\mathrm{PoP}}_{\mathrm{DYN}} < 0$ (cooperation is deteriorating), increase β to become more conservative; if $\dot{\mathrm{PoP}}_{\mathrm{DYN}} > 0$ (recovering), decrease β toward the optimal β^*.

5.2 The Unified Price of Paranoia Framework

We embed RATTL's adaptivity within a formal welfare framework.

Algorithm 1 RATTL with Adaptive $\beta(t)$

Require: Action a, reward r, partner action a_j
Require: Params: α, η, η_β, β^* (estimated), $\underline{\beta}, \bar{\beta}$ with $\underline{\beta} \frac{1}{4} > -1$ and $\bar{\beta} < \varepsilon^{-1} - 1$

// **Online partner model (EMA)**
$\hat{p} \leftarrow (1-\alpha)\hat{p} + \alpha \mathbf{1}[a_j = S]$
$\sigma^2_{\hat{p}} \leftarrow \hat{p}(1-\hat{p})$
// **Estimate** $\mathrm{PoP}_{\mathrm{DYN}}$ **change from reward signal**
$\widehat{\mathrm{SW}} \leftarrow 2\, r_i$ ▷ Welfare proxy (symmetric game)
$\Delta\mathrm{PoP} \leftarrow \widehat{\mathrm{SW}} - \widehat{\mathrm{SW}}_{\mathrm{prev}}$
// **Adaptive** $\beta(t)$ **update**
if $\Delta\mathrm{PoP} < 0$ **then** ▷ Cooperation deteriorating
  $\beta \leftarrow \min(\beta + \eta_\beta\, |\Delta\mathrm{PoP}|\,,\ \bar{\beta})$
else ▷ Recovering toward β^*
  $\beta \leftarrow \beta - \eta_\beta(\beta - \beta^*)_+$
end if
// **Adaptive trust factor**
$\tau \leftarrow (1 + \beta\, \sigma^2_{\hat{p}})^{-1}$
// **Trust-dampened advantage**
$b \leftarrow \mathrm{mean}(\mathcal{H})$; $A \leftarrow (r - b) \cdot \tau^{\mathbf{1}[a=S]}$
$\theta \leftarrow \theta + \eta\, A\, \nabla_\theta \log \pi_\theta(a)$
$\widehat{\mathrm{SW}}_{\mathrm{prev}} \leftarrow \widehat{\mathrm{SW}}$
return θ, β

Definition 5. $\mathrm{PoP}_{\mathrm{GT}}$ ***(Game-Theoretic):*** *Structural welfare ceiling:*

$$\mathrm{PoP}_{\mathrm{GT}}(G, \varepsilon) = \mathrm{SW}(\pi_{\mathrm{rat}}(\varepsilon), \bar{q}_\varepsilon) \,/\, \mathrm{SW}(\pi_{\mathrm{mm}}, \bar{q}_\varepsilon) \tag{17}$$

$\mathrm{PoP}_{\mathrm{ALG}}$ ***(Algorithmic):*** *Sample-complexity overhead:*

$$\mathrm{PoP}_{\mathrm{ALG}}(\mathcal{A}, \beta) = N_{\mathrm{robust}}(\beta) \,/\, N_{\mathrm{standard}} \le O(|\mathcal{A}_j|\, e^{\beta}) \tag{18}$$

$\mathrm{PoP}_{\mathrm{DYN}}(t)$ ***(Dynamic):*** *Adaptive diagnostic (Def. 4).*
Separation: $\mathrm{PoP}_{\mathrm{GT}}$ *is a property of the* game, $\mathrm{PoP}_{\mathrm{ALG}}$ *of the* algorithm, *and* $\mathrm{PoP}_{\mathrm{DYN}}$ *of the* learning trajectory. *Conflating them yields dimensionally inconsistent formulas; separating them yields the corrected* β^*.

Theorem 5. *Define the* Cooperation Window: $\mathrm{CW}(G, \varepsilon) = W^* - W_{\mathrm{mm}}(\varepsilon) = W^*(1 - 1/\mathrm{PoP}_{\mathrm{GT}}(G, \varepsilon))$. *Any algorithm achieves welfare* $W(\pi) \in [W^*/\mathrm{PoA},\ W^*]$. *Its* cooperation efficiency $\eta(\pi) = (W(\pi) - W_{\mathrm{mm}})/\mathrm{CW} \in [0, 1]$ *measures exploitation of* CW. *The dynamic diagnostic* $\mathrm{PoP}_{\mathrm{DYN}}(t)$ *in Def. 4 is monotone in* $\eta(\pi_t)$, *with* $\mathrm{PoP}_{\mathrm{DYN}}(t) = 1$ *iff* $\eta(\pi_t) = 0$ *(full collapse to maximin) and* $\mathrm{PoP}_{\mathrm{DYN}}(t) = \mathrm{PoP}_{\mathrm{GT}}$ *iff* $\eta(\pi_t) = 1$ *(full cooperative welfare). Its sign of change therefore tracks adaptive progress in real time.*

5.3 Theoretical Guarantees

Now, we state the theoretic guarantees of RATTL with adaptive $\beta(t)$ and provide proof sketches.

5.3.1 Adaptive Basin Expansion.

Theorem 6. *For RATTL with adaptive* $\beta(t)$, *payoff spread* $\Delta = r_c - r_s$, *and risk-neutral threshold* $p^* = (r_h - r_s)/\Delta$, *the effective cooperation threshold at episode t satisfies:*

$$p^*_{\beta(t)} = p^* - \frac{\beta(t)\, p^*(1-p^*)}{\Delta} + O((\beta(t))^2 (p^*)^2/\Delta^2) \tag{19}$$

The cooperation basin tracks $\beta(t)$ *in real time. When* $\beta(t)$ *increases during a Hangover event, the basin expands immediately, providing a self-correcting robustness mechanism. The expansion* $\Delta p^* = \beta(t) p^*(1-p^*)/\Delta$ *is maximised at* $p^* = \frac{1}{2}$, *i.e. when* $r_h = (r_c + r_s)/2$.

Proof Sketch. The trust-dampened cooperation condition is $b + \tau(q)(Q_S(q) - b) \ge r_h$. Taylor-expanding at $q = p^*$ with $Q_S(p^*) = r_h$, using $Q'_S = \Delta$ and $\tau'(q)|_{p^*} = -\beta(t)(1 - 2p^*)(1 + \beta(t)\sigma^2_{\hat{p}})^{-2}$, the first-order perturbation gives $\delta q = -\beta(t) p^*(1-p^*)/\Delta$. Since $\beta(t)$ updates online, $p^*_{\beta(t)}$ tracks all changes in $\beta(t)$ with the same episode lag as the EMA update. □

Corollary 7. *When the adaptive update rule (Algorithm 1) detects* $\Delta\mathrm{PoP} < 0$ *and increases* $\beta(t)$ *by* $\eta_\beta\, |\Delta\mathrm{PoP}|$, *the cooperation basin expands by* $\delta(\Delta p^*) = \eta_\beta\, |\Delta\mathrm{PoP}|\, p^*(1-p^*)/\Delta > 0$. *The algorithm is self-correcting: cooperation deterioration triggers immediate basin expansion.*

Corollary 8. *Under replicator dynamics* $\dot{p} = p(1-p)(Q_S(p) - r_h)$, *RATTL shifts the unstable fixed point from* p^* *to* $p^*_{\beta(t)}$, *widening the basin of attraction of* $p = 1$. *The basin grows in real time as* $\beta(t)$ *increases, until cooperation is re-established and* $\beta(t)$ *relaxes back toward* β^*.

5.3.2 Robustness Guarantees.

Theorem 9 (PAC Sample Complexity). *Under Assumptions 1–2, for any accuracy* $\epsilon_{\mathrm{PAC}} > 0$ *and confidence* $\delta \in (0, 1)$, *RATTL returns an* ϵ_{PAC}*-optimal robust policy with probability at least* $1 - \delta$ *using at most*

$$N_{\mathrm{RATTL}} = \widetilde{O}\left(\frac{|\mathcal{S}|^2\, |\mathcal{A}|\, H^4}{\epsilon^2_{\mathrm{PAC}}} \cdot |\mathcal{A}_j|\, e^{\beta} \cdot \log \tfrac{1}{\delta} \right) \tag{20}$$

episodes (the accuracy parameter ϵ_{PAC} *is distinct from the non-stationarity radius* ε *of Assumption 2). The robustness overhead is* $\mathrm{PoP}_{\mathrm{ALG}} \le O(|\mathcal{A}_j|\, e^{\beta})$*—polynomial in* $|\mathcal{A}_j|$, *not exponential in the full state-action space.*

Theorem 10 (Welfare Loss Decomposition). *Expected per-episode welfare loss decomposes into two adaptive-robustness trade-off terms:*

$$\mathbb{E}[\mathrm{Loss}] \le \underbrace{\mathrm{CW}(G, \varepsilon)(1 - e^{-\beta\varepsilon})}_{\text{(i) adaptation cost}} + \underbrace{c\, |\mathcal{A}_j|\, e^{\beta}/T}_{\text{(ii) robustness overhead}} \tag{21}$$

Term (i) is the welfare lost to insufficient adaptation (decreases in β*); term (ii) is the sample cost of robustness (increases in* β*). The adaptive* $\beta(t)$ *rule minimizes this bound online.*

Corollary 11. *The optimal fixed risk parameter is:*

$$\beta^* = \frac{1}{1 + \varepsilon} \log\left(\frac{\mathrm{CW}(G, \varepsilon) \cdot \varepsilon \cdot T}{c_{\mathrm{eff}}} \right) \tag{22}$$

where $c_{\mathrm{eff}} = c\, |\mathcal{A}_j|$ *calibrated empirically. Under the adaptive* $\beta(t)$ *rule,* $\beta(t) \to \beta^*$ *in expectation as* $T \to \infty$ *under stationary non-stationarity, and* $\beta(t)$ *tracks changes in* ε *and* CW *with lag* $O(1/\eta_\beta)$.

6 EXPERIMENTS

We first demonstrate that our theoretical framework generalizes across coordination games via predicted threshold analysis (§6.1), and then validate RATTL empirically on the Iterated Stag Hunt (§6.2).

Game	(r_c, r_h, r_s)	p^*	$p^*_{\beta=1}$	Expansion
Stag Hunt	$(5, 2, -5)$	0.70	0.58	+17%
Chicken	$(4, 2, -1)$	0.60	0.45	+25%
Pure Coord.	$(3, 1, 0)$	0.33	0.26	+21%

Table 1: Predicted basin expansion across coordination games. p^*: risk-neutral threshold; $p^*_{\beta=1}$: RATTL threshold (Theorem 6). Expansion denotes the percentage increase in the cooperation basin $[p^*, 1]$.

	No noise		Reward noise	
Algorithm	PoP↑	PoA↓	PoP↑	PoA↓
RATTL (β=−1)	1.27	4.23	1.34	3.96
RATTL (β=1)	**2.88**	**1.85**	**2.93**	**1.82**
PPO	2.51	2.13	2.49	2.14

Table 2: PoP and PoA in the Iterated Stag Hunt, with and without reward noise. Risk-seeking RATTL ($\beta = 1.0$) dominates in both conditions.

6.1 Theoretical Generalisation Across Games

While our empirical evaluation focuses on the Stag Hunt, Theorem 6 applies to any symmetric coordination game satisfying Definition 1. Table 1 instantiates the basin-expansion prediction $p^*_\beta = p^* - \beta p^*(1-p^*)/\Delta$ for three canonical games at $\beta = 1.0$.

The Chicken game (*Hawk-Dove*) has a smaller payoff spread $\Delta = 5$ vs. $\Delta = 10$ for the Stag Hunt, making the basin more sensitive to trust-factor dampening. The pure coordination game ($r_s = 0$) has the lowest threshold and hence the widest initial basin; RATTL still provides a meaningful expansion. These predictions are directly testable: each row requires only the payoff triple and Theorem 6, with no game-specific tuning.

Baseline positioning. Our empirical evaluation uses vanilla PPO as the risk-neutral baseline. Hysteretic Q-learning [26] and lenient learning [31] filter negative updates to aid cooperation *discovery*, but they offer no principled mechanism for cooperation *retention* under sustained partner noise—the regime RATTL targets. LOLA [17] addresses co-learning noise via second-order gradient corrections but requires differentiable access to the partner's learning rule (see §2). RATTL's $O(1)$ scalar trust factor achieves retention guarantees (Theorem 6) under strictly weaker information assumptions.

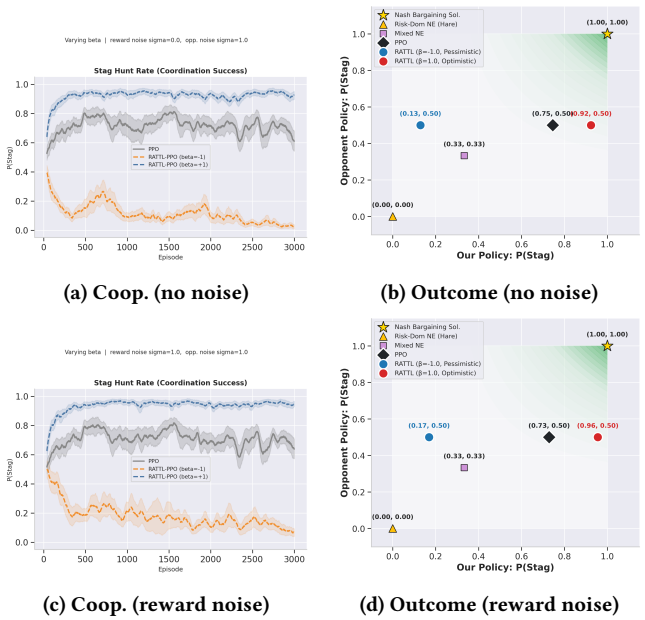
6.2 Iterated Stag Hunt

We train RATTL-PPO in the Iterated Stag Hunt against a stochastic opponent whose mixed strategy is sampled from a standard normal distribution and projected onto the simplex at each timestep. Training runs for 3000 episodes under two configurations: $\beta = -1.0$ (risk-averse, amplifies gradient signals) and $\beta = 1.0$ (risk-seeking, dampens noisy gradients via the trust factor). We compare against vanilla PPO [38]. Note that in the trust factor $\tau = (1 + \beta\sigma^2_{\hat{p}})^{-1}$, positive β dampens the Stag gradient (filtering partner noise), while negative β amplifies it.

As shown in Figure 1a, risk-seeking RATTL-PPO ($\beta = 1.0$) establishes and maintains stable cooperation, while the risk-averse variant ($\beta = -1.0$) converges to Hare, and vanilla PPO settles on a volatile mixed strategy (≈63% Stag). Figure 1b projects policies into the outcome space against the Nash Bargaining Solution (NBS)—the unique Pareto-optimal point maximizing surplus gains over the disagreement point [22]. Risk-seeking RATTL reliably isolates the cooperative equilibrium near the NBS. The bottom row of Figure 1 confirms that all configurations are robust to standard normal reward perturbations, maintaining near-identical equilibrium dynamics. Table 2 reports PoP and PoA under both conditions: risk-seeking RATTL ($\beta = 1.0$) dominates both metrics regardless of reward noise. Figure 2 (Appendix) further sweeps five risk profiles across stationary, mild (20%), and high (40%) partner noise: risk-aversion ($\beta \in \{1, 2\}$) collapses sharply under high non-stationarity (final cooperation drops to ∼0.5), while risk-seeking and risk-neutral retain > 0.9. The Pareto frontier (cooperation vs. stability) confirms that risk-seeking dominates the upper-left quadrant under heavy noise—directly evidencing the EVaR Paradox (Proposition 4).

7 CONCLUSION

We address the collapse of cooperation in non-stationary MARL driven by co-learning partner stochasticity. The EVaR Paradox (Proposition 4) shows that standard distributional robustness applied to returns widens the basin of instability; resolving this requires targeting gradient variance instead. This yields RATTL, whose adaptive trust factor $\tau(t)$ provably expands the cooperation basin and is unified under the Price of Paranoia framework.



(a) Coop. (no noise) **(b) Outcome (no noise)**

(c) Coop. (reward noise) **(d) Outcome (reward noise)**

Figure 1: RATTL-PPO in the Iterated Stag Hunt. Top row: no reward noise; bottom row: standard-normal reward perturbations. Risk-seeking RATTL ($\beta = 1.0$) converges to stable cooperation near the NBS in both conditions; risk-averse ($\beta = -1.0$) defaults to Hare; vanilla PPO oscillates.

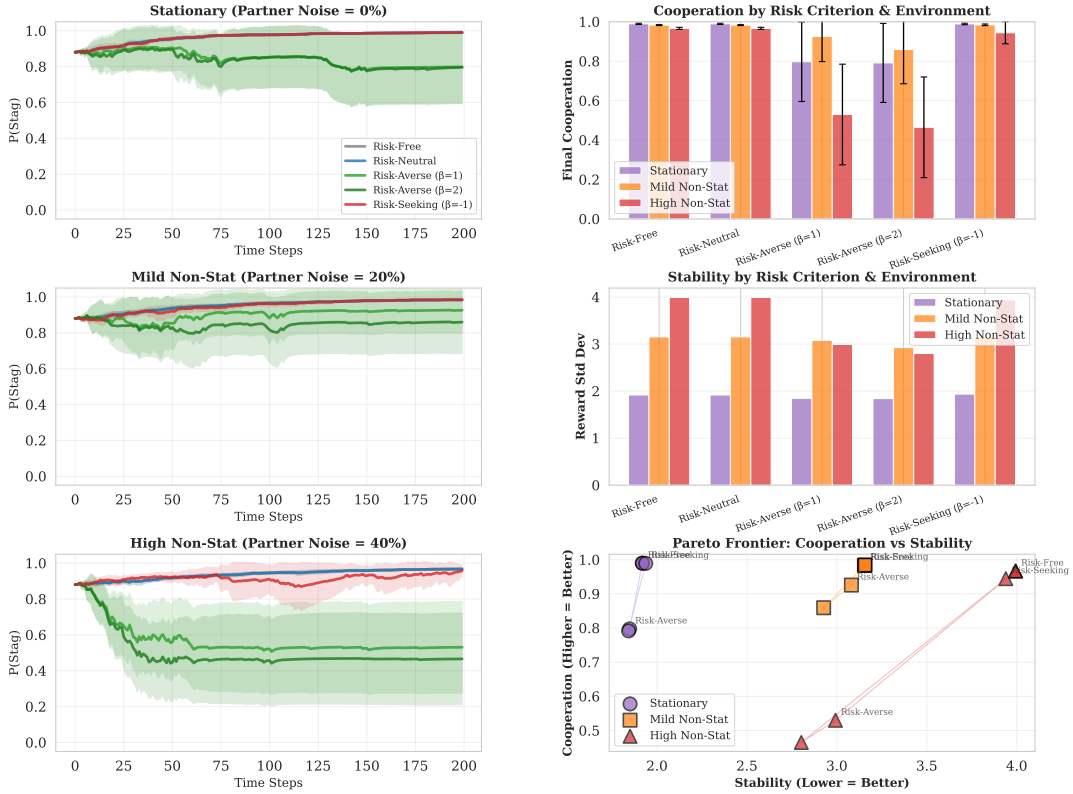


Figure 2: Risk-criteria performance across stationary and non-stationary partner dynamics. Left column: cooperation probability $P(\text{Stag})$ over 200 episodes for five risk profiles (Risk-Free, Risk-Neutral, Risk-Averse $\beta \in \{1, 2\}$, Risk-Seeking $\beta = -1$) under partner noise $\in \{0\%, 20\%, 40\%\}$. Top-right: final cooperation by criterion and environment; risk-aversion collapses sharply under high non-stationarity, while risk-seeking and risk-neutral retain > 0.9. Middle-right: reward standard deviation; non-stationarity raises variance uniformly. Bottom-right: Pareto frontier of cooperation vs. stability; risk-seeking dominates the upper-left (high cooperation, low variance) under heavy noise, whereas risk-aversion drops to ~ 0.5 cooperation. This empirically validates the EVaR Paradox (Proposition 4): return-level risk-aversion is counterproductive for cooperation retention.

Empirically, risk-seeking RATTL achieves near-100% cooperation retention where standard baselines collapse. Our framework demonstrates that stable cooperation requires neither opponent modelling nor prosocial priors—only calibrated uncertainty. Future directions include adaptive $\beta(t)$ meta-learning, scalable deep RL implementations, and human-AI cooperation studies.

REFERENCES

[1] Daron Acemoglu and Alexander Wolitzky. 2015. *Sustaining Cooperation: Community Enforcement vs. Specialized Enforcement.* Working Paper 21457. National Bureau of Economic Research. https://doi.org/10.3386/w21457

[2] Amir Ahmadi-Javid. 2012. Entropic Value-at-Risk: A New Coherent Risk Measure. *Journal of Optimization Theory and Applications* 155, 3 (2012), 1105–1123.

[3] Stefano V. Albrecht and Peter Stone. 2018. Autonomous Agents Modelling Other Agents: A Comprehensive Survey and Open Problems. *Artificial Intelligence* 258 (2018), 66–95.

[4] Luciano Andreozzi, Matteo Ploner, and Ali Seyhun Saral. 2020. The stability of conditional cooperation: beliefs alone cannot explain the decline of cooperation in social dilemmas. *Scientific reports* 10, 1 (2020), 13610.

[5] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. 1999. Coherent measures of risk. *Mathematical finance* 9, 3 (1999), 203–228.

[6] Baruch Awerbuch, Yossi Azar, Adi Richman, and Bart Tsur. 2005. Tradeoffs in Worst-Case Equilibria. In *Proceedings of APPROX/RANDOM.*

[7] Avrim Blum and Yishay Mansour. 2008. From External to Internal Regret. *Journal of Machine Learning Research* 8 (2008), 1307–1324.

[8] Michael Bowling and Manuela Veloso. 2002. Multiagent Learning Using a Variable Learning Rate. *Artificial Intelligence* 136, 2 (2002), 215–250.

[9] Robert Boyd and Peter J Richerson. 2009. Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1533 (2009), 3281–3288.

[10] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. 2017. Risk-Constrained Reinforcement Learning with Percentile Risk Criteria. *Journal of Machine Learning Research* 18, 1 (2017), 167–208.

[11] Caroline Claus and Craig Boutilier. 1998. The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems. In *Proceedings of the AAAI Conference on Artificial Intelligence.* 746–752.

[12] Pedro Dal Bó and Guillaume R. Fréchette. 2011. The Evolution of Cooperation in Infinitely Repeated Games: Experimental Evidence. *American Economic Review* 101, 1 (February 2011), 411–29. https://doi.org/10.1257/aer.101.1.411

[13] Ernst Fehr and Simon Gächter. 2000. Cooperation and Punishment in Public Goods Experiments. *American Economic Review* 90, 4 (2000), 980–994.

[14] Ernst Fehr and Herbert Gintis. 2007. Human motivation and social cooperation: Experimental and analytical foundations. *Annu. Rev. Sociol.* 33, 1 (2007), 43–64.

[15] Ernst Fehr and Simon Gächter. 2000. Cooperation and Punishment in Public Goods Experiments. *American Economic Review* 90, 4 (September 2000), 980–994. https://doi.org/10.1257/aer.90.4.980

[16] Tawni Hunt Ferrarini. 2013. The Economics of Government and the Fall of Rome. *Social Education* 77, 2 (2013), 60–63.

[17] Jakob Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, and Pieter Abbeel. 2018. Learning with Opponent-Learning Awareness. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. 122–130.

[18] Piotr J. Gmytrasiewicz and Prashant Doshi. 2005. A Framework for Sequential Planning in Multi-Agent Settings. *Journal of Artificial Intelligence Research* 24 (2005), 49–79.

[19] Ido Greenberg, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. 2022. Efficient Risk-Averse Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[20] Michele Griessmair and Patrick Hippmann. 2022. Anger, guilt, and repeated cooperation in social dilemmas. *Emotion* 22, 3 (2022), 444.

[21] Carla Handley and Sarah Mathew. 2020. Human large-scale cooperation as a product of competition between cultural groups. *Nature communications* 11, 1 (2020), 702.

[22] John C. Harsanyi and Reinhard Selten. 1988. *A General Theory of Equilibrium Selection in Games*. MIT Press.

[23] Edward Hughes, Joel Z. Leibo, Matthew G. Phillips, Karl Tuyls, Edgar A. Duéñez-Guzmán, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin R. McKee, Raphael Koster, Heather Roff, and Thore Graepel. 2018. Inequity Aversion Improves Cooperation in Intertemporal Social Dilemmas. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[24] Garud N. Iyengar. 2005. Robust Dynamic Programming. *Mathematics of Operations Research* 30, 2 (2005), 257–280.

[25] Miao Lu, Han Zhong, Tong Zhang, and Jose Blanchet. 2024. Distributionally Robust Reinforcement Learning with Interactive Data Collection: Fundamental Hardness and Near-Optimal Algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[26] Laëtitia Matignon, Guillaume J. Laurent, and Nadège Le Fort-Piat. 2007. Hysteretic Q-Learning: An Algorithm for Decentralised Reinforcement Learning in Cooperative Multi-Agent Teams. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 64–69.

[27] Kevin R. McKee, Ian Gemp, Brian McWilliams, Edgar A. Duéñez-Guzmán, Edward Hughes, and Joel Z. Leibo. 2020. Social Diversity and Social Preferences in Mixed-Motive Reinforcement Learning. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

[28] Arnab Nilim and Laurent El Ghaoui. 2005. Robust Control of Markov Decision Processes with Uncertain Transition Matrices. *Operations Research* 53, 5 (2005), 780–798.

[29] Shayegan Omidshafiei, Dong-Ki Kim, Miao Liu, Gerald Tesauro, Matthew Riemer, Christopher Amato, Murray Campbell, and Jonathan P. How. 2019. Learning to Teach in Cooperative Multiagent Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[30] Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P. How, and John Vian. 2017. Deep Decentralised Multi-Task Multi-Agent Reinforcement Learning under Partial Observability. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[31] Gregory Palmer, Karl Tuyls, Daan Bloembergen, and Rahul Savani. 2018. Lenient Multi-Agent Deep Reinforcement Learning. *arXiv preprint arXiv:1707.04402* (2018).

[32] Kishan Panaganti and Dileep Kalathil. 2022. Sample Complexity of Robust Reinforcement Learning with a Generative Model. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*. 9582–9602.

[33] Christos H. Papadimitriou and Tim Roughgarden. 2005. Computing Correlated Equilibria in Multi-Player Games. In *Proceedings of the ACM Symposium on Theory of Computing (STOC)*.

[34] Alexander Peysakhovich and Adam Lerer. 2017. Prosocial Learning Agents Solve Generalised Stag Hunts Better than Selfish Ones. *arXiv preprint arXiv:1709.02865* (2017).

[35] Roberta Raileanu, Emily Denton, Arthur Szlam, and Rob Fergus. 2018. Modeling Others Using Oneself in Multi-Agent Reinforcement Learning. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[36] R. Tyrrell Rockafellar and Stanislav Uryasev. 2000. Optimization of Conditional Value-at-Risk. *Journal of Risk* 2, 3 (2000), 21–41.

[37] Tim Roughgarden. 2015. Intrinsic Robustness of the Price of Anarchy. *J. ACM* 62, 5 (2015), 32:1–32:42.

[38] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[39] Elham Semsar-Kazerooni and Khashayar Khorasani. 2009. Multi-agent team cooperation: A game theory approach. *Automatica* 45, 10 (2009), 2205–2213.

[40] Laixi Shi, Eric Mazumdar, Yuejie Chi, and Adam Wierman. 2024. Sample-Efficient Robust Multi-Agent Reinforcement Learning in the Face of Environmental Uncertainty. *arXiv preprint arXiv:2404.18909* (2024).

[41] Brian Skyrms. 2004. *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press.

[42] Aviv Tamar, Yonatan Glassner, and Shie Mannor. 2015. Optimizing the CVaR via Sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[43] Woodrow Z Wang, Mark Beliaev, Erdem Biyik, Daniel A Lazar, Ramtin Pedarsani, and Dorsa Sadigh. 2021. Emergent Prosociality in Multi-Agent Games Through Gifting. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*.

[44] Lucas Wardil, Ivair R Silva, and Jafferson KL da Silva. 2019. Positive interactions may decrease cooperation in social dilemma experiments. *Scientific Reports* 9, 1 (2019), 1017.

[45] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3 (1992), 229–256.

[46] Zhongwen Xu, Hado van Hasselt, and David Silver. 2018. Meta-Gradient Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[47] Kaiqing Zhang, Tao Sun, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Başar. 2020. Robust Multi-Agent Reinforcement Learning with Model Uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[48] Runyu Zhang, Na Li, Asuman Ozdaglar, Jeff Shamma, and Gioele Zardini. 2025. Optimism as Risk-Seeking in Multi-Agent Reinforcement Learning. *arXiv preprint arXiv:2509.24047* (2025). Accepted to IEEE L-CSS / ACC 2026.

A APPENDIX: EXTENDED EXPERIMENTAL RESULTS

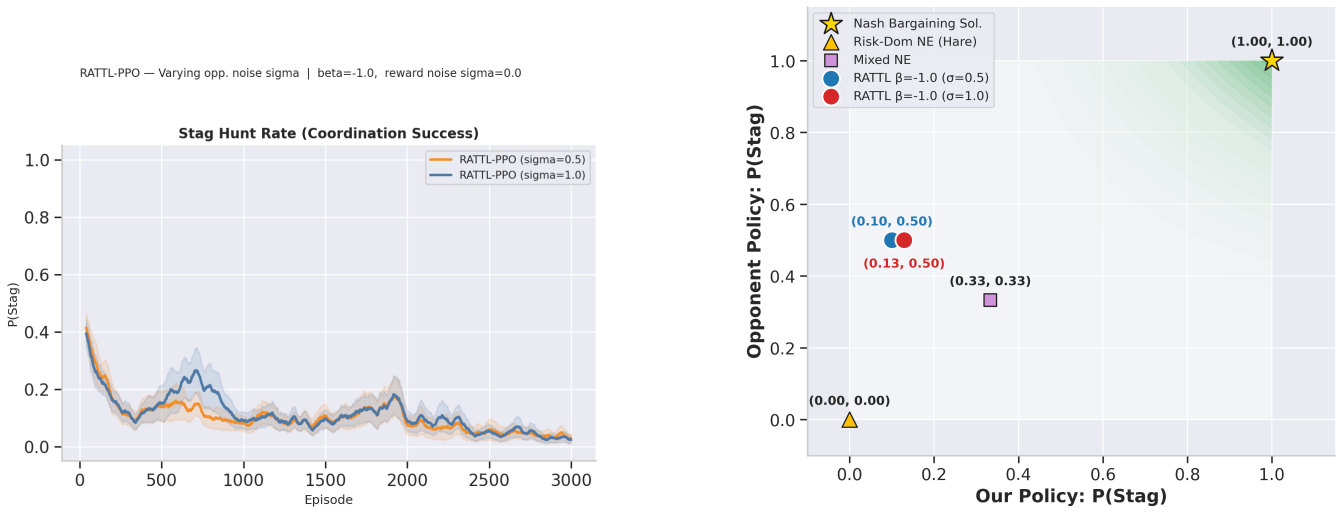
We next inject standard normal noise ($\sigma \in \{0.5, 1.0\}$) directly into the opponent's mixed strategy, testing resilience to behavioral stochasticity beyond reward perturbations. All configurations show resilience to partner noise, though the nature of that resilience differs sharply across objectives.

Risk-seeking RATTL.. As shown in Figure 4, RATTL-PPO ($\beta = 1.0$) maintains cooperation (Stag rate > 92%) despite heavy partner noise (Table 4: PoP peaks at 2.88, PoA drops to 1.82 at $\sigma = 1.0$). The trust factor absorbs partner stochasticity, keeping the policy anchored near the NBS.

Risk-averse RATTL and PPO.. Risk-averse RATTL ($\beta = -1.0$, Figure 3) collapses to Hare (Stag rate ≈0.10) across all noise levels (Table 3), confirming that gradient amplification prevents trust formation under partner unpredictability. Vanilla PPO (Figure 5, Table 5) converges to a volatile mixed strategy (≈75% Stag) but fails to commit to full cooperation.

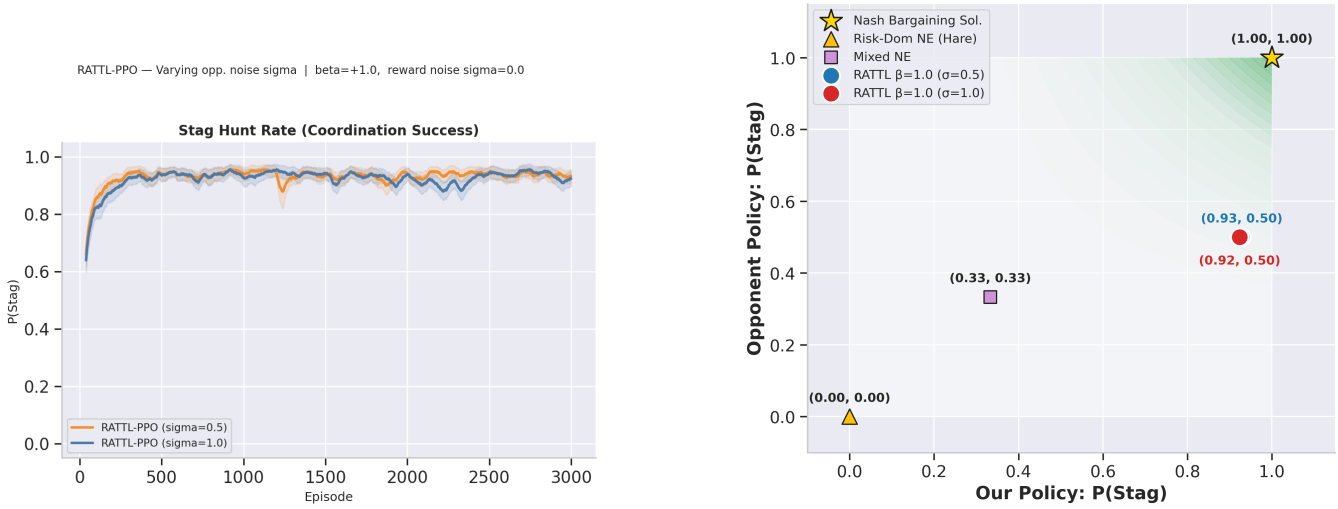
σ	PoP (↑)	PoA (↓)
0.5	**1.16**	**4.59**
1.0	1.27	4.23

Table 3: Empirical evaluation of the Price of Paranoia (PoP) and Price of Anarchy (PoA) across noise added to the stochastic opponent's mixed strategy in the Iterated Stag Hunt game for RATTL-PPO ($\beta = -1.0$).



(a) Attempt at coordination **(b) Mixed Strategy Nash Equilibria**

Figure 3: Evaluation of RATTL-PPO ($\beta = -1.0$) against a stochastic opponent in the Iterated Stag Hunt game where the partner strategy is perturbed by noise from a standard normal distribution.



(a) Attempt at coordination **(b) Mixed Strategy Nash Equilibria**

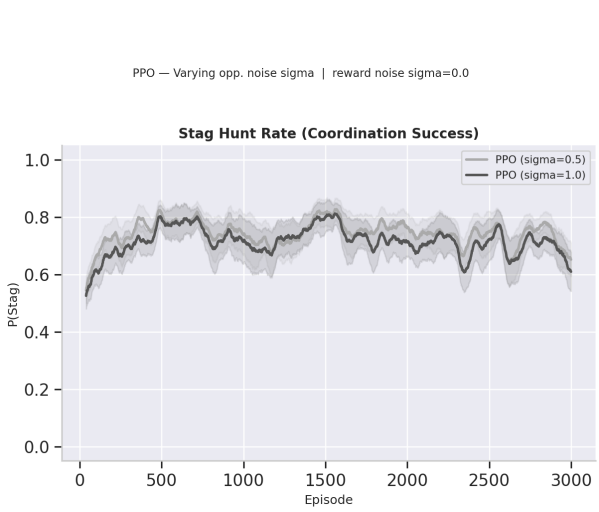
Figure 4: Evaluation of RATTL-PPO ($\beta = 1.0$) against a stochastic opponent in the Iterated Stag Hunt game where the partner strategy is perturbed by noise from a standard normal distribution.

σ	PoP (↑)	PoA (↓)
0.5	2.86	1.86
1.0	**2.88**	**1.85**

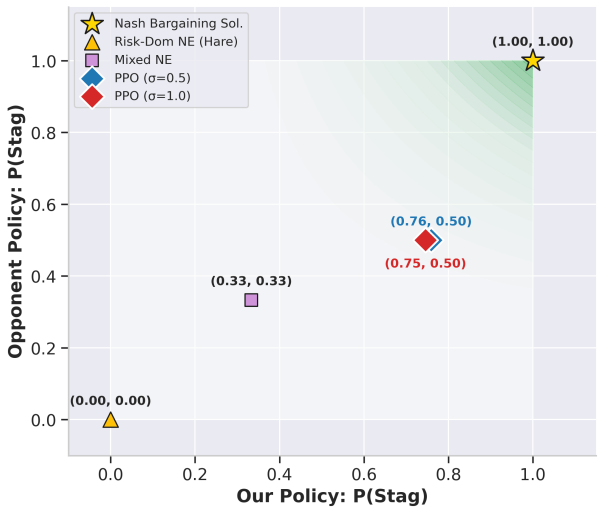
Table 4: Empirical evaluation of the Price of Paranoia (PoP) and Price of Anarchy (PoA) across noise added to the stochastic opponent's mixed strategy in the Iterated Stag Hunt game for RATTL-PPO ($\beta = 1.0$).

σ	PoP (↑)	PoA (↓)
0.5	2.53	2.11
1.0	**2.51**	**2.13**

Table 5: Empirical evaluation of the Price of Paranoia (PoP) and Price of Anarchy (PoA) across noise added to the stochastic opponent's mixed strategy in the Iterated Stag Hunt game for PPO.



(a) Attempt at coordination



(b) Mixed Strategy Nash Equilibria

Figure 5: Evaluation of PPO against a stochastic opponent in the Iterated Stag Hunt game where the partner strategy is perturbed by noise from a standard normal distribution.