ROBUST LARGE-SCALE MULTI-AGENT REINFORCEMENT LEARNING AGAINST BYZANTINE FAULTS

Anonymous authors
Paper under double-blind review

ABSTRACT

In a large-scale multi-agent system, several corrupted agents with unknown identities can derail the system by spreading abnormal behaviors that harm coordination. Such failures are known as *Byzantine faults*, under which defenders know neither the number nor the identities of agents being attacked. The classic solution concept is a Bayesian equilibrium, where defenders act on beliefs about whether each agent is cooperative or adversarial. However, this inference becomes intractable as the agent population grows, since the number of possible corrupted subsets increases exponentially. Our key insight is that individual defenders do not need to identify the type of the whole team, but only to aggregate their type uncertainty into the mean field, capturing how the population is distorted rather than by whom. Here we propose Belief over Mean Field (BeMF), which realizes this as robust mean-field control over two agent types. BeMF assigns each agent a Byzantine risk score in linear time, uses these scores to build a belief-weighted mean field, and learns a shared policy that acts on it. We prove that BeMF's belief-weighted mean-field model approximates both population behavior and worst-case performance with explicit error bounds that shrink as the population grows, so the attacker–defender equilibrium remains approximately valid at finite population sizes. We evaluate BeMF on red-versus-blue team combat with up to 144 agents and power-grid control with 20 to 40 agents. In both domains, BeMF outperforms representative baselines even when half of the agents are corrupted. These results establish BeMF as a scalable approach to robust multi-agent cooperation.

1 INTRODUCTION

Large-scale cooperative multi-agent reinforcement learning (MARL) is increasingly used for many agent control in domains such as robot swarm control (Zheng et al., 2018), voltage control (Wang et al., 2021b), and traffic control (Nguyen et al., 2018). Once deployed, however, a few agents may stop behaving as they were trained to, whether through action perturbations caused by software or hardware faults (Wan et al., 2024; Zhang et al., 2026) or through adversaries that hijack them and execute worst-case actions (Tessler et al., 2019; Ly & Ly, 2021). Such corrupted agents do not merely fail on their own: they keep taking part in the coordination, and as their teammates adapt to their behavior, the damage spreads well beyond where it started. As a result, attacks on a small group of agents can cause system-wide failure (Li et al., 2026b). Since these agents cannot be told apart from the rest in advance, robustness to such failures is a precondition for real-world deployment.

On the defense side, this is the setting of *Byzantine faults*, where defenders know neither how many teammates are corrupted nor which. A principled solution is a Bayesian equilibrium (Li et al., 2024), in which each defender acts optimally against worst-case attackers under its belief about which teammates are corrupted. However, this approach does not scale. With N agents, an exact belief ranges over 2^N combinations of corrupted agents, and even one belief per teammate totals $O(N^2)$. Our key insight is that, a defender never needs to enumerate every possible set of corrupted agents under mean-field interactions, only to aggregate their type uncertainty into the mean field, capturing how the population is distorted rather than by whom. In mean-field models (Huang et al., 2006; Lasry & Lions, 2007; Yang et al., 2018; Gu et al., 2021), agents are anonymous and interact with the population only through its empirical distribution, the mean field. Corrupted agents therefore affect the mean field only through their number and the states they occupy, not their identities. The same principle underlies Byzantine-robust distributed learning (Yin et al., 2018) and robust statistics

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Diakonikolas & Kane, 2019), whose guarantees depend only on the fraction of faulty workers or outliers. Misjudging individual agents thus matters only in aggregate. Weighting each agent in the mean field by how likely it is to be benign yields a field that departs from the uncorrupted one by an amount controlled by the average error of these weights.

Here we propose *Belief over Mean Field* (BeMF), which builds this insight into robust mean-field control with two agent types, cooperative defenders and Byzantine attackers. Unlike the exogenous types in multi-type mean-field MARL (Subramanian et al., 2020a), these types are latent and adversarially assigned. BeMF infers them with a shared detector that scores each agent's Byzantine risk from its own trajectory, using the estimated attack fraction as a prior. These scores define a belief-weighted mean field, which a shared defender policy combines with the raw mean field and the agent's local state and risk score. The policy is trained against a worst-case attacker by alternating updates toward a robust mean-field Bayesian equilibrium, while the detector is supervised with training-time type labels. Inference is therefore linear in N with network sizes independent of N, and deployment requires neither attacker identities nor the attack fraction.

Our analysis quantifies how imperfect type inference affects this control framework. We first bound the discrepancy between the belief-weighted field and the true cooperative field by the population-average type-score error. We then propagate this error through the closed-loop dynamics to bound the finite-horizon return gap relative to the same controller supplied with true-type inputs. The type-score error is itself controlled by the distinguishability of cooperative and Byzantine behavior and the detector's excess risk. Finally, for a fixed selection mechanism and under uniform regularity conditions, we prove that the belief-augmented population dynamics and robust values converge to their mean-field limits with explicit finite-population error bounds. These bounds transfer approximate mean-field saddle points to finite populations, with an additional error that vanishes as the population grows. Experiments on Battle and PowerGrid support these guarantees empirically. BeMF achieves the highest mean return in 16 of 18 evaluated attack configurations, including all five configurations with half of the agents compromised, and sustains its advantage in Battle teams of up to 144 agents. Ablations further show that belief information improves robustness beyond adversarial training alone. Shared inference and population aggregation yield $O(N)$ inference cost per step, with network dimensions independent of population size.

Contributions. Our contributions are threefold. Theoretically, we formulate Byzantine-robust mean-field control with latent agent types and establish belief-dependent performance bounds, mean-field approximation guarantees, and the transfer of approximate saddles to finite populations. Technically, BeMF combines shared agent-wise type inference, belief-weighted population aggregation, and adversarial policy training, enabling linear-time inference with population-size-independent trainable interfaces and requiring neither attacker identities nor the attack fraction at deployment. Empirically, we evaluate on a large-population discrete benchmark and a continuous safety-critical grid task, under a published attack policy and against both non-robust and robust mean-field baselines.

2 Related Work

Mean-Field MARL. Mean-field approximation offers a principled method of controlling agents at scale. Mean-field game approximate non-cooperative games representative agent with a self-consistent population distribution (Huang et al., 2006; Lasry & Lions, 2007), with recent RL implementations (Guo et al., 2019; Laurière et al., 2022). Mean-field control is a cooperative variant of mean field games, with recent RL solutions on Q learning and error bounds (Angiuli et al., 2022; Gu et al., 2021; Mondal et al., 2022). Beyond theoretical approach, mean-field approximation are applied to large-scale MARL which accounts for neighborhood action averages, local estimates and decentralized deployment (Yang et al., 2018; Subramanian & Mahajan, 2019; Li et al., 2021; Subramanian et al., 2020b; 2022). A similar work to our setting is multi-type mean-field MARL (Subramanian et al., 2020a), where agents have different hard-assigned types and make decisions conditioned on information of its own type and all other types. Later works extend the multi-type setup to imitation learning (Yang et al., 2020), and a major agent that is not aggregated in the mean field (Cui et al., 2024). Comparing with multi-type setup, our types are semantically grounded and latent in a strictly adversarial setting, requiring distinct theoretical and technical solutions.

Robustness in small-scale MARL. According to (Yang & Wang, 2020), we define small-scale MARL as agent number $N < 30$, where modeling the impact on each agent is tractable. Existing

works study robustness in different modalities via max-min methods, including observation (Lin et al., 2020), action (Li et al., 2019), environment dynamics (Zhang et al., 2020) and communications Sun et al. (2022), with uncertainties applied at different agent proportions(Li et al., 2026a). Our work belongs to attack in action modality, with papers showing a minor group of attackers can induce team-wide failure (Li et al., 2023). Defenses either enumerate over all threat scenarios via adversarial training (Li et al., 2019; Phan et al., 2020; 2021), or derive robust regularization as light-weight approach (Bukharin et al., 2023; Li et al., 2025). The paper closest to our setting is EIR-MAPPO, which show the optimal defense is a Bayesian equilibrium, where each agent identify if other agents are attacker and act optimally (Li et al., 2024). However, the method assumes only one attacker, and do not scale to large number of agents.

Robust Mean-Field Decision Making. We consider robustness in mean-field games and mean-field control. On the theoretical side, the first line of work perturbs each agent's own dynamics. Earlier works assume linear quadratic settings for closed-form solution via indefinite Riccati equations (Bauso et al., 2016; Moon & Başar, 2017; Huang & Huang, 2017). Later works relaxes linearity assumption (Tirumalai & Baras, 2022; Delarue & Lavigne, 2026b), and studies cooperative instead of Nash setting (Liang et al., 2022; Wang et al., 2021a). The second line consider global uncertainties shared across all agents, in both mean-field games (Wang, 2026; Delarue & Lavigne, 2026a; Aydin & Saldi, 2025) and mean-field control (Huang et al., 2021; Laurière et al., 2025). Our research belongs to a practical line on large-scale RL with mean-field approximations, where robustness is studied against uncertainties in observations (Zhou & Liu, 2022), environment transitions (Zaman et al., 2024), or as a regularization during training (Bukharin et al., 2023). Closest to ours is VAI (Li et al., 2026b) that finds the most vulnerable subset of agents to attack, which our defense must withstand. To our knowledge, no prior work addresses Byzantine robustness in mean-field control, where an unknown subset of agents may deviate arbitrarily and must be identified and countered at scale.

3 Problem Formulation

3.1 Byzantine mean-field game

We formulate robust large-scale MARL against Byzantine errors as a Byzantine mean-field game with an unknown portion and identity of adversaries, defined by a tuple:

$$\mathcal{G}_N = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, \Omega, P, \Theta, r, \gamma, \beta \rangle,$$

where $\mathcal{N} = \{1, \dots, N\}$ is the agent set, $\mathcal{S}$, $\mathcal{A}$ are the state space and action spaces, $P(\mathrm{d}s' \mid s, a, \mu, \nu)$ is a stochastic kernel given $\mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S}) \times \Delta(\mathcal{A}) \to \Delta(\mathcal{S})$, with $\Delta(\cdot)$ the probability measure. $r(s, a, \mu, \nu)$ is the cooperative team reward, and $\gamma \in (0, 1)$ is the discount factor. At the beginning of an episode, each agent is assigned a type θ_i from type space $\Theta_i = \{0, 1\}$, with $\theta_i = 0$ denotes a cooperative agent and $\theta_i = 1$ denotes a Byzantine adversary. The joint type space is $\Theta = \Pi_{i \in N} \Theta^i$. The *attack fraction* $\beta = \sum_{i=1}^{N} \theta^i / N$ specify the adversary proportion in agents. We assume β belongs to a compact set $\beta \in [0, \bar{\beta}]$ for theoretical convenience, with $\bar{\beta} < 1$ an unknown constant.

When game begins, the attacker selects β fraction of agents as adversary via a selection mechanism $\alpha(\theta \mid s_0^1, \dots, s_0^N)$, where s_0^i are initial states of agent i. At time t, each agent receives a local state s_t^i and an empirical mean-field states estimate $\mu_t := \frac{1}{N} \sum_{i=1}^{N} \delta_{s_t^i}$. For cooperative agents, it make decisions via $\pi(a_t^i | s_t^i, p_t^i, \mu_t, \hat{\mu}_t)$, where $p_t^i(\theta)$ is the inferred probability of the type of itself, $\hat{\mu}_t$ is the averaged mean field over cooperative agents. We will discuss how to obtain these in method section. For Byzantine attackers, it make decisions via $\pi_\alpha(a_t^i \mid s_t^i, \theta^i, \mu_t^N, \mu_t^{\mathrm{adv}})$, with $\theta^i = 1$ the type of itself and $\mu_t^{\mathrm{adv}} = \frac{1}{\lfloor \beta N \rfloor} \sum_{i=1}^{N} \theta^i \delta_{s_t^i}$ the mean-field state of adversaries. The joint actions taken by cooperative and Byzantine agents results in an empirical mean-field action $\nu_t := \frac{1}{N} \sum_{i=1}^{N} \delta_{a_t^i}$. The environment yields a shared cooperative reward $r \equiv \frac{1}{N} \sum_{i=1}^{N} r^i(s_t^i, a_t^i, \mu_t^N, \nu_t^N)$ for all agents. The expected return is:

$$J_{N,T}(\pi, \alpha, \pi_\alpha) := \mathbb{E}_{\pi, \alpha, \pi_\alpha} \left[\sum_{t=0}^{T} \gamma^t \frac{1}{N} \sum_{i=1}^{N} r^i(s_t^i, a_t^i, \mu_t^N, \nu_t^N) \right]. \tag{1}$$

Relation to existing formulations. In small-scale robust MARL, our work is closest to EIR-MAPPO (Li et al., 2024), which requires identifying the adversary and adapting defender policies for optimal

defense. However, they assumes a single attacker and does not scale to large agent populations. In large-scale MARL, multi-type mean-field MARL (Subramanian et al., 2020a) captures behavioral diversity by assigning agents known types, or hard-inferring them via clustering. In contrast, our types carry latent semantic meaning, which must be inferred consistently during training. Moreover, they assume a coordinated structure in which each agent pursues its own non-conflicting goal, whereas we take a fully adversarial setting. Also related is VAI (Li et al., 2026b) in large-scale robust MARL, which selects the most vulnerable subset of agents to attack. The paper exposes the threat of Byzantine agents in large-scale systems and thus specifies a strong threat model our method must withstand.

3.2 Threat Model

We next provide the threat model for both attackers and defenders. We consider an adversary may corrupt a fraction β of the agents in the system at the beginning of an episode. The compromised agents remain fixed throughout the episode, while their identities and the attack fraction β are unknown to the defenders. We specify the capabilities and information available to both sides below.

Assumption 3.1 (Attacker's capability)**.** At the beginning of the episode, attacker can assign arbitrary agents as attacker via a selection mechanism $\alpha(\theta \mid s_0^1, \ldots, s_0^N)$, as long as $\sum_{i=1}^N \theta^i \leq \lfloor \beta N \rfloor$. The type θ^i are fixed afterwards and attacker knows its own type. In the episode, the attacker can make arbitrary decisions via its Byzantine action policy $\pi_\alpha(a_t^i \mid s_t^i, \theta^i, \mu_t^N, \mu_t^{\text{adv}})$. We assume a zero-sum setting such that adversary minimize the expected payoff of cooperative agents over both levels, $\min_\alpha \min_{\pi_\alpha} J_{N,T}(\pi, \alpha, \pi_\alpha)$.

We assume a strong adversary that can affect the large-scale system by participating in the decision process, similar to broken or compromised nodes. However, its capability is restricted to changing the action it takes, and cannot directly perturb the environment or the input of other agents. We next specify the capabilities of the defenders.

Assumption 3.2 (Defender's capability)**.** The defender controls its own policy $\pi(a_t^i | s_t^i, p_t^i, \mu_t, \hat{\mu}_t)$. It does not has access to adversarial proportion β, the identity θ^i of each agent being attacked, the selection mechanism α and the Byzantine action policy π_α. Each node is associated with an external detector that estimates the reliability p^i of the node from its trajectories, which is not controlled by adversaries. The detector is not assumed to be perfect. A mean-field aggregates state information conditioned on the reliability reported by each agent.

The detect-then-aggregate architecture is standard in large-scale distributed systems. In federated learning, the gradient of each worker is summarized and aggregated only those that survives trimming (Yin et al., 2018). In trusted cloud platform, the node of each agent is continuously attested and those failed are excluded from the shared state (Schear et al., 2016). In our setting, the aggregated mean field is itself an input to the policy, so the trusted agents must additionally achieve a robust equilibrium in the presence of adversaries, rather than merely computing a single aggregated statistic.

3.3 Solution Concept

For a fixed N, let Π be the admissible class of shared belief-conditioned protected policies, $\mathfrak{A}_{\text{sel}}$ the class of selection methods, and $\mathfrak{A}_{\text{adv}}$ the class of Byzantine action policies. The belief used by policies in Π is Bayes-consistent with the prior and the observations available to the defender. The learned score p_t^i introduced later approximates this ideal Bayesian belief. The robust objective is

$$V_{N,T}^{\text{wc}}(\Pi) := \sup_{\pi \in \Pi} \inf_{\alpha \in \mathfrak{A}_{\text{sel}}} \inf_{\pi_\alpha \in \mathfrak{A}_{\text{adv}}} J_{N,T}(\pi, \alpha, \pi_\alpha). \tag{2}$$

Definition 3.1 (Mean-field robust Bayesian equilibrium)**.** *A strategy profile $(\pi^\star, \alpha^\star, \pi_\alpha^\star)$ is a mean-field robust Bayesian equilibrium if, for every admissible π, α, and π_α,*

$$J_{N,T}(\pi, \alpha^\star, \pi_\alpha^\star) \leq J_{N,T}(\pi^\star, \alpha^\star, \pi_\alpha^\star) \leq J_{N,T}(\pi^\star, \alpha, \pi_\alpha). \tag{3}$$

Moreover, after every on-path history, with the realized selection fixed, $\pi^\star$ and $\pi_\alpha^\star$ remain mutual best responses in the continuation game conditional on the current Bayesian belief.

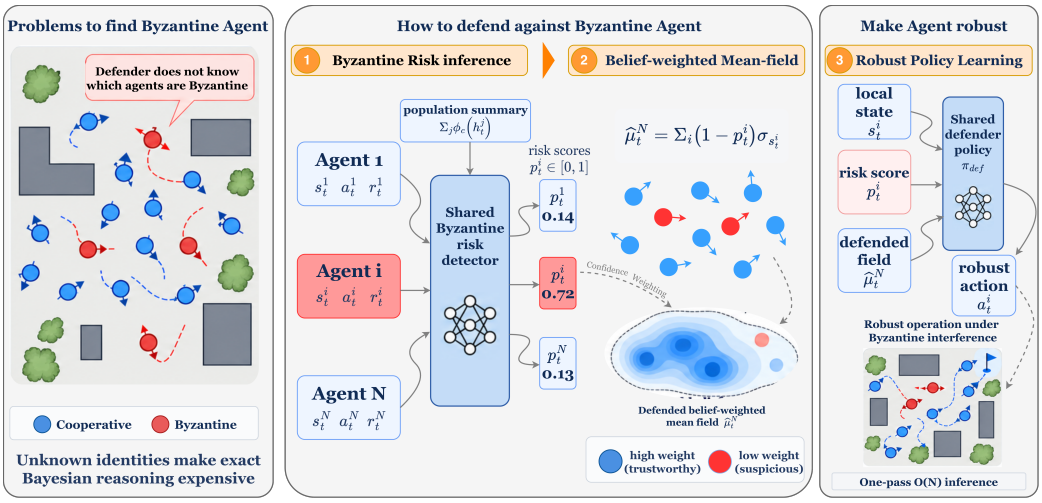


Figure 1: **Method overview**. Each agent has an external monitor that assign each agent a Byzantine risk, then aggregate belief-weighted mean-fields by merging states from trustworthy agents only. Adversarial training with adversaries additionally conditions on risk and belief-weighted mean-field.

4 METHOD

In this section, we introduce Belief over Mean Field (BeMF) as a scalable and practical defense against Byzantine adversaries. As shown in Fig. 1, the overall pipeline is illustrated in three steps: (1) inferring Bayesian type belief of each agent, (2) aggregating clean mean field for defenders and (3) make decisions conditioned on agent-wise risk and clean mean field.

4.1 BAYESIAN TYPE BELIEF

As the identity of each agent is unknown to defenders, optimal defense admits a Bayesian equilibrium, which requires each agent inferring the type of other agents. This creates a $O(N^2)$ complexity with N agent numbers, which is intractable with large N. In this paper, we show that it is possible to attach an external monitor to each agent, which monitor its trustworthiness and report this local result to the mean field. Formally, for agent-wise interaction history $h_t^i = \{s_1^i, a_1^i, ..., s_t^i\}$. Define likelihood of h_t^i under mean-field state and actions $m_{0:t} = \{\mu_s, \nu_s\}_{s=0}^t$ as $L_{\theta,t}(h; m_{0:t}) := \Pr\left(h_t^i = h \,\middle|\, \theta_i = \theta,\, m_{0:t}\right)$ under type $\theta \in \{0, 1\}$. The empirical-Bayes belief of population-wise attack fraction $\hat{\beta}_t^N = p(\beta|h_t^i; m_{0:t})$ and agent-wise risk of being adversary $b_t^{i,N} = p(\theta|h_t^i; m_{0:t})$ is:

$$\widehat{\beta}_t^N \in \arg\max \frac{1}{N} \sum_{j=1}^{N} \log\Big[(1-u)L_{0,t}(h_t^j; m_{0:t}) + uL_{1,t}(h_t^j; m_{0:t})\Big], \tag{4}$$

$$b_t^{i,N} := \frac{\widehat{\beta}_t^N L_{1,t}(h_t^i; m_{0:t})}{(1-\widehat{\beta}_t^N)L_{0,t}(h_t^i; m_{0:t}) + \widehat{\beta}_t^N L_{1,t}(h_t^i; m_{0:t})}. \tag{5}$$

In practice, both equations requires reasoning over the trajectory-level likelihood and is intractable. In practice, we estimate the population-level belief $\hat{\beta}_t$ and agent-level belief b_t^i via neural network with input and output as follows:

$$\zeta_t^N \to \widehat{\beta}_t^N = p(\beta|\mu_t, \nu_t), \quad p_t^i \to b_t^{i,N} = p(\theta^i|s_t^i, a_t^i, \hat{\beta}_t), \tag{6}$$

with both neural networks implemented to incorporate historical data. For agent-wise risk, we use a shared neural network for all agents. In this way, the central network first inform the attacker proportion ζ_t^N, which provides population context for a shared detector, then use it as a prior to infer the agent-wise risk p_t^i, with the risk of each agent inferred locally in one pass with $O(N)$ complexity.

Remark 1 (When the learned belief is accurate.) As Byzantine agents can only depress the team return by producing trajectory statistics a cooperative one would not, the average type-score error is bounded by the irreducible overlap of the two types' behavioral history plus the excess risk of the learned risk detector over the Bayes predictor on that evidence. See Proposition E.1.

4.2 BELIEF-AUGMENTED MEAN-FIELD CONTROL

With the risk of each agent clear, it is crucial for defenders to aggregate the information of its own type as defender-specific mean-field. The approach avoids aggregating the information of adversary and models the max-min structure between defenders and adversaries, with each type of agent making decisions using their own representations while maintaining scalability. Formally, considering the risk p_t^i of each agent, the belief-aggregate mean field $\hat{\mu}_t^N$ for defenders is:

$$\hat{\mu}_t^N := \frac{\sum_{i=1}^N (1-p_t^i)\delta_{s_t^i}}{\sum_{i=1}^N (1-p_t^i)}. \tag{7}$$

To understand, $1-p_t^i$ is the estimated probability that agent i is cooperative, and the mean field of defenders suppresses agents in proportion to their inferred risk with soft type assignment. Additionally, define the mean-field of defenders with known type as $\hat{\mu}_t^{\text{type}} = \sum_{i=1}^N (1-\theta^i)\delta_{s_t^i} / \sum_{i=1}^N (1-\theta^i)$. We further show the error between $\hat{\mu}_t^{\text{type}}$ and $\hat{\mu}_t^N$ are bounded.

Proposition 4.1 (Population-field stability under type-score error)**.** *Under the attack budget $\beta \le \bar{\beta} < 1$ and Assumption D.3,*

$$\mathbb{E} W_1\big(\widehat{\mu}(\eta_t^N), \mu^{\text{type}}(\eta_t^N)\big) \le \frac{2D_{\mathcal{S}}}{1-\bar{\beta}}\, e_t, \tag{8}$$

where W_1 is the 1-Wasserstein distance induced by the state metric $d_{\mathcal{S}}$, $D_{\mathcal{S}} := \operatorname{diam}(\mathcal{S}) = \sup_{s,s' \in \mathcal{S}} d_{\mathcal{S}}(s,s')$, and $e_t := \mathbb{E}[N^{-1}\sum_i |p_t^i - \theta_i|]$ is the expected population-average type-score error under the evaluated closed loop.

Proof sketch. Apply Kantorovich–Rubinstein duality to the normalized weight difference, use the cooperative-mass lower bound $1-\bar{\beta}$, and take expectations, see Appendix E.3.

4.3 ROBUST POLICY LEARNING AND DEPLOYMENT

Given the inferred risk of each agent and the belief-weighted mean field for defender, the next step is to learn equilibrium policies for optimal defense. We first introduce the learning process of value function that evaluates the current game status.

Value evaluation. Fix the detector, the selector α, and the current policy pair (π, π_α). The closed loop is then Markov in the joint pre-action augmented state $X_t := ((s_t^i, h_t^i, p_t^i, \theta_i))_{i=1}^N$, so the cooperative continuation value satisfies:

$$\begin{aligned} V_t(X_t) = \sum_{\mathbf{a}_t \in \mathcal{A}^N} \prod_{i:\theta_i=0} \pi(a_t^i \mid (s_t^i, p_t^i, \mu_t^N, \widehat{\mu}_t^N)) \prod_{i:\theta_i=1} \pi_\alpha(a_t^i \mid (s_t^i, \theta_i, \mu_t^N, \mu_t^{\text{adv}})) \\ \times \mathbb{E}[r_t + \gamma V_{t+1}(X_{t+1}) \mid X_t, \mathbf{a}_t] \end{aligned} \tag{9}$$

with r_t the team reward in equation 1. The evaluation operator behind equation 9 is a γ-contraction in the sup norm, so this policy-pair value is its unique fixed point and backward induction returns it exactly over the finite horizon (Proposition B.1). A shared centralized critic $v_\xi(C_t)$ is trained by minimizing $\mathcal{L}_{\text{TD}}(\xi) := \widehat{\mathbb{E}}[(v_\xi(C_t) - y_t)^2]$, where $y_t := r_t + \gamma \mathbf{1}_{\{t<T\}} v'_\xi(C_{t+1})$, C_t is a fixed-dimensional pre-action encoding of (t, X_t), v'_ξ is a frozen critic copy, $\widehat{\mathbb{E}}$ averages sampled transitions.

Policy Gradients. Given a Byzantine scenario θ, for attacker policy $\pi_\alpha(a_t^i \mid s_t^i, \theta^i, \mu_t^N, \mu_t^{\text{adv}})$, defender policy $\pi(a_t^i | s_t^i, p_t^i, \mu_t, \hat{\mu}_t)$, and $A_t(X_t, \mathbf{a}_t) := \mathbb{E}[r_t + \gamma V_{t+1}(X_{t+1}) \mid X_t, \mathbf{a}_t] - V_t(X_t)$. as advantage function, the policy gradient theorem gives the update of attackers and defenders as:

$$\nabla_\psi J_{N,T} = \mathbb{E}\Big[\sum_{t=0}^T \gamma^t \sum_{i:\theta_i=0} \nabla_\psi \log \pi_\psi(a_t^i \mid I_t^{b,i})\, A_t\Big], \tag{10}$$

$$\nabla_\omega J_{N,T} = \mathbb{E}\Big[\sum_{t=0}^T \gamma^t \sum_{i:\theta_i=1} \nabla_\omega \log \pi_{\alpha,\omega}(a_t^i \mid I_t^{i,\text{adv}})\, A_t\Big], \tag{11}$$

which is standard policy gradient under their own objectives. Note that the defender conditions on its own inferred risk and belief-weighted mean field, which is tractable and do not depend on agent number N. For fixed α and (π, π_α), let $J_{N,T}^{\text{type}}$ be the return obtained by replacing p_t^i with θ_i in defender inputs and mean-field aggregation, using the same policy maps and initial conditions. The following theorem bounds its gap from the learned return $J_{N,T}$.

Theorem 4.1 (Finite-horizon performance stability)**.** *Under Assumptions D.3 and D.7, for every admissible* (π, π_α) *and every finite horizon* T*,*

$$\left| J_{N,T}(\pi, \alpha, \pi_\alpha) - J^{\text{type}}_{N,T}(\pi, \alpha, \pi_\alpha) \right| \le K_T \sum_{t=0}^{T} \gamma^t e_t \le K_T \sum_{t=0}^{T} \gamma^t \mathcal{L}_t, \tag{12}$$

where

$$K_T := L_{\text{oc}} \left(1 + \frac{2D_{\mathcal{S}}}{1-\bar{\beta}}\right) \sum_{j=0}^{T} \left(\gamma L_P \left(1 + \frac{1}{1-\bar{\beta}}\right)\right)^j. \tag{13}$$

Here L_P *and* L_{oc} *are the coupled-transition and reward Lipschitz constants in Assumption D.7;* e_t *and* $\mathcal{L}_t := \mathbb{E}[N^{-1}\sum_i \text{BCE}(p_t^i, \theta_i)] = \mathbb{E}[N^{-1}\sum_i -\theta \log p - (1-\theta)\log(1-p)]$ *are evaluated under the learned loop, with* BCE *denoting binary cross entropy. Proposition E.1 gives* $e_t \le \mathcal{L}_t$*. Thus small detector loss controls the gap to the same controller with true-type inputs. Corollary F.0.1 gives the infinite-horizon extension.*

Proof sketch. Couple both systems with identical initial states and types. Propagating the field-error bound through the Lipschitz dynamics and summing errors yields $K_T \sum_{t=0}^{T} \gamma^t e_t$. $e_t \le \mathcal{L}_t$. See E.4

Adversarial training. The robust objective in equation 2 accounts for both the selection of compromised agents through α and their action policy π_α. During training, each episode samples an attack fraction β and selects $\lfloor \beta N \rfloor$ agents through α, exposing the detector to diverse type configurations. The defender policy π and adversarial policy π_α follow a two-timescale update schedule, with multiple adversary update blocks for each defender update block. The opposing actor is held fixed within each block. Each block collects fresh rollouts, which also provide data for critic and detector updates according to the training schedule (Algorithm 1).

Extending to infinite agents. Mean-field methods are defined in the limit $N \to \infty$, but deployment always faces a finite population, so the main text states everything at finite N and quantifies the limit as an approximation rather than assuming it. Appendix F develops the limiting theory and connects the two. The population flow is no longer generated mechanically by the agents but must be required to be consistent with the one the equilibrium prescriptions induce, so the solution concept becomes an ex-interim equilibrium of a distribution-space Bellman–Isaacs recursion (Definition F.1). And at a fixed stage, with that flow held fixed, the prevalence estimate converges in probability to the unique maximizer of a population log-likelihood, so the plug-in belief becomes a deterministic map of a tagged history, though the belief itself remains random through that history (Proposition F.1). Finally, with a fixed detector and selection mechanism, the empirical population flow converges over any fixed finite horizon at rate $O(\rho_N)$ under the stated assumptions (Theorem F.1), specializing to $O(N^{-1/2})$ in finite augmented-state models with matching initialization and sampling bounds.

Proposition 4.2 (Robust value under nested information)**.** *The robust values satisfy*

$$V^{\text{wc}}_{0,N,T} \le V^{\text{wc}}_{b,N,T} \le V^{\text{wc}}_{\theta,N,T}. \tag{14}$$

When the corresponding saddle points exist, $V^{\text{wc}}_{b,N,T}$ *is the partial-information robust Bayesian equilibrium payoff, whereas* $V^{\text{wc}}_{\theta,N,T} = J_{N,T}(\pi^\star_\theta, \pi^\star_{\alpha,\theta})$ *is the full-information robust equilibrium payoff and the oracle upper benchmark within the declared policy and adversary classes. The type-informed return in Theorem 4.1 need not equal this optimized oracle value; equality requires additional representation and optimization conditions.*

Proof sketch. Policies may ignore additional inputs, so $\Pi_0 \subseteq \Pi_b \subseteq \Pi_\theta$, where Π_0, Π_b, Π_θ be the policy classes using raw, belief-augmented, and full-type information, respectively. With common adversary classes, maximizing the worst-case return over these nested classes gives the ordering. When saddle points exist, their payoffs equal the corresponding values.

5 Experiment

Environments. We evaluate our algorithms on two mean-field environments. The first is **Battle** (Zheng et al., 2018), where a group of agents combat with another group. Adversaries are introduced in the group controlled by mean-field MARL which mimics partial agent failures. The

second is **PowerGrid** (Chen et al., 2022), where each agent is a node in the power grid that outputs continuous voltage. The failure of a subset of nodes mimics disaster that destroys part of the grid. We handle partial observability by aggregating the belief-weighted mean field of each agent by using its neighbor only. More experiment details are given in Appendix G.

Baselines. We consider baseline from robust small-scale MARL and large-scale mean-field MARL. For small-scale MARL, MIR3 (Li et al., 2025) gives a lightweight agent-wise regularization for MAPPO-size MARL. For large-scale MARL, we consider RoMFAC (Zhou & Liu, 2022) as state-robust mean-field MARL, MTMFQ (Subramanian et al., 2020a) as a variant of our method with hard type assignment, Robust MFTG (RMFTG) (Zaman et al., 2024) as a dynamic-robust mean-field MARL under linear-quadratic settings. We also include an ablation of our method without considering belief of agent types as *w/o belief*, which do not consider agent risk and the belief-aggregated mean-field in training. We do not compare EIR-MAPPO (Li et al., 2024) since the computation is intractable under our population sizes, with computational complexity reported in our discussions.

Evaluation protocol. We select an attack fraction β up to 50% of overall agents. Using the solution derived by VAI (Li et al., 2026b), we select the agents to attack via policy α at $t = 0$, and learn the attacker policy π_α, which serves as a worst-case attack against the system. For Battle, the adversary is selected to be $\beta N \in \{32, 64, 144\}$ agents and PowerGrid with $\beta N \in \{20, 40\}$ agents. For generality, we sweep $\beta \in [0.1, 0.5]$ uniformly at training time. We evaluate all methods with 5 random seeds, with results evaluated over 50 rollouts for each seeds.

5.1 Simulation Results

We first show our main result. As shown in Table 1, our BeMF shows consistently higher robustness in 16 out of 18 configurations with Byzantine agents. The advantage is consistent across environments with different interaction dynamics. Battle requires agents to coordinate movement and combat, whereas PowerGrid requires coordinated voltage regulation through physically coupled distributed energy resources. Additionally, we find BeMF effective even with large number of agents being compromised (*e.g.*, $\beta = 0.5$), showing the effectiveness under extreme conditions. In contrast, other defenses consider threats other than Byzantine faults, resulting in reduced defense capability.

Agent Num	Adv Num	MTMFQ	MIR3	RoMFAC	RMFTG	BeMF
Environment: Battle (↑)						
32	4	-510.50±43.15	-260.00±42.74	-214.67±65.33	-513.46±57.59	**-160.19±11.55**
	8	-613.46±29.03	-370.00±52.97	-596.67±36.19	-603.33±25.80	**-230.35±18.26**
	16	-704.34±43.70	-356.03±51.47	-690.00±61.71	-690.00±39.99	**-270.24±12.95**
64	8	-538.63±89.33	-1287.99±16.00	-1567.99±9.22	-487.65±24.37	**-347.68±9.72**
	16	-2407.76±48.53	-2183.19±15.80	-2150.81±15.46	-2045.80±13.52	**-1057.91±14.43**
	32	-1602.88±79.48	-2687.99±20.69	-1373.08±14.22	-1408.00±17.97	**-935.7±19.62**
144	16	-2207.94±70.31	-1957.30±57.30	-874.71±50.91	-840.94±16.51	**-783.23±19.79**
	32	-1546.60±25.43	-2258.54±71.59	**-1402.14±21.10**	-1568.00±127.10	-1450.11±21.03
	36	-1816.79±135.04	-2305.68±101.41	-1676.89±54.72	-1917.39±67.08	**-1593.89±22.50**
	48	-3132.48±250.14	-3126.77±48.47	-4113.16±218.77	-2208.01±14.96	**-2162.20±53.74**
	72	-3174.83±93.35	-3061.10±47.60	-3167.99±57.10	-3220.69±20.69	**-2555.94±26.53**
Environment: PowerGrid (↑)						
20	4	8.69±1.22	8.75±0.68	11.36±2.91	7.20±2.73	**12.97±1.54**
	8	7.75±1.02	6.64±1.38	7.58±1.42	**10.51±1.96**	10.30±1.80
	10	-5.92±2.83	-1.85±1.69	0.78±1.14	-1.88±1.27	**8.54±2.74**
40	8	14.29±2.14	9.74±2.95	21.17±3.28	16.56±2.66	**23.72±2.11**
	12	10.46±2.85	18.45±1.96	18.18±1.20	15.42±2.90	**19.75±1.81**
	16	16.21±2.78	14.47±1.91	3.33±1.08	13.21±0.78	**19.21±2.12**
	20	8.86±1.86	13.54±1.07	-1.95±2.67	13.77±0.99	**17.38±1.85**

Table 1: Defense result of BeMF and baselines in Battle and PowerGrid environment, under different agent numbers and attacker numbers. Bold fonts highlights the best result over baselines.

5.2 Discussions

Evolution of belief dynamics. To show how our algorithm works, we illustrate how the belief of each agent evolves over time. The setting is in Battle with 64 agents and 32 Byzantine agents. At t=0, as shown in Figure 2 (a,b), while adversary has been selected, no prior information has been observed by the defender. Every agent therefore carries the same prior belief of 50% to prepare for the worst-case. After only five step, as shown in Figure 2 (c,d), defenders quickly update their belief, reaching 90.36% accuracy for the true type. Defenders are then able to adapt their behaviors according to agent types. Benign agents advance to engage the opponent which demonstrate themselves in faithful types, while

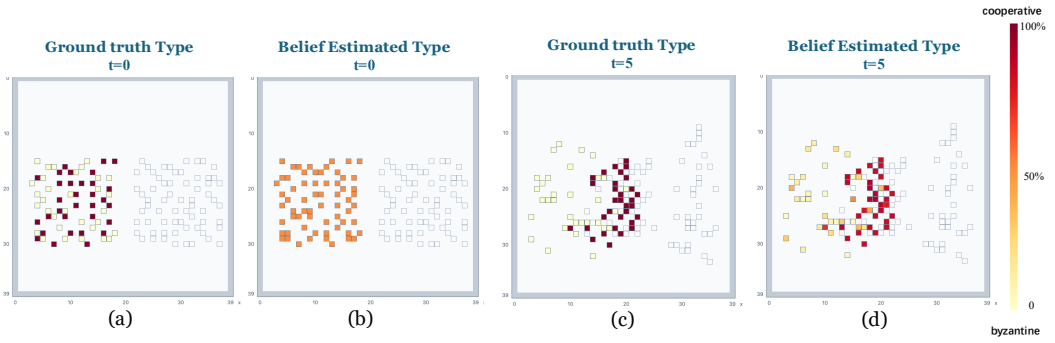


Figure 2: Belief-weighted mean field compare to ground-truth weighted mean field. The map is 40×40, cooperators controlling 64 agents, in which 32 agents are Byzantine.

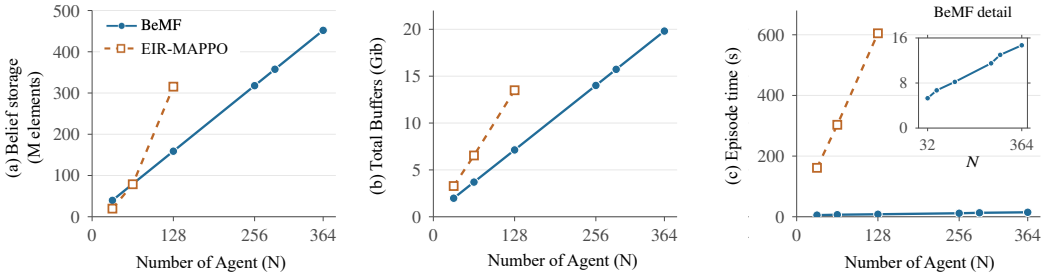


Figure 3: Computational cost of BeMF. BeMF reaches complexity of $O(N)$.

attackers stay behind or scatter, resulting in highly suspicious behaviors. At current stage, the type of several agents still remain uncertainty as shown in orange. As game proceeds further, the type of each agents become more certain and defenders are then able to separate the effect of attackers clearly, reaching score of 97.02% and take robust combat actions for victory.

BeMF scales linearly with the number of agents. Although BeMF models the type of every agent in the game, it never enumerates the 2^N possible attack configurations (*i.e.*, every combination of adversarial agents). It also avoids the $O(N^2)$ observer–target beliefs that EIR-MAPPO (Li et al., 2024) must maintain. Each agent stores only one additional belief bit and the aggregated belief mean field. Neither quantity depends on the number of agents, so the per-agent cost is $O(1)$ and the total cost is $O(N)$. This makes BeMF suitable for large agent populations. As shown in Figure 3, the training cost of BeMF grows linearly and stays tractable with up to 364 agents, while EIR-MAPPO quickly becomes intractable.

Ablations over belief. We first analyze the result of MTMFQ and the *w/o belief* baseline as our ablation. As shown in Table 4, the defense capability of *w/o belief* is lower than our method in all settings, showing robustness cannot be achieved without modeling compromised agents, consistent with Proposition 4.2. Additionally, MTMFQ represents our BeMF without explicit inference of defenders and adversaries. As a consequence, defenders are unable to accurately identify the true types of each agents and thus unable to achieve valid defense comparing with our BeMF.

6 Conclusion

We studied large-scale cooperative MARL in which an unknown subset of agents suffers Byzantine faults. Instead of hard type assignment, BeMF maintains continuous beliefs: a shared detector assigns each agent a Byzantine risk score, these scores form a belief-weighted mean field, and a shared policy acts on its local state, its risk score, and both the raw and belief-weighted mean fields. The policy is trained by alternating minimax against a worst-case attacker, and its per-step cost grows linearly in N with network sizes independent of N. On Battle and PowerGrid, BeMF achieves the highest return in 16 of 18 attacked configurations, including every configuration with half of the population compromised, and ablations attribute the gain to the belief mechanism rather than to adversarial training alone. Our evaluation fixes the Byzantine set within each episode, attackers that switch behavior over time or explicitly evade the detector remain future work.

AI USE STATEMENT

Generative AI tools were utilized to assist in organizing literature outlines write polishing and coding syntax for data preprocessing. No core data analysis, statistical evaluation, or qualitative interpretations were delegated to AI systems. All outputs were independently audited by our team.

REPRODUCIBILITY STATEMENT

We have open sourced all our code for robustness evaluation and enhancement in `https://anonymous.4open.science/r/BeMF-10E0-5649/`. Additional experiment details are available in Appendix A and G. Environment details and hyperparameters are provided in Appendix G. Our experiments involve environmental stochasticity, stochastic action selection, and random parameter initialization. These sources of randomness can propagate through multi-agent interactions and affect learning trajectories and final performance. Reproduction should therefore account for variability across random seeds and assess the consistency of aggregate results over repeated runs.

For computational efficiency, we use a JAX-based reimplementation of the Battle environment. This implementation is part of our experimental setup, and reproducing the reported Battle results should use the same environment implementation and configuration. Although we have run consistent test but differences due to implementations still maybe inevitable, including their handling of random events and simulation updates, may affect the resulting trajectories and scores; numerical agreement with the original Battle implementation is therefore not guaranteed.

REFERENCES

Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Laurière. Unified reinforcement q-learning for mean field game and control problems. *Mathematics of Control, Signals, and Systems*, 34(2): 217–271, 2022. doi: 10.1007/s00498-021-00310-1.

Uğur Aydin and Naci Saldi. Robustness and approximation of discrete-time mean-field games under discounted cost criterion. *Mathematics of Operations Research*, 51(1):185–217, 2025. doi: 10.1287/moor.2023.0316.

Dario Bauso, Hamidou Tembine, and Tamer Başar. Robust mean field games. *Dynamic Games and Applications*, 6(3):277–303, 2016. doi: 10.1007/s13235-015-0160-4.

Alexander Bukharin, Yan Li, Yue Yu, Qingru Zhang, Zhehui Chen, Simiao Zuo, Chao Zhang, Songan Zhang, and Tuo Zhao. Robust multi-agent reinforcement learning via adversarial regularization: Theoretical foundation and stable algorithms. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Dong Chen, Kaian Chen, Zhaojian Li, Tianshu Chu, Rui Yao, Feng Qiu, and Kaixiang Lin. PowerNet: Multi-agent deep reinforcement learning for scalable powergrid control. *IEEE Transactions on Power Systems*, 37(2):1007–1017, 2022.

Kai Cui, Gökçe Dayanıklı, Mathieu Laurière, Matthieu Geist, Olivier Pietquin, and Heinz Koeppl. Learning discrete-time major-minor mean field games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(9):9616–9625, 2024. doi: 10.1609/aaai.v38i9.28818.

François Delarue and Pierre Lavigne. Robust mean-field games under entropy-based uncertainty. *arXiv preprint arXiv:2603.18628*, 2026a. URL `https://arxiv.org/abs/2603.18628`.

François Delarue and Pierre Lavigne. Robust mean field control: Stochastic maximum principle and variational mean field games. *arXiv preprint arXiv:2604.21641*, 2026b. URL `https://arxiv.org/abs/2604.21641`.

Ilias Diakonikolas and Daniel M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019.

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-field controls with Q-learning for cooperative MARL: Convergence and complexity analysis. *SIAM Journal on Mathematics of Data Science*, 3(4):1168–1196, 2021. doi: 10.1137/20M1360700.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/030e65da2b1c944090548d36b244b28d-Abstract.html.

Jianhui Huang and Minyi Huang. Robust mean field linear-quadratic-gaussian games with unknown L^2-disturbance. *SIAM Journal on Control and Optimization*, 55(5):2811–2840, 2017. doi: 10.1137/15M1014437.

Jianhui Huang, Bing-Chang Wang, and Jiongmin Yong. Social optima in mean field linear-quadratic-gaussian control with volatility uncertainty. *SIAM Journal on Control and Optimization*, 59(2): 825–856, 2021. doi: 10.1137/19M1306737.

Minyi Huang, Roland P. Malhamé, and Peter E. Caines. Large population stochastic dynamic games: Closed-loop McKean–Vlasov systems and the nash certainty equivalence principle. *Communications in Information and Systems*, 6(3):221–252, 2006. doi: 10.4310/CIS.2006.v6.n3.a5.

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese Journal of Mathematics*, 2 (1):229–260, 2007. doi: 10.1007/s11537-007-0657-8.

Mathieu Laurière, Sarah Perrin, Sertan Girgin, Paul Muller, Ayush Jain, Théophile Cabannes, Georgios Piliouras, Julien Pérolat, Romuald Elie, Olivier Pietquin, and Matthieu Geist. Scalable deep reinforcement learning algorithms for mean field games. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12078–12095. PMLR, 2022. URL https://proceedings.mlr.press/v162/lauriere22a.html.

Mathieu Laurière, Ariel Neufeld, and Kyunghyun Park. Robust mean-field control under common noise uncertainty. *arXiv preprint arXiv:2511.04515*, 2025. URL https://arxiv.org/abs/2511.04515.

Shihui Li, Yi Wu, Xinyue Cui, Honghua Dong, Fei Fang, and Stuart Russell. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1):4213–4220, 2019. doi: 10.1609/aaai.v33i01.33014213.

Simin Li, Jun Guo, Jingqiao Xiu, Pu Feng, Xin Yu, Aishan Liu, Wenjun Wu, and Xianglong Liu. Attacking cooperative multi-agent reinforcement learning by adversarial minority influence. *arXiv preprint arXiv:2302.03322*, 2023.

Simin Li, Jun Guo, Jingqiao Xiu, Ruixiao Xu, Xin Yu, Jiakai Wang, Aishan Liu, Yaodong Yang, and Xianglong Liu. Byzantine robust cooperative multi-agent reinforcement learning as a bayesian game. 2024.

Simin Li, Ruixiao Xu, Jingqiao Xiu, Yuwei Zheng, Pu Feng, Yaodong Yang, and Xianglong Liu. Robust multi-agent reinforcement learning by mutual information regularization. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Simin Li, Zihao Mao, Hanxiao Li, Zonglei Jing, Jun Guo, Li Wang, Zhuoran Han, Ruixiao Xu, Xin Yu, Chengdong Ma, et al. Empirical study on robustness and resilience in cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 38:44009–44052, 2026a.

Simin Li, Zihao Mao, Zheng Yuwei, Linhao Wang, Ruixiao Xu, Chengdong Ma, Zhiqian Liu, Xin Yu, Yuqing Ma, Xin Wang, Jie Luo, Bo An, Yaodong Yang, Weifeng Lv, and Xianglong Liu. Vulnerable agent identification in large-scale multi-agent reinforcement learning. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=Q1twBCjpUt.

Yan Li, Lingxiao Wang, Jiachen Yang, Ethan Wang, Zhaoran Wang, Tuo Zhao, and Hongyuan Zha. Permutation invariant policy optimization for mean-field multi-agent reinforcement learning: A principled approach. *arXiv preprint arXiv:2105.08268*, 2021.

Yong Liang, Bing-Chang Wang, and Huanshui Zhang. Robust mean field linear quadratic social control: Open-loop and closed-loop strategies. *SIAM Journal on Control and Optimization*, 60(4): 2184–2213, 2022. doi: 10.1137/21M1414140.

Jieyu Lin, Kristina Dzeparoska, Sai Qian Zhang, Alberto Leon-Garcia, and Nicolas Papernot. On the robustness of cooperative multi-agent reinforcement learning. In *2020 IEEE Security and Privacy Workshops (SPW)*, pp. 62–68. IEEE, 2020.

Bora Ly and Romny Ly. Cybersecurity in unmanned aerial vehicles (uavs). *Journal of Cyber Security Technology*, 5(2):120–137, 2021.

Washim Uddin Mondal, Vaneet Aggarwal, and Satish V. Ukkusuri. Can mean field control (MFC) approximate cooperative multi agent reinforcement learning (MARL) with non-uniform interaction? In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1371–1380. PMLR, 2022. URL https://proceedings.mlr.press/v180/mondal22a.html.

Jun Moon and Tamer Başar. Linear quadratic risk-sensitive and robust mean field games. *IEEE Transactions on Automatic Control*, 62(3):1062–1077, 2017. doi: 10.1109/TAC.2016.2579264.

Duc Thien Nguyen, Akshat Kumar, and Hoong Chuin Lau. Credit assignment for collective multiagent rl with global rewards. *Advances in neural information processing systems*, 31, 2018.

Thomy Phan, Thomas Gabor, Andreas Sedlmeier, Fabian Ritz, Bernhard Kempter, Cornel Klein, Horst Sauer, Reiner Schmid, Jan Wieghardt, Marc Zeller, and Claudia Linnhoff-Popien. Learning and testing resilience in cooperative multi-agent systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1055–1063. International Foundation for Autonomous Agents and Multiagent Systems, 2020. URL https://www.ifaamas.org/Proceedings/aamas2020/pdfs/p1055.pdf.

Thomy Phan, Lenz Belzner, Thomas Gabor, Andreas Sedlmeier, Fabian Ritz, and Claudia Linnhoff-Popien. Resilient multi-agent reinforcement learning with adversarial value decomposition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11308–11316, 2021. doi: 10.1609/aaai.v35i13.17348.

Nabil Schear, Patrick T Cable, Thomas M Moyer, Bryan Richard, and Robert Rudd. Bootstrapping and maintaining trust in the cloud. In *Proceedings of the 32nd Annual Conference on Computer Security Applications*, pp. 65–77, 2016.

Jayakumar Subramanian and Aditya Mahajan. Reinforcement learning in stationary mean-field games. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 251–259. International Foundation for Autonomous Agents and Multiagent Systems, 2019. URL https://www.ifaamas.org/Proceedings/aamas2019/pdfs/p251.pdf.

Sriram Ganapathi Subramanian, Pascal Poupart, Matthew E Taylor, and Nidhi Hegde. Multi type mean field reinforcement learning. *arXiv preprint arXiv:2002.02513*, 2020a.

Sriram Ganapathi Subramanian, Matthew E Taylor, Mark Crowley, and Pascal Poupart. Partially observable mean field reinforcement learning. *arXiv preprint arXiv:2012.15791*, 2020b.

Sriram Ganapathi Subramanian, Matthew E. Taylor, Mark Crowley, and Pascal Poupart. Decentralized mean field games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):9439–9447, 2022. doi: 10.1609/aaai.v36i9.21176.

Yanchao Sun, Ruijie Zheng, Parisa Hassanzadeh, Yongyuan Liang, Soheil Feizi, Sumitra Ganesh, and Furong Huang. Certifiably robust policy learning against adversarial communication in multi-agent systems. *arXiv preprint arXiv:2206.10158*, 2022.

Chen Tessler, Yonathan Efroni, and Shie Mannor. Action robust reinforcement learning and applications in continuous control. In *International Conference on Machine Learning*, pp. 6215–6224. PMLR, 2019.

Amoolya Tirumalai and John S. Baras. A robust mean-field game of Boltzmann–Vlasov-like traffic flow. In *2022 American Control Conference*, pp. 556–561. IEEE, 2022. doi: 10.23919/ACC53348.2022.9867331.

Zishen Wan, Nandhini Chandramoorthy, Karthik Swaminathan, Pin-Yu Chen, Kshitij Bhardwaj, Vijay Janapa Reddi, and Arijit Raychowdhury. Mulberry: Enabling bit-error robustness for energy-efficient multi-agent autonomous systems. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pp. 746–762, 2024.

Bing-Chang Wang, Jianhui Huang, and Ji-Feng Zhang. Social optima in robust mean field LQG control: From finite to infinite horizon. *IEEE Transactions on Automatic Control*, 66(4):1529–1544, 2021a. doi: 10.1109/TAC.2020.2996189.

Jianhong Wang, Wangkun Xu, Yunjie Gu, Wenbin Song, and Tim C Green. Multi-agent reinforcement learning for active voltage control on power distribution networks. *Advances in Neural Information Processing Systems*, 34:3271–3284, 2021b.

Yue Wang. Stationary robust mean-field games under model mismatches. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 7229–7272. PMLR, 2026. URL `https://proceedings.mlr.press/v337/wang26i.html`.

Fan Yang, Alina Vereshchaka, Changyou Chen, and Wen Dong. Bayesian multi-type mean field multi-agent imitation learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yaodong Yang and Jun Wang. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*, 1, 2020.

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *International conference on machine learning*, pp. 5571–5580. PMLR, 2018.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International conference on machine learning*, pp. 5650–5659. Pmlr, 2018.

Muhammad Aneeq uz Zaman, Mathieu Laurière, Alec Koppel, and Tamer Başar. Robust cooperative multi-agent reinforcement learning: A mean-field type game perspective. In *Proceedings of the 6th Annual Learning for Dynamics & Control Conference*, pp. 770–783. PMLR, 2024.

Kaiqing Zhang, Tao Sun, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Basar. Robust multi-agent reinforcement learning with model uncertainty. *Advances in neural information processing systems*, 33:10571–10583, 2020.

Lingzhe Zhang, Tong Jia, Mingyu Wang, Weijie Hong, Chiming Duan, Minghua He, Rongqian Wang, Xi Peng, Meiling Wang, Nicholas Zhang, et al. Efficient failure management for multi-agent systems with reasoning trace representation. In *Proceedings of the 34th ACM International Conference on the Foundations of Software Engineering*, pp. 1222–1226, 2026.

Lianmin Zheng, Jiacheng Yang, Han Cai, Ming Zhou, Weinan Zhang, Jun Wang, and Yong Yu. Magent: A many-agent reinforcement learning platform for artificial collective intelligence. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Ziyuan Zhou and Guanjun Liu. Romfac: A robust mean-field actor-critic reinforcement learning against adversarial perturbations on states. *arXiv preprint arXiv:2205.07229*, 2022.

APPENDIX FOR "ROBUST LARGE-SCALE MULTI-AGENT REINFORCEMENT LEARNING AGAINST BYZANTINE CORRUPTION"

A DEPLOYMENT AND IMPLEMENTATION DETAILS

A.1 FIXED-DIMENSIONAL POPULATION INTERFACE

Let $\varphi_s : \mathcal{S} \to \mathbb{R}^{d_s}$ be a shared, possibly learned state encoder whose dimension d_s is independent of N. The trusted monitor computes

$$c_t^N := \frac{1}{N}\sum_{j=1}^{N} \varphi_s(s_t^j), \qquad \widehat{c}_t^N := \frac{\sum_{j=1}^{N}(1-p_t^j)\varphi_s(s_t^j)}{\sum_{j=1}^{N}(1-p_t^j)}. \tag{15}$$

These summaries are computed once and broadcast to all cooperative agents; no agent receives the latent types or an identity-indexed belief table. The deployed protected controller uses

$$I_{t,\mathrm{dep}}^{b,i} := (s_t^i, p_t^i, c_t^N, \widehat{c}_t^N),$$

and samples its own action from $\pi_{\mathrm{def}}(\cdot \mid I_{t,\mathrm{dep}}^{b,i})$.

A.2 COMPLEXITY AND SIZE-INDEPENDENT ARCHITECTURE

Proposition A.1 (Linear deployment complexity)**.** *One step of Algorithm 2 costs $O(N)$ time and $O(Nd_h)$ recurrent memory, where $d_h = \dim(\mathcal{H})$ is independent of N. Its trainable parameter dimensions are independent of population size. If the detector additionally processes at most k neighbors per target, the time cost is $O(Nk)$ and remains linear for bounded k.*

Proof. The detector, pooled-memory average, state-summary encoder, policy, and weighted-summary accumulation each make one shared pass over the N target agents. The trusted monitor computes c_t^N and $\widehat{c}_t^N$ once per time step, after which every policy evaluation reads only these fixed-dimensional summaries. Each target stores one fixed-dimensional memory and one scalar posterior; no observer–target matrix or per-policy pass over the full empirical population is constructed. A bounded-neighborhood implementation performs at most k additional shared operations per target. Since all trainable maps accept fixed-dimensional local features and permutation-invariant summaries, neither their input dimensions nor their parameter counts depend on N. □

A.3 TRAINING-TIME COMPLEXITY OF THE BELIEF REPRESENTATION

Proposition A.1 bounds the cost of *deploying* a trained defender. Training carries a second cost that deployment does not: an on-policy learner must hold an entire rollout in memory before it can update, and the width of what it holds is fixed by the belief representation. This is where a pairwise representation becomes expensive, and the reason is arithmetic rather than implementation.

Fix a rollout of horizon T collected over E parallel copies of the environment. It contains $N\,T\,E$ agent-observations, one per agent per step per copy. Write d_{raw} for the width of an agent's local observation before any belief information is attached. Under BeMF an agent acts on its local observation together with a risk coordinate summarising the population, so each stored observation is $d_{\mathrm{raw}} + d_c$ numbers wide, where d_c is a constant set by the encoder and does not depend on how many agents there are. Under a pairwise-belief equilibrium such as Li et al. (2024), agent i instead acts on one belief entry per teammate, $b_i = (b_{i\to j})_{j\neq i}$, so each stored observation is $d_{\mathrm{raw}} + N$ numbers wide. Keeping only the part that carries belief information, the two rollouts occupy

$$\underbrace{N\,T\,E\cdot d_c}_{\text{BeMF}} \qquad \text{and} \qquad \underbrace{N\,T\,E\cdot N}_{\text{pairwise}}$$

numbers respectively. The first grows in proportion to the population; the second grows with its square, because every one of the N agents stores a vector whose length is itself N.

A pairwise learner in fact pays this twice. Besides the belief vector inside each observation, every agent retains the joint type θ of the whole population as a supervision target for its detector, a second $N\,T\,E \cdot N$ entries. The two representations therefore break even at $N = d_c$ if one counts only the belief vector, or at $N = d_c/2$ once the type labels are included, and past that point the ratio grows in direct proportion to the population,

$$\frac{\text{belief storage, pairwise}}{\text{belief storage, BeMF}} = \frac{2N}{d_c}.$$

A third quantity scales the same way whenever the adversary is trained against the rest of the population: its observation concatenates the actions of the other $N-1$ agents, so an update over all N agents carries on the order of $N^2|\mathcal{A}|$ action entries. Unlike the storage terms above this one cannot be reduced by splitting the update into smaller batches, since the width of a single adversary observation already grows with N.

Finally, the widths themselves differ in kind. Every BeMF map reads fixed-dimensional local features and permutation-invariant population summaries, so its input widths and parameter counts are the same at every population size. A pairwise policy reads b_i directly, so its input layer has width proportional to N: a policy trained at one population size cannot even be evaluated at another without being rebuilt and retrained.

Empirical verification. We instantiate both representations on Battle with an otherwise identical configuration (map $= 60$, $T = 400$, $E = 24$, $d_{\text{raw}} = 2231$, $d_c = 129$, EP critic, no adversaries, one A800-80GB) and measure the rollout buffers directly rather than sampling process-level peaks, which absorb transient allocations. Table 2 reports the result.

Table 2: Training cost against population size. Belief storage is reported in elements so that the comparison is independent of dtype (BeMF stores in `bfloat16`, which halves its byte cost again). The measured ratio matches $2N/d_c$ to three decimals at every population size: the factor two arises because the pairwise representation stores the belief row inside each observation *and* a full type-label array. Fitted log-log slopes of belief storage are 1.000 for BeMF across all four intervals and 2.000 for the pairwise baseline. A dash(—) marks a measurement that does not exist because the run did not produce one: at $N \geq 256$ the pairwise baseline exhausted the 80 GB device during its first actor update, before any training step completed. That failure is a limitation of the baseline's batching strategy rather than of the pairwise representation, and we do not count it as evidence;

		belief storage (M elements)				total buffers (GiB)		s / episode	
$\lvert\mathcal{N}\rvert$	N	BeMF	pairwise	ratio	$4\lvert\mathcal{N}\rvert/d_c$	BeMF	pairwise	BeMF	pairwise
16	32	39.7	19.7	0.496	0.496	1.98	3.28	5.3	161.3
32	64	79.5	78.8	0.992	0.992	3.70	6.54	6.7	303.2
64	128	158.9	315.4	1.984	1.984	7.13	13.50	8.2	604.9
128	256	317.8	—	—	3.969	14.01	—	11.5	—
144	288	357.6	—	—	4.465	15.73	—	13.0	—
182	364	451.9	—	—	5.643	19.81	—	14.7	—

Where the quadratic term is Aggregate storage is dominated by d_{raw} whenever $N \lesssim d_{\text{raw}}$, so the *total* buffer size grows approximately linearly under both representations (measured log-log slopes 0.98 and 1.05); at $N = 64$ the quadratic terms account for 8.7% of the pairwise baseline's storage. What separates the two is the belief-attributable term, which is exactly $\Theta(N)$ and $\Theta(N^2)$ as Table 2 shows. We therefore report the crossover $N = d_c$ rather than a fitted exponent on aggregate memory: on Battle the spawn geometry caps the population at $N = 182$, and reaching the regime where the quadratic term dominates aggregate storage would require $N \gtrsim d_{\text{raw}}$, which in turn needs a map large enough that the global state itself becomes the dominant cost.

B The BeMF Algorithm

B.1 Training and Optimization Details

The Bellman–Isaacs characterization in Appendix F requires optimization over augmented population laws. BeMF approximates this control problem through simulation and alternating PPO.

Actors and critic. Defenders share an actor π_ψ, Byzantine agents share an actor $\pi_{\alpha,\omega}$, and the detector has parameters ϕ. A centralized critic v_ξ, with parameters ξ, estimates the cooperative continuation return under the current policy pair:

$$v_\xi(C_t) \approx \mathbb{E}_{\pi_\psi,\pi_{\alpha,\omega}}\left[\sum_{\ell=t}^{T}\gamma^{\ell-t}r_\ell \,\middle|\, C_t\right], \tag{16}$$

where $C_t := \left(t, c_t^N, \widehat{c}_t^N, \zeta_t^N\right)$ is its input before the current actions are sampled and r_ℓ is the team reward in equation 1. Shared parameters can produce different values for different inputs. The critic evaluates the current policies, rather than the optimal Bellman–Isaacs value.

B.1.1 Policy evaluation and critic fitting

With the detector and the policy pair fixed, the closed loop is Markov in the augmented state X_t, which carries all physical states, detector memories, scores, and types (Appendix C); the explicit time index allows nonstationary finite-horizon policies. In equation 9, $V_t(x) := \mathbb{E}[G_t \mid X_t = x]$, with the dependence on the fixed pair suppressed.

Proposition B.1 (Convergence of exact policy evaluation)**.** *Suppose rewards are bounded and the detector, selector, and policy pair are fixed. For bounded value sequences $U = (U_t)_{t=0}^{T+1}$ with $U_{T+1} = 0$, define*

$$(\mathcal{B}U)_t(x) := \mathbb{E}[r_t + \gamma U_{t+1}(X_{t+1}) \mid X_t = x], \qquad (\mathcal{B}U)_{T+1} := 0.$$

Under $\|U\|_\infty := \max_t \sup_x |U_t(x)|$, the operator $\mathcal{B}$ is a γ-contraction whose unique fixed point is the continuation-value sequence V in equation 9, so m sweeps contract the error by γ^m. Over this finite horizon, backward induction computes V exactly and synchronous evaluation is exact after at most $T+1$ sweeps.

Proof. For any U, U' with zero terminal value, $|(\mathcal{B}U)_t(x) - (\mathcal{B}U')_t(x)| \le \gamma\,\mathbb{E}[|U_{t+1}(X_{t+1}) - U'_{t+1}(X_{t+1})| \mid X_t = x] \le \gamma\|U - U'\|_\infty$, and taking suprema gives the contraction. Conditioning G_t on the next state shows that V is a fixed point, and each sweep resolves one more stage from the terminal boundary. □

The neural critic reads a fixed-dimensional, permutation-invariant pre-action encoding C_t of (t, X_t), a compression that need not preserve the exact Markov state. Let v'_ξ be a copy of v_ξ frozen while targets are formed. The one-step TD fit is

$$\mathcal{L}_{\mathrm{TD}}(\xi) := \widehat{\mathbb{E}}\left[(v_\xi(C_t) - y_t)^2\right], \qquad y_t := r_t + \gamma \mathbf{1}_{\{t<T\}} v'_\xi(C_{t+1}), \tag{17}$$

where $\widehat{\mathbb{E}}$ averages sampled transitions and no gradient passes through y_t. PPO uses the generalized advantage built from the same residuals and its return target,

$$\widehat{A}_t := \sum_{\ell=0}^{T-t}(\gamma\lambda)^\ell\big(y_{t+\ell} - v'_\xi(C_{t+\ell})\big), \qquad \widehat{G}_t := v'_\xi(C_t) + \widehat{A}_t, \qquad \lambda \in [0,1], \tag{18}$$

so fitting v_ξ to $\widehat{G}_t$ is the λ-return form of equation 17, recovered at $\lambda = 0$. Both use a true terminal boundary at T; a rollout truncated before termination bootstraps at its endpoint instead. Proposition B.1 concerns the exact expectation operator on X_t, not a compressed neural critic or jointly changing actors and detector.

B.1.2 PPO, detector fitting, and update timescales

Let π_ψ^{old} be the behavior actor that collected the current block and $\rho_t^i := \pi_\psi(a_t^i \mid I_t^{b,i})/\pi_\psi^{\mathrm{old}}(a_t^i \mid I_t^{b,i})$. A defender block maximizes the clipped surrogate

$$\mathcal{L}_{\mathrm{PPO}} := \widehat{\mathbb{E}}\left[\sum_{t=0}^{T}\gamma^t \sum_{i:\theta_i=0} \min\Big\{\rho_t^i\widehat{A}_t,\ \mathrm{clip}(\rho_t^i, 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}})\widehat{A}_t\Big\}\right], \tag{19}$$

with clipping radius $\epsilon_{\mathrm{clip}} > 0$ and $\widehat{\mathbb{E}}$ the empirical average over fresh episodes. An adversary block is equation 19 with $\pi_{\alpha,\omega}$, $I_t^{i,\mathrm{adv}}$, the agents $\{i : \theta_i = 1\}$, and $-\widehat{A}_t$ replacing their defender counterparts.

Stored beliefs, behavior log probabilities, and advantage targets stay fixed during these updates, and the defender's population encoder, being part of its action kernel, is updated inside the defender block. Lemma B.1 gives the exact gradient identities for fixed inference and feature maps; clipping and multiple epochs are approximate policy-improvement steps.

The detector is fitted separately from the recorded histories and the training-time types by

$$\mathcal{L}_{\mathrm{det}}(\phi) := \frac{1}{N(T+1)} \sum_{t=0}^{T} \sum_{i=1}^{N} \mathrm{BCE}(p_t^i, \theta_i), \tag{20}$$

the empirical counterpart of $\mathcal{L}_t$ in Theorem 4.1, with every p_t^i recomputed under the current ϕ. No actor objective backpropagates into ϕ, and the type labels are withdrawn at deployment.

The schedule repeats $K_{\mathrm{adv}} > 1$ adversary blocks per defender block, each on fresh rollouts with the opposing actor held fixed, implementing the faster adversary and slower defender updates of EIR-MAPPO (Li et al., 2024); critic updates follow the configured schedule and the detector is refitted between actor blocks. Here K_{adv} is a configured interval, not a convergence rate: a two-timescale stochastic-approximation argument would additionally require a vanishing defender-to-adversary step-size ratio, suitable critic tracking, noise and boundedness conditions, and stable limiting dynamics, none of which a fixed block ratio, clipped neural updates, and a changing detector verify. Our guarantees therefore concern exact policy evaluation, detection-dependent performance stability, and population approximation; the optimization residual is retained in Corollary F.2.1. The selector α is supplied by the declared attack protocol and receives no gradient in these blocks: sampling β and attack subsets gives scenario coverage rather than a certificate over selectors or attack fractions, and a rollout containing no Byzantine agents contributes zero adversary gradient.

Algorithm 1 BEMF training with alternating PPO blocks

Require: detector $(\Phi_\phi, \varphi_{c,\phi}, q_\phi)$; population encoder φ_s; defender π_ψ; adversary $\pi_{\alpha,\omega}$; critic v_ξ; attack-sampling protocol; actor and critic update schedule; horizon T.
Ensure: trained detector, population encoder, and defender actor.

for each scheduled actor block **do**
  choose a defender or adversary block and hold the opposing actor's parameters fixed
  collect fresh episodes under the current policy pair:
  sample initial states and the configured attack budget β; select exactly $\lfloor \beta N \rfloor$ agents through α and assign their types θ_i
  initialize detector memories h_0^i
  for $t = 0, \ldots, T$ **do**
    $\zeta_t^N \leftarrow N^{-1} \sum_j \varphi_{c,\phi}(h_t^j)$; $p_t^i \leftarrow q_\phi(h_t^i, \zeta_t^N)$ for all i
    $c_t^N \leftarrow N^{-1} \sum_i \varphi_s(s_t^i)$; $\widehat{c}_t^N \leftarrow \frac{\sum_i (1-p_t^i)\varphi_s(s_t^i)}{\sum_i (1-p_t^i)}$
    sample cooperative actions from $\pi_\psi(\cdot \mid s_t^i, p_t^i, c_t^N, \widehat{c}_t^N)$ and Byzantine actions from $\pi_{\alpha,\omega}(\cdot \mid I_t^{i,\mathrm{adv}})$
    execute the joint action; record the team reward, critic inputs, beliefs, and behavior-policy log probabilities
    if $t < T$ **then**
      receive $z_{t+1}^i = (o_{t+1}^i, a_t^i, r_t^i)$ and update $h_{t+1}^i \leftarrow \Phi_\phi(h_t^i, z_{t+1}^i)$ for all i
    end if
  end for
  compute return targets and cooperative advantages from the rollout and critic, with the appropriate terminal or bootstrap masks
  if defender block **then**
    update π_ψ and its population encoder by the defender PPO surrogate, using stored beliefs as fixed inputs
  else
    update $\pi_{\alpha,\omega}$ by the adversary PPO surrogate using negated cooperative advantages
  end if
  if critic update scheduled **then**
    fit v_ξ to the cooperative return targets
  end if
  update ϕ by BCE on the recorded histories and type labels; no actor gradient is passed into the detector
end for
return frozen $(\Phi_\phi, \varphi_{c,\phi}, q_\phi)$, φ_s, and π_ψ

Algorithm 2 BEMF deployment: equivariant type belief and defended mean-field rollout

Require: frozen detector (Φ, φ_c, q) and protected policy π_{def} from Algorithm 1; initial memories $\{h_0^i\}_{i=1}^N$.

Ensure: target-wise posteriors and cooperative rollout.

Latent types $\{\theta_i\}$, the prior β, the selection mechanism α, and the Byzantine action policy π_α are unobserved; Byzantine agents act only through the environment.

for $t = 0, 1, 2, \ldots$ **do**
  $\zeta_t^N \leftarrow N^{-1} \sum_j \varphi_c(h_t^j)$
  for all target agents i **do**
    $p_t^i \leftarrow q(h_t^i, \zeta_t^N)$
  end for
  $c_t^N \leftarrow N^{-1} \sum_i \varphi_s(s_t^i)$
  $\widehat{c}_t^N \leftarrow \big(\sum_i (1 - p_t^i)\varphi_s(s_t^i)\big) / \big(\sum_i (1 - p_t^i)\big)$
  for all cooperative agents i **do**
    $a_t^i \sim \pi_{\mathrm{def}}(\cdot \mid s_t^i, p_t^i, c_t^N, \widehat{c}_t^N)$
  end for
  Byzantine agents emit $a_t^i \sim \pi_\alpha(\cdot \mid I_t^{i,\mathrm{adv}})$ through the environment
  execute actions and observe $z_{t+1}^i = (o_{t+1}^i, a_t^i, r_t^i)$ for all targets
  for all target agents i **do**
    update h_{t+1}^i by

$$h_{t+1}^i = \Phi\big(h_t^i, o_{t+1}^i, a_t^i, r_t^i, \zeta_t^N, \widehat{c}_t^N\big)$$

  end for
end for

B.2 Actor Updates and the Team Objective

For the identities below, fix the detector, population encoders, initial-state law, attack budget, and selection mechanism. The actor parameters ψ and ω affect only their respective conditional action kernels. Define the continuation return $G_t := \sum_{\ell=t}^T \gamma^{\ell-t} r_\ell$, where r_ℓ is the team reward in equation 1.

Lemma B.1 (Actor-gradient consistency)**.** *Suppose actions are sampled independently conditional on the joint pre-action history and types. Assume differentiable policy densities with parameter-independent support and integrable scores, so that differentiation and expectation may be interchanged. Derivatives include all actor computations defining the conditional action kernel. For fixed critic parameters ξ and a pre-action critic input C_t,*

$$\nabla_\psi J_{N,T} = \mathbb{E}\left[\sum_{t=0}^T \gamma^t \sum_{i:\theta_i=0} \nabla_\psi \log \pi_\psi(a_t^i \mid I_t^{b,i})\big(G_t - v_\xi(C_t)\big)\right], \tag{21}$$

$$\nabla_\omega(-J_{N,T}) = \mathbb{E}\left[\sum_{t=0}^T \gamma^t \sum_{i:\theta_i=1} \nabla_\omega \log \pi_{\alpha,\omega}(a_t^i \mid I_t^{i,\mathrm{adv}})\big(v_\xi(C_t) - G_t\big)\right]. \tag{22}$$

Here $J_{N,T} = J_{N,T}(\pi_\psi, \pi_{\alpha,\omega})$ and the expectation is under this policy pair. A separate pre-action input may be used for each agent without changing the identities.

Proof. In the joint trajectory likelihood, initial conditions, type selection, observations, and environment transitions have no direct dependence on the actor parameters. Differentiating its logarithm with respect to ψ therefore gives the sum of the cooperative action scores. The likelihood-ratio identity yields

$$\nabla_\psi J_{N,T} = \mathbb{E}\left[\left(\sum_{\ell=0}^T \gamma^\ell r_\ell\right) \sum_{t=0}^T \sum_{i:\theta_i=0} \nabla_\psi \log \pi_\psi(a_t^i \mid I_t^{b,i})\right].$$

Conditionally on the joint pre-action history and types, each action score has mean zero, since the corresponding action density integrates to one. Rewards before t are measurable with respect to that history, so their contribution vanishes. The remaining reward sum is $\gamma^t G_t$. The same zero-mean identity permits subtracting any integrable pre-action baseline $v_\xi(C_t)$, proving equation 21.

Differentiating with respect to ω instead selects the Byzantine action scores. Negating the objective and using the same baseline gives equation 22. □

Algorithm 1 approximates the control objective through alternating PPO and supervised type inference. Lemma B.1 explains the opposing actor directions under fixed inference and feature maps, while Corollary F.2.1 characterizes the transfer of residual strategic error to finite populations

C THE BELIEF-AUGMENTED POPULATION STATE

Let $(\mathcal{S}, d_{\mathcal{S}})$ and $(\mathcal{H}, d_{\mathcal{H}})$ be compact, write $D_{\mathcal{S}} := \operatorname{diam}(\mathcal{S})$, equip $\Theta = \{0, 1\}$ with the discrete metric, and set $\mathcal{X} := \mathcal{S} \times \mathcal{H} \times [0, 1] \times \Theta$ with the sum metric. Section 4.2 defines the augmented per-agent state $x_t^i := (s_t^i, h_t^i, p_t^i, \theta_i) \in \mathcal{X}$ and the augmented empirical population $\eta_t^N := N^{-1} \sum_{i=1}^N \delta_{x_t^i} \in \mathcal{P}(\mathcal{X})$. For any $\eta \in \mathcal{P}(\mathcal{X})$, the raw, belief-weighted, and type-weighted state fields are

$$\mu(\eta) := \int_{\mathcal{X}} \delta_s \, \eta(ds, dh, dp, d\theta), \tag{23}$$

$$\widehat{\mu}(\eta) := \frac{\int (1-p) \, \delta_s \, \eta(ds, dh, dp, d\theta)}{\int (1-p) \, \eta(ds, dh, dp, d\theta)}, \tag{24}$$

$$\mu^{\text{type}}(\eta) := \frac{\int (1-\theta) \, \delta_s \, \eta(ds, dh, dp, d\theta)}{\int (1-\theta) \, \eta(ds, dh, dp, d\theta)}. \tag{25}$$

Evaluating them at the empirical law recovers the finite-population objects of the main text: $\mu(\eta_t^N) = \mu_t^N$, $\widehat{\mu}(\eta_t^N) = \widehat{\mu}_t^N$ as defined in equation 7, and $\mu^{\text{type}}(\eta_t^N) = \mu_t^{\text{type},N}$. The adversarial field of Section 3.1 is the same construction with weights θ in place of $1 - \theta$, so it too is a functional of η_t^N. Only μ and $\widehat{\mu}$ are available at deployment. The type-weighted field reads the latent type coordinate and is used exclusively as the analysis-only comparator in Proposition 4.1 and Theorem 4.1; it changes the aggregation weights and does not re-optimize the protected policy.

The augmented population law is closed under the closed-loop update, which is what licenses treating it as a Markov state. The detector memory summarizes a target's own behavioral history by construction, $h_{t+1}^i = \Phi(h_t^i, z_{t+1}^i)$; the score is the shared deterministic map $p_t^i = q(h_t^i, \zeta_t^N) \in [0, 1]$ and the pooled summary $\zeta_t^N = N^{-1} \sum_j \varphi_c(h_t^j)$, which is itself a functional of η_t^N; and both types act on information that is a function of x_t^i together with $\mu(\eta_t^N)$ and $\widehat{\mu}(\eta_t^N)$. Hence the conditional law of x_{t+1}^i given the whole past depends on it only through (x_t^i, η_t^N): there is a shared kernel $Q(\cdot \mid x, \eta)$ applied agentwise, and η_{t+1}^N is generated from η_t^N alone. This is the closure that the mean-field flow equation 64 requires. It also explains why the belief has to live inside the population state rather than beside it: a controller that conditions on p_t^i but propagates only μ_t^N has no self-contained update, because the next round of scores depends on the entire memory profile, not on the state field.

The belief stays individual in the population limit. The belief coordinate does not collapse to a single scalar as N grows. The limiting object η_t is the joint law of physical state, detector memory, score, and latent type, so targets with different behavioral histories retain different scores, and $\widehat{\mu}(\eta_t)$ is a genuinely reweighted field rather than a globally rescaled copy of $\mu(\eta_t)$. This is what separates the construction from appending a finite-agent belief vector to a mean-field controller: such a vector has a dimension that grows with N and admits no population limit, whereas here the strategic state is the permutation-invariant law η_t, the closed-loop update is equivariant (Lemma F.1), and the empirical system converges to a self-consistent flow of that law (Theorem F.1).

D ASSUMPTIONS

Assumption D.1 (Empirically consistent initialization)**.** *Conditional on an episode-level attack fraction β, let*

$$x_0^i := (s_0^i, h_0^i, p_0^i, \theta_i), \qquad \eta_0^N := \frac{1}{N} \sum_{i=1}^N \delta_{x_0^i},$$

where the type profile follows the exact-count assignment in Definition 3.1, and

$$p_0^i = q\left(h_0^i, \frac{1}{N}\sum_{j=1}^{N}\varphi_c(h_0^j)\right).$$

There exist a deterministic population law $\eta_0 \in \mathcal{P}(\mathcal{X})$, a constant $C_{\mathrm{init}} < \infty$, and a deterministic sequence $\rho_N \downarrow 0$ such that

$$\mathbb{E}W_1(\eta_0^N, \eta_0) \le C_{\mathrm{init}}\rho_N. \tag{26}$$

Moreover,

$$\eta_0(\{(s,h,p,\theta) : \theta = 1\}) = \beta, \tag{27}$$

$$\eta_0\left(\left\{(s,h,p,\theta) : p = q\left(h, \int \varphi_c(h')\,\eta_0(ds', dh', dp', d\theta')\right)\right\}\right) = 1. \tag{28}$$

Assumption D.2 (Shared equivariant architecture)**.** *The joint initialization law is invariant under simultaneous permutations of the agent indices, and the declared exact-count selection mechanism α is permutation equivariant. The detector update function Φ, posterior head q, pooled-memory encoder φ_c, protected policy π, and Byzantine action policy π_α are shared across agent indices. They may depend on an agent only through its local trace and permutation-invariant empirical summaries.*

Assumption D.3 (Positive retained belief mass)**.** *There exists $\underline{z}_p > 0$, independent of N and of the admissible policy pair, such that every learned empirical or limiting population law encountered in the analysis satisfies*

$$Z_p(\eta) := \int (1-p)\,d\eta \ge \underline{z}_p.$$

Discussion of Assumption D.3. The lower bound on the true cooperative mass follows naturally from the threat model. If the admissible attack fractions satisfy $\beta \le \bar{\beta} < 1$ and exactly $\lfloor \beta N \rfloor$ agents are Byzantine, then, because agent types remain fixed throughout an episode,

$$Z_\theta(\eta_t^N) = 1 - \frac{\lfloor \beta N \rfloor}{N} \ge 1 - \bar{\beta}$$

for every t and N; the corresponding population law satisfies $Z_\theta(\eta_t) = 1 - \beta \ge 1 - \bar{\beta}$. Hence one may take $\underline{z}_\theta = 1 - \bar{\beta}$. The condition on Z_p rules out the degenerate failure mode in which the detector assigns nearly unit Byzantine probability to the entire population. It is implied, for example, by a uniform bound on the population-level type-score error, since

$$Z_p(\eta) = \int (1-p)\,d\eta \ge \int (1-\theta)\,d\eta - \int |p-\theta|\,d\eta = Z_\theta(\eta) - \varepsilon(\eta).$$

Consequently, if $\varepsilon(\eta) \le \bar{\varepsilon} < \underline{z}_\theta$ along the admissible trajectories, then $Z_p(\eta) \ge \underline{z}_\theta - \bar{\varepsilon} > 0$, and one may take $\underline{z}_p = \underline{z}_\theta - \bar{\varepsilon}$. Assumption D.3 therefore requires neither the realized attack fraction nor the latent types to be available to the deployed defender; it is a nondegeneracy condition on the threat class and the learned detector that ensures the normalized belief-weighted field remains well defined and uniformly stable.

Assumption D.4 (Augmented-state regularity and sampling error)**.** *Equip $\mathcal{X}$ with the sum metric $d_{\mathcal{X}}$, including the discrete metric on the type coordinate. The augmented transition kernel Q preserves the latent type and satisfies*

$$W_1(Q(\cdot \mid x, \eta), Q(\cdot \mid x', \eta')) \le L_Q(d_{\mathcal{X}}(x,x') + W_1(\eta, \eta')) \tag{29}$$

for all admissible $x, x' \in \mathcal{X}$ and $\eta, \eta' \in \mathcal{P}(\mathcal{X})$. There is a deterministic sequence $\rho_N \downarrow 0$ such that

$$\mathbb{E}\left[W_1\left(\eta_{t+1}^N, \int_{\mathcal{X}} Q(\cdot \mid x, \eta_t^N)\,\eta_t^N(dx)\right) \,\middle|\, \mathcal{F}_t\right] \le c_t\rho_N. \tag{30}$$

The constants L_Q and c_t do not depend on N.

Assumption D.5 (Stage reward and action sampling)**.** *The reward r is bounded and L_r-Lipschitz in (s, μ, ν), uniformly over actions. Conditionally on $\mathcal{F}_t$, the actions $\{a_t^i\}_{i=1}^N$ are independent with shared type-wise kernels $a_t^i \sim K_{\theta_i}(\cdot \mid x_t^i, \eta_t^N)$, obtained by composing the observation kernel with the type-θ_i policy. For this composition to inherit the regularity required by equation 29, we assume the observation kernel is Lipschitz in the W_1 sense,*

$$W_1\big(\Omega(\cdot \mid s, \mu), \Omega(\cdot \mid s', \mu')\big) \le L_\Omega\big(d_{\mathcal{S}}(s, s') + W_1(\mu, \mu')\big), \tag{31}$$

and that every admissible policy is Lipschitz in its inputs. The observation is therefore integrated out in K_θ and never appears as a coordinate of $\mathcal{X}$. For each admissible policy pair we suppress its dependence in K_θ, $\bar{\nu}$, and $\mathcal{R}$. The resulting mean action measure and mean-field stage reward are

$$\bar{\nu}(\eta) := \int K_\theta(\cdot \mid x, \eta)\, \eta(dx), \qquad \mathcal{R}(\eta) := \iint r\big(s, a, \mu(\eta), \bar{\nu}(\eta)\big)\, K_\theta(da \mid x, \eta)\, \eta(dx). \tag{32}$$

We assume that $\mathcal{R}$ is $L_{\mathcal{R}}$-Lipschitz on $\mathcal{P}(\mathcal{X})$ and that the conditional empirical action measure concentrates at the sampling rate,

$$\mathbb{E}\big[W_1\big(\nu_t^N, \bar{\nu}(\eta_t^N)\big) \,\big|\, \mathcal{F}_t\big] \le c' \rho_N. \tag{33}$$

None of L_r, $L_{\mathcal{R}}$, c' depends on N.

Assumption D.6 (Uniform regularity over the theoretical strategy classes)**.** *The constants L_G, $L_{\mathcal{R}}$, c_t, and c' in D.4, D.5, and hence the finite-horizon constant C_T in F.1, can be chosen independently of $(\pi, \pi_\alpha) \in \Pi_b \times \mathfrak{A}_{\mathrm{adv}}$, uniformly over admissible selection mechanisms $\alpha \in \mathfrak{A}_{\mathrm{sel}}$. The shared Markov prescriptions in these classes have a common finite Lipschitz bound in their continuous local-information and population-law inputs. The class $\mathfrak{A}_{\mathrm{adv}}$ is the regular adversary class covered by the theory. Discontinuous, unrestricted-history, or otherwise out-of-class attacks may be included as empirical stress tests, but the equilibrium and convergence guarantees below do not cover them.*

Assumption D.7 (Oracle-comparison regularity)**.** *Rewards are uniformly bounded. There are constants $L_P, L_{\mathrm{oc}} < \infty$ independent of N and t, such that the learned and oracle next states can be coupled to satisfy*

$$\mathbb{E}[d_{\mathcal{S}}(S_{t+1}^{i,L}, S_{t+1}^{i,O}) \mid \mathcal{F}_t] \le L_P\Big(d_{\mathcal{S}}(S_t^{i,L}, S_t^{i,O}) + |p_t^{i,L} - \theta_i| + W_1(\widehat{\mu}_t^L, \mu_t^O)\Big), \tag{34}$$

$$|r_t^{i,L} - r_t^{i,O}| \le L_{\mathrm{oc}}\Big(d_{\mathcal{S}}(S_t^{i,L}, S_t^{i,O}) + |p_t^{i,L} - \theta_i| + W_1(\widehat{\mu}_t^L, \mu_t^O)\Big). \tag{35}$$

E Proofs of the Detection and Performance Guarantees

E.1 When Can the Learned Type Belief Be Accurate?

Fix a time t and a frozen closed-loop evaluation law. Let U be uniform on $\{1, \dots, N\}$ and independent of the rollout, and write

$$Y_t := \left(h_t^U, \frac{1}{N}\sum_{j=1}^N \varphi_c(h_t^j)\right), \qquad \theta := \theta_U, \qquad p_t := p_t^U = q(Y_t).$$

Then

$$e_t = \mathbb{E}|p_t - \theta|, \qquad \mathcal{L}_t = \mathbb{E}[-\theta \log p_t - (1-\theta)\log(1-p_t)]\,.$$

Let $b_t(y) := \Pr(\theta = 1 \mid Y_t = y)$ be the Bayes posterior based on the detector input. Define the excess detector risk and the type-overlap error by

$$\varepsilon_t := \mathcal{L}_t - H(\theta \mid Y_t), \qquad A_t := \mathrm{Aff}(\mathrm{Law}(Y_t \mid \theta = 0), \mathrm{Law}(Y_t \mid \theta = 1))\,. \tag{36}$$

Here $\mathrm{Aff}(P, Q) := \int \sqrt{dP\, dQ}$ is the Hellinger affinity, $H(\theta \mid Y_t)$ denotes the conditional Shannon entropy (with natural logarithms), namely

$$H(\theta \mid Y_t) := \mathbb{E}\big[-b_t(Y_t)\log b_t(Y_t) - \big(1 - b_t(Y_t)\big)\log\big(1 - b_t(Y_t)\big)\big]\,.$$

Proposition E.1 (Belief accuracy). *Let $p_t^i = q(h_t^i, \zeta_t^N) \in [0,1]$, for every detector q:*

$$\mathbb{E}|p_t - b_t(Y_t)| \le \sqrt{\frac{\varepsilon_t}{2}}, \tag{37}$$

$$e_t \le \mathcal{L}_t = H(\theta \mid Y_t) + \varepsilon_t \le A_t + \varepsilon_t. \tag{38}$$

Consequently, if the two type-conditional detector-input laws become separable, so that $A_t \to 0$, and the detector excess risk satisfies $\varepsilon_t \to 0$, then $e_t \to 0$.

Proof. Conditioning the logarithmic loss on Y_t gives the proper-loss decomposition

$$\mathcal{L}_t = H(\theta \mid Y_t) + \mathbb{E}\,\mathrm{KL}(\mathrm{Ber}(b_t(Y_t)) \,\|\, \mathrm{Ber}(p_t))\,. \tag{39}$$

By the definition of ε_t, the expected KL divergence in equation 39 equals ε_t. Since the total-variation distance between two Bernoulli distributions equals the absolute difference between their parameters, Pinsker's inequality and Jensen's inequality yield

$$\begin{aligned}\mathbb{E}|p_t - b_t(Y_t)| &\le \mathbb{E}\sqrt{\frac{1}{2}\,\mathrm{KL}(\mathrm{Ber}(b_t(Y_t)) \,\|\, \mathrm{Ber}(p_t))} \\ &\le \sqrt{\frac{\varepsilon_t}{2}},\end{aligned}$$

which proves equation 37.

Moreover, pointwise,

$$|p_t - \theta| \le -\theta \log p_t - (1-\theta)\log(1-p_t).$$

Taking expectations gives

$$e_t \le \mathcal{L}_t = H(\theta \mid Y_t) + \varepsilon_t. \tag{40}$$

Let h_2 denote the binary entropy function. Using $h_2(u) \le 2\sqrt{u(1-u)}$ and Bayes' rule,

$$\begin{aligned}H(\theta \mid Y_t) &= \mathbb{E}[h_2(b_t(Y_t))] \\ &\le 2\mathbb{E}\sqrt{b_t(Y_t)\big(1-b_t(Y_t)\big)} \\ &= 2\sqrt{\Pr(\theta=0)\Pr(\theta=1)}\,A_t \\ &\le A_t,\end{aligned}$$

where the last inequality follows from $2\sqrt{x(1-x)} \le 1$ for every $x \in [0,1]$. Combining this bound with equation 40 proves equation 38. □

The overlap A_t bounds the irreducible uncertainty in the available evidence, whereas ε_t is the detector's population excess risk. The latter can be bounded in terms of approximation, generalization, and optimization errors.

Effect of finite training. The excess risk ε_t is defined under the frozen evaluation law. For a fixed input representation and this same sampling law, it tends to zero if the detector class approximates the Bayes posterior, the empirical BCE risk converges uniformly to its population risk, and empirical optimization error vanishes. Together with $A_t \to 0$, Proposition E.1 then gives $e_t \to 0$. These are sufficient conditions, not a convergence guarantee for the implemented detector optimizer. Independent sampling units are complete episodes; agents within an episode are dependent. Changes in the representation or training distribution require additional control.

E.2 Attack-Model Coverage and Uniform Type Identifiability

The above assumption makes explicit the type-identifiability condition for the regular attack class under which the Bayes component of the detector loss contracts. The conditional stability result in Theorem 4.1 remains valid without this assumption whenever the realized errors e_t are supplied directly.

For this appendix only, let U be uniform on $\{1, \dots, N\}$ and independent of the closed loop, and write

$$Y_t := \left(h_t^U, \frac{1}{N} \sum_{j=1}^{N} \varphi_c(h_t^j) \right), \qquad p_t := p_t^U, \qquad \theta := \theta_U. \tag{41}$$

Uniform tagging gives $e_t = \mathbb{E}|p_t - \theta|$ and expresses $\mathcal{L}_t$ as the corresponding tagged-agent detector risk. The exact conditional probability under the evaluated closed-loop law,

$$b_t(Y_t) := \Pr(\theta = 1 \mid Y_t), \tag{42}$$

is introduced only as a proof device; it is not assumed to equal the explicit empirical-Bayes plug-in quantity $b_t^{i,N}$ in equation 5.

E.3 Proof of the Population-Field Stability Bound

Proof. Fix a realization and write $\eta = \eta_t^N$. Reuse the retained-mass notation

$$Z_p := \int (1 - p)\, d\eta, \qquad Z_\theta := \int (1 - \theta)\, d\eta.$$

Assumption D.3 gives $Z_p > 0$, so the learned field is well defined, while the attack budget gives

$$Z_\theta = 1 - \frac{1}{N} \sum_{i=1}^{N} \theta_i \geq 1 - \bar{\beta}. \tag{43}$$

Let f be 1-Lipschitz on $\mathcal{S}$. Since both fields are probability measures, subtracting a constant from f does not change the difference of their integrals; because $\mathcal{S}$ has diameter $D_{\mathcal{S}}$, we may therefore take $|f| \leq D_{\mathcal{S}}$. Adding and subtracting $Z_\theta^{-1} \int (1 - p) f(s)\, d\eta$ yields

$$\begin{aligned}
\left| \int f\, d\widehat{\mu}(\eta) - \int f\, d\mu^{\text{type}}(\eta) \right| &\leq \left| \frac{1}{Z_p} - \frac{1}{Z_\theta} \right| \left| \int (1 - p) f(s)\, d\eta \right| \\
&\quad + \frac{1}{Z_\theta} \left| \int (\theta - p) f(s)\, d\eta \right| \\
&\leq \frac{D_{\mathcal{S}}}{Z_\theta} |Z_p - Z_\theta| + \frac{D_{\mathcal{S}}}{Z_\theta} \int |p - \theta|\, d\eta \\
&\leq \frac{2 D_{\mathcal{S}}}{Z_\theta} \int |p - \theta|\, d\eta.
\end{aligned}$$

The last step uses $|Z_p - Z_\theta| \leq \int |p - \theta|\, d\eta$. By Kantorovich–Rubinstein duality and equation 43, we obtain

$$W_1\big(\widehat{\mu}(\eta_t^N), \mu^{\text{type}}(\eta_t^N)\big) \leq \frac{2 D_{\mathcal{S}}}{1 - \bar{\beta}} \frac{1}{N} \sum_{i=1}^{N} |p_t^i - \theta_i|.$$

Taking expectations and using the definition of e_t proves equation 8. □

E.4 Proof of the Closed-Loop Performance Bound

Proof of Theorem 4.1. Write S_t^i and $S_t^{i,\text{type}}$ for the physical states in the actual system and its perfect-detection reference, and write their stage rewards as r_t^i and $r_t^{i,\text{type}}$. This only relabels the two coupled systems in Assumption D.7. The type-weighted field in the reference system is

$$\mu_t^{N,\text{type}} := \frac{\sum_{i=1}^{N} (1 - \theta_i) \delta_{S_t^{i,\text{type}}}}{\sum_{i=1}^{N} (1 - \theta_i)}. \tag{44}$$

The only error quantity needed in the recursion is the average physical-state distance

$$d_t := \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} d_{\mathcal{S}}(S_t^i, S_t^{i,\text{type}}). \tag{45}$$

First compare the two population fields. Proposition 4.1 applied to the learned-state empirical law gives

$$\mathbb{E}W_1\big(\widehat{\mu}_t^N, \mu^{\text{type}}(\eta_t^N)\big) \le \frac{2D_{\mathcal{S}}}{1-\bar{\beta}} e_t.$$

Coupling agents with equal indices under the common true-type weights gives

$$\begin{aligned} W_1\Big(\mu^{\text{type}}(\eta_t^N), \mu_t^{N,\text{type}}\Big) &\le \frac{\frac{1}{N}\sum_{i=1}^N (1-\theta_i) d_{\mathcal{S}}(S_t^i, S_t^{i,\text{type}})}{\frac{1}{N}\sum_{i=1}^N (1-\theta_i)} \\ &\le \frac{1}{1-\bar{\beta}} \frac{1}{N} \sum_{i=1}^N d_{\mathcal{S}}(S_t^i, S_t^{i,\text{type}}). \end{aligned}$$

The triangle inequality and expectation therefore imply

$$\mathbb{E}W_1\Big(\widehat{\mu}_t^N, \mu_t^{N,\text{type}}\Big) \le \frac{2D_{\mathcal{S}}}{1-\bar{\beta}} e_t + \frac{1}{1-\bar{\beta}} d_t. \tag{46}$$

Averaging the transition inequality in Assumption D.7 over agents, taking expectations, and using equation 46 yields

$$d_{t+1} \le L_P \left[\left(1 + \frac{1}{1-\bar{\beta}}\right) d_t + \left(1 + \frac{2D_{\mathcal{S}}}{1-\bar{\beta}}\right) e_t \right]. \tag{47}$$

The two systems have the same initial physical states, so $d_0 = 0$. Iterating equation 47 gives, for $t \ge 1$,

$$d_t \le L_P \left(1 + \frac{2D_{\mathcal{S}}}{1-\bar{\beta}}\right) \sum_{k=0}^{t-1} \left[L_P \left(1 + \frac{1}{1-\bar{\beta}}\right) \right]^{t-1-k} e_k. \tag{48}$$

The reward inequality in Assumption D.7, averaged over agents and combined again with equation 46, gives

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}|r_t^i - r_t^{i,\text{type}}| \le L_{\text{oc}} \left[\left(1 + \frac{1}{1-\bar{\beta}}\right) d_t + \left(1 + \frac{2D_{\mathcal{S}}}{1-\bar{\beta}}\right) e_t \right]. \tag{49}$$

By the definition of the two returns and the triangle inequality, their gap is at most the discounted sum of the left-hand side of equation 49. Substituting equation 48, the contribution of an error e_k to rewards after time k is bounded by

$$\begin{aligned} &L_{\text{oc}} \left(1 + \frac{2D_{\mathcal{S}}}{1-\bar{\beta}}\right) \sum_{t=k+1}^{T} \gamma^t L_P \left(1 + \frac{1}{1-\bar{\beta}}\right) \left[L_P \left(1 + \frac{1}{1-\bar{\beta}}\right) \right]^{t-1-k} e_k \\ &\quad = L_{\text{oc}} \left(1 + \frac{2D_{\mathcal{S}}}{1-\bar{\beta}}\right) \gamma^k e_k \sum_{j=1}^{T-k} \left[\gamma L_P \left(1 + \frac{1}{1-\bar{\beta}}\right) \right]^j. \end{aligned}$$

The direct effect of e_k on the reward at time k supplies the $j = 0$ term. Extending each nonnegative geometric sum from $T-k$ to T, summing over k, and using equation 13 now give

$$\Big| J_{N,T}(\pi, \pi_\alpha) - J_{N,T}^{\text{type}}(\pi, \pi_\alpha) \Big| \le K_T \sum_{k=0}^{T} \gamma^k e_k.$$

Finally, Proposition E.1 gives $e_k \le \mathcal{L}_k$ for every k, which proves the second inequality in equation 12. □

E.5 Oracle Information, Value Ordering, and Equilibrium

Proof of Proposition 4.2. Because $\Pi_0 \subseteq \Pi_b \subseteq \Pi_\theta$, monotonicity of the supremum in the already-defined robust objective equation 2 immediately gives equation 14. Whenever a saddle point exists in one of these classes, Definition 3.1 identifies its payoff with $V_{N,T}^{\text{wc}}(\Pi)$ for that class. □

F EXTEND TO INFINITY AGENT LIMIT

First, we have the proposition illustrate that as the number of agents $N \to \infty$, the empirical Bayes method for estimating the attack proportion based on full agent trajectories converges to a fixed global attack rate.

Proposition F.1 (Population limit of the empirical-Bayes belief)**.** *Fix t and a flow of population measures $m_{0:t} := (\mu_s, \nu_s)_{s=0}^{t}$, held fixed as in the standard two-step formulation of mean-field equilibria: individual quantities are first evaluated against a given flow, which is then required to be consistent with the one the prescriptions generate. Suppose the normalized objective in equation 4 converges uniformly in probability to*

$$\mathcal{M}_t(u; m_{0:t}) := \int \log[(1-u)L_{0,t}(h; m_{0:t}) + uL_{1,t}(h; m_{0:t})]\, \chi_t(dh), \tag{50}$$

and that $\mathcal{M}_t$ has a unique maximizer $\beta_t^\star$. Then $\widehat{\beta}_t^N \to \beta_t^\star$ in probability. For every tagged history at which the denominator below is positive,

$$b_t^\infty(h; m_{0:t}) = \frac{\beta_t^\star L_{1,t}(h; m_{0:t})}{(1-\beta_t^\star)L_{0,t}(h; m_{0:t}) + \beta_t^\star L_{1,t}(h; m_{0:t})}. \tag{51}$$

If the maximizer is not unique, the attack fraction is not identifiable from the available population traces and no unique plug-in belief exists.

Proof. Uniform convergence and uniqueness give consistency of the argmax in equation 4. Substitution into equation 5 and continuity of the displayed ratio prove equation 51. □

When $L_{0,t}$ and $L_{1,t}$ are the correct type-conditional memory laws under $m_{0:t}$ and the two-component mixture is identifiable, $\mathcal{M}_t(u)$ is the expected log-likelihood of a weight-u mixture under the true marginal, so the information inequality makes $u = \beta$ its unique maximizer and $\beta_t^\star = \beta$. Under misspecification $\beta_t^\star$ is the pseudo-true value that the estimator still tracks, and equation 51 is the corresponding projected posterior. Next we illstrate the infinite agent type equilibrium.

F.1 EX-INTERIM ROBUST MEAN-FIELD BAYESIAN EQUILIBRIUM

Centralized training may use the type labels through the detector loss $N^{-1}\sum_i \mathrm{BCE}(p_t^i, \theta_i)$, whereas the deployed policy never observes those labels. The concrete training workflow, deployed information interface, ablations, and rollout algorithms are collected in Appendices A and B.

We consider a fixed selection mechanism and its initial augmented population law η_0. The recursion below optimizes the subsequent action policies; it does not optimize selection through η_0. Let

$$\mathcal{T}(\eta, \pi_t, \pi_{\alpha,t}) := \int Q^{\pi_t, \pi_{\alpha,t}}(\cdot \mid x, \eta)\, \eta(dx) \tag{52}$$

be the one-step population flow induced by the current policies, where $Q^{\pi_t, \pi_{\alpha,t}}(\cdot \mid x, \eta)$ is the transition kernel of a representative agent's augmented state, conditional on its current augmented state x and the current population law η, under the cooperative and adversarial prescriptions $(\pi_t, \pi_{\alpha,t})$. Integrating this individual transition kernel against η gives the next population law. For a population law η and current prescriptions $(\pi_t, \pi_{\alpha,t})$, let

$$\mathcal{R}(\eta; \pi_t, \pi_{\alpha,t}) := \mathbb{E}\big[r\big(S, A, \mu(\eta), \mathcal{L}(A)\big)\big] \tag{53}$$

be the one-step mean-field team reward, where $X \sim \eta$, S and Θ are its physical-state and latent agent type, the action A is generated by π_t when $\Theta = 0$ and by $\pi_{\alpha,t}$ when $\Theta = 1$ (including the observation kernel), and $\mathcal{L}(A)$ is the resulting population action law. Let $V_t^{\mathrm{mf}}(\eta)$ denote the Bellman–Isaacs continuation value for the conditional mean-field game. Its identification with a robust rollout value requires the same admissible information and prescription classes in the two formulations.

When the joint state–memory–belief law is a sufficient common-information state, the finite-horizon mean-field control game has the Bellman–Isaacs recursion

$$V_{T+1}^{\mathrm{mf}}(\eta) := 0, \tag{54}$$

$$V_t^{\mathrm{mf}}(\eta) := \sup_{\pi_t \in \Pi_b} \inf_{\pi_{\alpha,t} \in \mathfrak{A}_{\mathrm{adv}}} \left\{ \mathcal{R}(\eta; \pi_t, \pi_{\alpha,t}) + \gamma V_{t+1}^{\mathrm{mf}}\big(\mathcal{T}(\eta, \pi_t, \pi_{\alpha,t})\big) \right\}. \tag{55}$$

Here the actions of the distribution-space game are shared decentralized prescriptions, rather than individual physical actions.

Definition F.1 (Robust mean-field Bayesian equilibrium)**.** *Markov prescription sequences $(\pi_t^\star, \pi_{\alpha,t}^\star)_{t=0}^T$ together with their induced flow $(\eta_t^\star)_{t=0}^T$ form a robust belief-augmented mean-field equilibrium if, for every $t \le T$, every admissible population law η, and every pair of admissible prescriptions $(\pi_t, \pi_{\alpha,t})$,*

$$\mathcal{R}(\eta; \pi_t, \pi_{\alpha,t}^\star) + \gamma V_{t+1}^{\mathrm{mf}}\big(\mathcal{T}(\eta, \pi_t, \pi_{\alpha,t}^\star)\big) \le V_t^{\mathrm{mf}}(\eta) \le \mathcal{R}(\eta; \pi_t^\star, \pi_{\alpha,t}) + \gamma V_{t+1}^{\mathrm{mf}}\big(\mathcal{T}(\eta, \pi_t^\star, \pi_{\alpha,t})\big), \tag{56}$$

$$\eta_{t+1}^\star = \mathcal{T}(\eta_t^\star, \pi_t^\star, \pi_{\alpha,t}^\star), \qquad t < T. \tag{57}$$

Here $\pi_t^\star$ and $\pi_{\alpha,t}^\star$ are Markov prescriptions; their dependence on the current population law η is suppressed in the notation.

Proposition F.2 (Conditional equilibrium existence)**.** *Suppose the augmented state space and the relaxed cooperative and adversarial prescription spaces are compact, the two prescription spaces are convex, and the stage reward and flow operator are continuous. Suppose further that, at each stage of 54, the Hamiltonian is concave in the cooperative prescription and convex in the adversarial prescription. If the augmented law is a sufficient common-information state, then a finite-horizon equilibrium in F.1 exists within the Markov prescription class.*

Proof. At T, compactness, continuity, and the concave-convex structure give a saddle prescription by the minimax theorem. If V_{t+1}^{mf} is continuous, continuity of $\mathcal{R}$ and $\mathcal{T}$ makes the stage Hamiltonian continuous; the maximum theorem preserves continuity of V_t^{mf}, and a measurable-selection theorem provides Markov saddle prescriptions. Backward induction gives such prescriptions at every stage. Forward propagation under the selected prescriptions gives equation 57. □

Existence of a finite-horizon equilibrium does not by itself prove existence, uniqueness, or convergence to this stationary fixed point. For a stationary infinite-horizon population law, the additional condition is

$$\eta^\star = \mathcal{T}(\eta^\star, \pi^\star, \pi_\alpha^\star). \tag{58}$$

Corollary F.0.1 (Infinite-horizon performance stability)**.** *Under Assumptions D.3 and D.7, define*

$$E_\infty := \sum_{t=0}^{\infty} \gamma^t e_t, \quad c_0 := \frac{2D_{\mathcal{S}}}{1-\bar{\beta}}, \quad q_0 := 1 + \frac{1}{1-\bar{\beta}}.$$

If

$$\gamma L_P q_0 = \gamma L_P \left(1 + \frac{1}{\underline{z}_\theta}\right) < 1, \tag{59}$$

then

$$\left| J_N^L - J_N^{\mathrm{type}} \right| \le K_\infty E_\infty, \tag{60}$$

where

$$K_\infty := \frac{L_{\mathrm{oc}}(1+c_0)}{1-\gamma L_P q_0}. \tag{61}$$

Consequently,

$$\left| J_N^L - J_N^{\mathrm{type}} \right| \le K_\infty \sum_{t=0}^{\infty} \gamma^t \mathcal{L}_t.$$

If $e_t \le \bar{e}$ for all t, then

$$\left| J_N^L - J_N^{\mathrm{type}} \right| \le \frac{K_\infty}{1-\gamma} \bar{e}.$$

Proof. For every T, Theorem 4.1 gives

$$\left|J_{N,T}^{L} - J_{N,T}^{\text{type}}\right| \le K_T \sum_{t=0}^{T} \gamma^t e_t.$$

Under equation 59, $K_T \uparrow K_\infty$. Since rewards are uniformly bounded, $J_{N,T}^{L} \to J_N^{L}$ and $J_{N,T}^{\text{type}} \to J_N^{\text{type}}$. Letting $T \to \infty$ proves the first bound. The remaining statements follow from $e_t \le \mathcal{L}_t$ and $\sum_{t\ge 0} \gamma^t e_t \le \bar{e}/(1-\gamma)$. □

F.2 Mean-Field Consistency of Belief-Augmented Control

The following analysis relies on standard conditions: empirically consistent, shared equivariance, Lipschitz dynamics and empirical convergence, regular rewards and action sampling, and strategy-uniform regularity. Precise mathematical statements, sufficient primitive conditions, and scope limitations are collected in Appendix D.

Fix the detector and selection method throughout this analysis. The finite and limiting systems use the same admissible policy maps and information interfaces. For each policy pair, define

$$J_T^{\text{mf}}(\pi, \pi_\alpha) := \sum_{t=0}^{T} \gamma^t \mathcal{R}(\eta_t; \pi_t, \pi_{\alpha,t}), \qquad \eta_{t+1} = \mathcal{T}(\eta_t, \pi_t, \pi_{\alpha,t}). \tag{62}$$

The flow is generated anew for each policy pair, including unilateral deviations. For trained networks, the bounds are applied conditionally on the frozen parameters using independent evaluation rollouts. If different networks are used at different population sizes, the regularity constants must remain uniform for the stated rate to hold.

Lemma F.1 (Equivariance of target-wise belief)**.** *The closed-loop update is permutation equivariant: relabeling agents by any permutation σ relabels the posterior vector in the same way,*

$$(p_t^1, \dots, p_t^N) \mapsto (p_t^{\sigma(1)}, \dots, p_t^{\sigma(N)}).$$

Consequently, η_t^N and $\widehat{\mu}(\eta_t^N)$ are invariant to agent labels. The law of the closed loop is also invariant.

Proof. At $t = 0$, every target is processed by the same map q, so relabeling inputs relabels outputs. Suppose the claim holds at time t. Empirical summaries such as $\widehat{\mu}(\eta_t^N)$ are unchanged by the permutation. Since the same detector update, posterior head, and within-type policies are applied to every relabeled local trace, the outputs at $t + 1$ are relabeled by σ as well. Induction proves equivariance. Empirical measures are sums over indices and are therefore invariant. Finally, D.1 and the iid noises are invariant in law under relabeling; applying the equivariant update preserves this invariance at every time. □

Lemma F.2 (Stability of the belief-weighted field)**.** *If $Z_p(\eta), Z_p(\eta') \ge \underline{z}_p > 0$, then*

$$W_1(\widehat{\mu}(\eta), \widehat{\mu}(\eta')) \le \frac{1 + 2D_{\mathcal{S}}}{\underline{z}_p} W_1(\eta, \eta'). \tag{63}$$

Proof. Write $Z = Z_p(\eta)$ and $Z' = Z_p(\eta')$. Coupling the two laws in the augmented sum metric gives $|Z - Z'| \le W_1(\eta, \eta')$. For any 1-Lipschitz f on $\mathcal{S}$, subtract a constant so that $|f| \le D_{\mathcal{S}}$. With $A = \int (1-p) f(s)\, d\eta$ and $A' = \int (1-p') f(s')\, d\eta'$, the same coupling gives

$$|A - A'| \le (1 + D_{\mathcal{S}}) W_1(\eta, \eta'), \qquad |A'| \le D_{\mathcal{S}} Z'.$$

Consequently,

$$\left|\frac{A}{Z} - \frac{A'}{Z'}\right| \le \frac{|A - A'|}{Z} + \frac{|A'||Z - Z'|}{ZZ'} \le \frac{1 + 2D_{\mathcal{S}}}{\underline{z}_p} W_1(\eta, \eta').$$

Taking the supremum over f proves the claim. □

Define the nonlinear mean-field flow by

$$\eta_{t+1}(B) := \int_{\mathcal{X}} Q(B \mid x, \eta_t)\, \eta_t(dx), \qquad B \subseteq \mathcal{X}. \tag{64}$$

Theorem F.1 (Finite-Population Convergence and Accumulation of Errors)**.** *Under empirical consistency, shared equivariance, Lipschitz dynamics and empirical convergence, let η_0 be the population law in Assumption D.1, and let $(\eta_t)_{t=0}^T$ be the limiting population flow generated by equation 64. Then*

$$\mathbb{E}W_1(\eta_{t+1}^N, \eta_{t+1}) \le 2L_Q\, \mathbb{E}W_1(\eta_t^N, \eta_t) + c_t\rho_N. \tag{65}$$

Consequently, for every finite horizon T, there is a constant $C_T < \infty$, independent of N, such that

$$\max_{0\le t\le T} \mathbb{E}W_1(\eta_t^N, \eta_t) \le C_T\rho_N. \tag{66}$$

Since $\rho_N \to 0$,

$$\lim_{N\to\infty} \max_{0\le t\le T} \mathbb{E}W_1(\eta_t^N, \eta_t) = 0.$$

For a finite augmented state space, one may take $\rho_N = N^{-1/2}$.

Proof. By the triangle inequality and equation 64,

$$\mathbb{E}W_1(\eta_{t+1}^N, \eta_{t+1}) \le \mathbb{E}W_1\left(\eta_{t+1}^N, \int Q(\cdot \mid x, \eta_t^N)\, \eta_t^N(dx)\right) + \mathbb{E}W_1\left(\int Q(\cdot \mid x, \eta_t^N)\, \eta_t^N(dx), \int Q(\cdot \mid x, \eta_t)\, \eta_t(dx)\right).$$

The first term is at most $c_t\rho_N$ by equation 30. Coupling the two integrals through an optimal coupling of η_t^N and η_t, and then applying equation 29, bounds the second term by

$$2L_Q\, \mathbb{E}W_1(\eta_t^N, \eta_t).$$

This proves equation 65. Iterating the recursion from the initial bound in Assumption D.1 gives equation 66. □

Corollary F.1.1 (Finite-population scaling of team performance)**.** *Under D.1D.2D.4D.5,D.6, for any finite T, let*

$$\Delta_{N,T} := (L_r c' + L_{\mathcal{R}} C_T)\rho_N \sum_{t=0}^{T} \gamma^t, \tag{67}$$

$$\sup_{\substack{\pi\in\Pi_b \\ \pi_\alpha\in\mathfrak{A}_{\mathrm{adv}}}} \left|J_{N,T}(\pi, \pi_\alpha) - J_T^{\mathrm{mf}}(\pi, \pi_\alpha)\right| \le \Delta_{N,T}. \tag{68}$$

Proof. Write $R_t^N := N^{-1}\sum_i r(s_t^i, a_t^i, \mu_t^N, \nu_t^N)$, so that $J_{N,T} = \mathbb{E}\sum_{t=0}^T \gamma^t R_t^N$. Conditionally on $\mathcal{F}_t$, replacing ν_t^N by $\bar\nu(\eta_t^N)$ in each summand changes R_t^N by at most $L_r W_1(\nu_t^N, \bar\nu(\eta_t^N))$, whose conditional expectation equation 33 bounds by $L_r c'\rho_N$. After the replacement, conditional independence of the actions gives

$$\mathbb{E}\left[\frac{1}{N}\sum_i r\big(s_t^i, a_t^i, \mu_t^N, \bar\nu(\eta_t^N)\big) \,\middle|\, \mathcal{F}_t\right] = \frac{1}{N}\sum_i \int r\big(s_t^i, a, \mu_t^N, \bar\nu(\eta_t^N)\big) K_{\theta_i}(da \mid x_t^i, \eta_t^N) = \mathcal{R}(\eta_t^N),$$

the definition equation 32 evaluated at the empirical measure. Hence $|\mathbb{E}R_t^N - \mathbb{E}\,\mathcal{R}(\eta_t^N)| \le L_r c'\rho_N$. The Lipschitz property of $\mathcal{R}$ and F.1 give $|\mathbb{E}\,\mathcal{R}(\eta_t^N) - \mathcal{R}(\eta_t)| \le L_{\mathcal{R}} C_T \rho_N$. Combining the two bounds and summing with weights γ^t gives the stated estimate for each policy pair. The constants are uniform over $\Pi_b \times \mathfrak{A}_{\mathrm{adv}}$ by D.6, so taking the supremum proves equation 68. □

Theorem F.2 (Finite-population robust-value convergence)**.** *Under the assumptions of Corollary F.1.1, with the fixed selection mechanism and common policy classes above,*

$$\left|\sup_{\pi\in\Pi_b}\inf_{\pi_\alpha\in\mathfrak{A}_{\mathrm{adv}}} J_{N,T}(\pi, \pi_\alpha) - \sup_{\pi\in\Pi_b}\inf_{\pi_\alpha\in\mathfrak{A}_{\mathrm{adv}}} J_T^{\mathrm{mf}}(\pi, \pi_\alpha)\right| \le \Delta_{N,T}. \tag{69}$$

The rate is $O(\rho_N)$ for fixed T, and $O(N^{-1/2})$ for a finite augmented state space. When the Bellman–Isaacs formulation uses the same admissible information and prescription classes and its dynamic programming conditions hold, the mean-field lower value above equals $V_0^{\mathrm{mf}}(\eta_0)$.

Proof. Equation equation 68 gives, uniformly over the common policy classes,

$$J_T^{\mathrm{mf}}(\pi, \pi_\alpha) - \Delta_{N,T} \le J_{N,T}(\pi, \pi_\alpha) \le J_T^{\mathrm{mf}}(\pi, \pi_\alpha) + \Delta_{N,T}.$$

Taking $\inf_{\pi_\alpha}$ and then $\sup_\pi$ preserves the inequalities and proves equation 69. This argument does not require that either optimum is attained. Under the stated dynamic programming conditions, the same mean-field optimization is represented by equation 54. □

The payoff approximation also transfers candidate policies that solve the mean-field game approximately.

Corollary F.2.1 (Approximate mean-field saddles in finite populations)**.** *Under the assumptions of Corollary F.1.1, fix the detector, selection mechanism, and corresponding initial law. Let (π, π_α) be a candidate pair in the common policy classes. Suppose that, for every admissible π' and π'_α,*

$$\begin{aligned} J_T^{\mathrm{mf}}(\pi', \pi_\alpha) &\le J_T^{\mathrm{mf}}(\pi, \pi_\alpha) + \varepsilon_{\mathrm{opt}}, \\ J_T^{\mathrm{mf}}(\pi, \pi'_\alpha) &\ge J_T^{\mathrm{mf}}(\pi, \pi_\alpha) - \varepsilon_{\mathrm{opt}}, \end{aligned} \tag{70}$$

where $\varepsilon_{\mathrm{opt}} \ge 0$ bounds each player's unilateral improvement. Then

$$\begin{aligned} J_{N,T}(\pi', \pi_\alpha) &\le J_{N,T}(\pi, \pi_\alpha) + \varepsilon_{\mathrm{opt}} + 2\Delta_{N,T}, \\ J_{N,T}(\pi, \pi'_\alpha) &\ge J_{N,T}(\pi, \pi_\alpha) - \varepsilon_{\mathrm{opt}} - 2\Delta_{N,T}. \end{aligned} \tag{71}$$

Thus the pair is an $(\varepsilon_{\mathrm{opt}} + 2\Delta_{N,T})$-saddle of the finite game.

Proof. For a defender deviation, the uniform payoff bound and equation 70 give

$$\begin{aligned} J_{N,T}(\pi', \pi_\alpha) &\le J_T^{\mathrm{mf}}(\pi', \pi_\alpha) + \Delta_{N,T} \\ &\le J_T^{\mathrm{mf}}(\pi, \pi_\alpha) + \varepsilon_{\mathrm{opt}} + \Delta_{N,T} \\ &\le J_{N,T}(\pi, \pi_\alpha) + \varepsilon_{\mathrm{opt}} + 2\Delta_{N,T}. \end{aligned}$$

For an adversary deviation, the same argument yields

$$\begin{aligned} J_{N,T}(\pi, \pi'_\alpha) &\ge J_T^{\mathrm{mf}}(\pi, \pi'_\alpha) - \Delta_{N,T} \\ &\ge J_T^{\mathrm{mf}}(\pi, \pi_\alpha) - \varepsilon_{\mathrm{opt}} - \Delta_{N,T} \\ &\ge J_{N,T}(\pi, \pi_\alpha) - \varepsilon_{\mathrm{opt}} - 2\Delta_{N,T}. \end{aligned}$$

Each comparison uses the flow induced by its own policy pair. The shared prescriptions are admissible at both population scales, so both inequalities hold over the same unilateral deviations. □

Setting $\varepsilon_{\mathrm{opt}} = 0$ recovers the exact-saddle transfer. For learned policies, $\varepsilon_{\mathrm{opt}}$ describes their residual strategic error in the specified game; the corollary does not bound it as a function of the number of PPO updates. This residual is distinct from the detector excess risk ε_t in equation 36.

G Additional Experimental Details

G.1 Environment Details

Battle. Our main environment is Battle (Zheng et al., 2018), the standard testbed for mean-field MARL (Yang et al., 2018): two armies of N homogeneous agents on an $L \times L$ grid, each agent choosing among 21 discrete actions (moves and directional attacks) from a 13×13 egocentric window, so interactions are local, anonymous and partially observed. We evaluate at $N \in \{32, 64, 144\}$ per army with $L = 40, 40, 60$ respectively; the $64 \to 144$ pair holds the agent density N/L^2 fixed and therefore isolates population size, while $N = 32$ shares a map with $N = 64$ and probes the small-population regime in which a mean-field approximation is least favourable. The opposing army is a fixed pre-trained policy throughout, so that changes in performance are attributable to the Byzantine agents inside our own army rather than to opponent adaptation.

PowerGrid. PowerGrid (Chen et al., 2022) is a secondary voltage-control task on a microgrid with $N \in \{20, 40\}$ inverter-interfaced distributed energy resources (DERs). Each agent receives a nine-dimensional local observation and selects one of ten discrete voltage-reference set-points. The simulator integrates the coupled electrical dynamics between control steps, and the reward encourages voltage regulation within $[0.95, 1.05]$ p.u. Byzantine agents disrupt coordination through adversarial choices of their DER control actions. It tests transfer to continuous control under physical rather than spatial coupling, and carries a concrete threat model: a compromised DER, inverter or local controller issues malicious set-points that push neighbouring buses outside the band, eroding damping margins and, in the worst case, triggering cascading protection trips. We report episode return and the fraction of steps at which all bus voltages lie in band.

Methods we could not run. Robust MARL methods that train an explicit adversary, such as EIR-MAPPO, enumerate or approximate worst-case scenarios whose number grows with the population, a limitation Li et al. (2025) identify as the motivation for adversary-free regularisation. We measured the wall-clock cost of EIR-MAPPO at $|\mathcal{N}| \in \{16, 32, 64\}$ and extrapolate to the populations we study in Table 3. Its throughput falls in inverse proportion to the population, so a single run at $|\mathcal{N}| = 144$ is on the order of 60 GPU-days against 0.6 for BeMF; at $|\mathcal{N}| \geq 128$ it additionally exceeded the memory of an 80 GB device in our setup. We therefore report the measured scaling in place of scores.

	env. steps / second		GPU-days per run		
$\lvert\mathcal{N}\rvert$	BeMF	EIR-MAPPO	BeMF	EIR-MAPPO	ratio
16	1809	59.5	0.26	7.8	30×
32	1433	31.7	0.32	14.6	45×
64	1170	15.9	0.40	29.2	74×
128	835	8.0*	0.55	58.2*	—
144	736	7.1*	0.63	65.4*	—
182	652	5.6*	0.71	82.6*	—

Table 3: Cost of a single training run against controlled-population size $|\mathcal{N}|$, on one A800-80GB with $T = 400$ and $E = 24$ parallel environment copies. Throughput is reported in environment steps per second, which is independent of E and therefore converts to any step budget; GPU-days assume the 40M-step budget of our Battle configuration. EIR-MAPPO's throughput falls in inverse proportion to the population (fitted log-log slope -1.00) against -0.70 for BeMF. Entries marked * are extrapolated along that fit because the run did not complete: beyond $|\mathcal{N}| = 64$ EIR-MAPPO exhausted the device during its actor update, an implementation limit discussed in Section A.3. The extrapolations therefore bound the cost of a run that did not fit, and we use them only to justify omitting the baseline at scale.

G.2 Ablation Study Additional result

Agent Num	Adv Num	w/o Belief	BeMF
32	4	-334.57 ± 120.18	$\mathbf{-160.19} \pm 11.55$
	8	-496.94 ± 45.14	$\mathbf{-230.35} \pm 18.26$
	16	-824.18 ± 64.84	$\mathbf{-270.24} \pm 12.95$
64	8	-783.12 ± 137.33	$\mathbf{-347.68} \pm 9.72$
	16	-1292.71 ± 71.58	$\mathbf{-1057.91} \pm 14.43$
	32	-1403.20 ± 94.26	$\mathbf{-935.72} \pm 19.62$
144	16	-2450.65 ± 122.52	$\mathbf{-783.23} \pm 19.79$
	32	-2739.41 ± 122.58	$\mathbf{-1450.11} \pm 21.03$
	36	-2809.25 ± 74.69	$\mathbf{-1593.89} \pm 22.50$
	48	-2569.90 ± 116.71	$\mathbf{-2162.20} \pm 53.74$
	72	-2905.39 ± 101.47	$\mathbf{-2555.94} \pm 26.53$

Table 4: Ablation of the belief component on Battle. Results are team returns (mean $\pm$ standard deviation; higher is better), with the best mean in each configuration highlighted in bold.

G.3 Hyperparameters

Stealth level λ	0.0	0.1	0.2	0.3	0.4	0.5	0.8	0.9	0.95
ROC-AUC	0.999	0.997	0.979	0.894	0.760	0.642	0.501	0.492	0.491
Score, Byzantine	0.956	0.944	0.872	0.684	0.479	0.330	0.219	0.215	0.214
Score, honest	0.005	0.013	0.032	0.085	0.145	0.189	0.219	0.215	0.215
Recall @ 0.5	0.951	0.936	0.865	0.662	0.420	0.200	0.001	0.000	0.000

Table 5: Detector quality versus attacker stealth on Battle ($N = 64$, $\beta N = 16$, so a score of 0.25 is uninformative). Scores are the detector's mean Byzantine probability for each group. AUC is threshold-free; recall is at the deployed threshold 0.5.

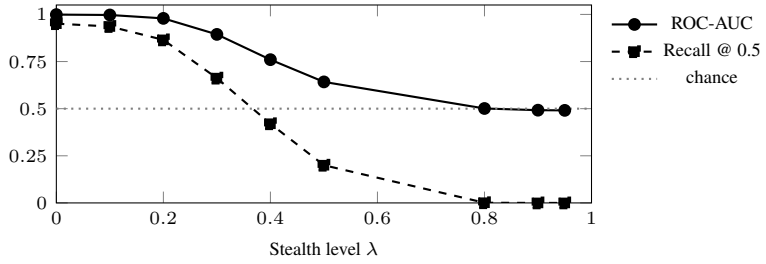


Figure 4: Detector quality versus attacker stealth (Battle, $N = 64$, $\beta N = 16$).

Hyperparameter	Value	Hyperparameter	Value
Training and Optimization			
Env. steps	4×10^7	Episode length	400
Rollout threads	64	PPO epochs	5
Critic epochs	5	Actor mini-batches	8
Critic mini-batches	4	Discount γ	0.99
GAE λ	0.95	PPO clip ratio	0.2
Entropy coef.	0.01	Value loss coef.	1.0
Max grad norm	10.0	Huber threshold	10.0
Actor LR	5×10^{-4}	Critic LR	5×10^{-4}
Belief LR	5×10^{-4}	Adversary LR	5×10^{-4}
Optimizer ϵ	10^{-5}	Weight decay	0
Network and BeMF			
Hidden dimensions	$[128, 128, 128]$	Activation	ReLU
Initialization	Orthogonal	Parameter sharing	Yes
Recurrent belief	Yes	Belief recurrent layers	1
BeMF hidden dim.	128	Memory embed. dim.	64
State embed. dim.	64	Detector epochs	1
Detector mini-batches	8	Hard-belief threshold	0.5
Central belief agg.	Mean	Teacher forcing	Yes
Adversarial Training and Environment			
Action aggregation	Product	Adversarial prob.	0.5
Victim interval	3	Adversary epochs	1
Adv. entropy coef.	0.01	Map size	40×40
View range	10	Attack range	2
Agent HP	10	Attack damage	2
Kill reward	5		

Table 6: Hyperparameters used for BeMF in the Battle environment.

Hyperparameter	Value	Hyperparameter	Value
Training and Optimization			
Env. steps	1×10^6	Episode length	20
Rollout threads	32	Eval. episodes	20
PPO epochs	5	Critic epochs	5
Actor mini-batches	2	Critic mini-batches	2
Discount γ	0.99	GAE λ	0.95
PPO clip ratio	0.2	Entropy coef.	0.01
Value loss coef.	1.0	Max grad norm	10.0
Huber threshold	10.0	Optimizer ϵ	10^{-5}
Actor LR	5×10^{-4}	Critic LR	5×10^{-4}
Belief LR	5×10^{-4}	Adversary LR	5×10^{-4}
Weight decay	0	Parameter sharing	Yes
Network and BeMF			
Hidden dimensions	$[128, 128, 128]$	Activation	ReLU
Initialization	Orthogonal	Recurrent belief	Yes
Belief recurrent layers	1	BeMF hidden dim.	128
Memory embed. dim.	64	State embed. dim.	64
Detector epochs	1	Detector mini-batches	2
Hard-belief threshold	0.5	Central belief agg.	Separate
Teacher forcing	Yes	Action aggregation	Product
Adversarial Training and Environment			
Adversarial prob.	0.5	Victim interval	3
Adversary epochs	1	Adv. entropy coef.	0.01
Number of DERs	40	Number of agents	40
Episode length	20	State type	FP

Table 7: Hyperparameters used for BeMF in the PowerGrid environment.