Robust Coordination under Misaligned Communication via Power Regularization

Nancirose Piazza, Amirhossein Karimi†, Behnia Soleymani†, Vahid Behzadan, Stefan Sarkadi

{npiazza,akari9}@newhaven.edu, ibehnia.s@gmail.com, vbehzadan@newhaven.edu, stefan.sarkadi@kcl.ac.uk

† Equal Contribution

Abstract

Effective communication in Multi-Agent Reinforcement Learning (MARL) can significantly enhance coordination and collaborative performance in complex and partially observable environments. However, reliance on communication can also introduce vulnerabilities when agents are misaligned, potentially leading to adversarial interactions that exploit implicit assumptions of cooperative intent. Prior work has addressed adversarial behavior through power regularization by controlling the influence one agent exerts over another, but has largely overlooked the role of communication in these dynamics. This paper introduces communicative power regularization (CPR), which extends power regularization specifically to communication channels. By explicitly quantifying and constraining agents' communicative influence during training, CPR actively mitigates vulnerabilities arising from misaligned or adversarial communications. Evaluations in the Grid Coverage benchmark environment demonstrate that our approach significantly enhances robustness to adversarial communication while preserving cooperative performance, offering a practical framework for secure and resilient cooperative MARL systems.

1 Introduction

Effective coordination among agents in Multi-Agent Reinforcement Learning (MARL) is crucial for achieving collective goals. Communication, as an explicit exchange of information, is often employed to facilitate this coordination, particularly in Cooperative MARL (CoMARL), where agents collaborate. However, common CoMARL approaches emphasizing parameter sharing for training efficiency can lead to joint policies vulnerable to issues such as agent free-riding (Ueshima et al., 2023) or over-reliance on learned conventions (Köster et al., 2020). Such vulnerabilities are exacerbated when agents are misaligned or face adversarial interactions, especially if policies implicitly assume cooperative intent.

Objective misalignment, where agents pursue self-interested goals, makes public communication channels susceptible to sabotage, particularly against cooperative agents. Resilience against such misalignment is critical for deploying autonomous agent teams, requiring evaluation under non-standard conditions, considering both team and individual contexts.

Communication remains a key research area in MARL (OroojlooyJadid & Hajinezhad, 2021), often modeled with protocol controllers like CommNet (Sukhbaatar et al., 2016) and IC3Net (Singh et al., 2018). When agents learn communication and environment policies concurrently, they may develop uncontrolled regularization against misaligned messages, whether from co-learning errors or intentional adversarial actions. Such self-learned communication, however, can create vulnerabilities if

naive agents are targeted. We define misaligned communication as any message negatively affecting a recipient's performance, irrespective of explicit adversarial intent. Fostering resilience requires policies robust enough for mixed settings, differing from adversarial attacks that inject malicious payloads (Tu et al., 2021; Dong et al., 2022).

Power, one agent's influence over another's utility and decisions, offers a mechanism to enhance policy robustness when incorporated into training. While agents typically do not explicitly optimize for power, decomposing utility functions, akin to intrinsic rewards (Du et al., 2019), could offer greater control. This paper introduces Communicative Power Regularization (CPR), extending the concept of power regularization (Li & Dennis, 2024) specifically to communication channels. By quantifying and constraining communicative influence, CPR mitigates vulnerabilities from misaligned communication.

Our contributions are: (1) We propose CPR as a technique to control power dynamics within learned communication policies. (2) We evaluate CPR in the Grid Coverage benchmark, demonstrating significantly enhanced robustness to adversarial communication while preserving cooperative performance. CPR thus offers a practical framework for more secure and resilient cooperative MARL.

This paper is organized as follows: Section 2 reviews related work. Section 3 covers preliminaries. Section 4 details our CPR approach. Section 5 presents experimental results, followed by conclusions in Section 6.

2 Related Work

Adversarial communication in MARL settings is often highlighted by its emergence in non-cooperative settings (Blumenkamp & Prorok, 2020), which may be the product of misaligned agents. Adversarial attacks in MARL settings are diverse in their methodology, ranging from sparse targeted attacks (Hu & Zhang, 2022) to attacks that exploit vulnerabilities in mechanism design, such as consensus-based mechanisms (Figura et al., 2021) and adversarial minority influence (Li et al., 2024).

Adversarial training, an approach to mitigating against adversarial interests, is an umbrella-term for incorporating adversarial interactions into training for hardening and better resilience against adversarial opponents. In support, there are works on the robustness of CoMARL, such as Lin et al. (2020) and Guo et al. (2022). There are diverse defenses against adversarial communication, including works that consider test-time settings with theory of mind inspired mechanisms (Piazza & Behzadan, 2023). In our work, adversarial training is used to address misaligned communication.

Many CoMARL works that address credit assignment between global reward and local reward can be viewed as a means for regularizing agent behaviors and dynamics. For example, a reward-shaping mechanism was proposed by Ibrahim et al. (2020) to portion out the team reward based on individual contributions in order to address free-riders. Foerster et al. (2018) proposed COMA, Counterfactual Multi Agent Policy Gradients, which marginalizes out single agent actions and also addresses credit assignment with the incentive that agents will maximize their contribution to the global reward. The motivation behind COMA is complementary in the sense that it quantifies how much influence an individual agent's action has on the joint action, whereas this work quantifies how much influence other agents' actions have upon an individual agent's policy.

Additionally, investigative work by Jaques et al. (2019), for example, explores causal relationships among agents in MARL through counterfactual reasoning. While promoted for more efficient communication and coordination, this approach can also quantify the contribution of other agents to a self-agent's return.

Some other related works on explicit regularization in MARL originate from maximum-entropy MARL, which reconstructs the return as the reward and the entropy of the policy distribution, weighed by a temperature parameter. An example would be FOP (Zhang et al., 2021), an actor-critic method that factorizes the optimal joint policy from maximum-entropy MARL. The existing

work on quantifying power in MARL by Li & Dennis (2024) studies adversarial power, defined as power associated with an adversarial opponent. The authors discuss various fine-tune parameters for implementing power and measuring power in multi-opponent settings. This work investigates power in settings with communication.

3 Preliminaries

3.1 Communicative MARL

Multi-Agent Reinforcement Learning (MARL) provides a framework for sequential decision-making problems involving multiple interacting agents. Cooperative MARL scenarios are often formalized as Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs), representable by the tuple:

$$\langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, \{\mathcal{O}^i\}_{i\in\mathcal{N}}, P, \{\mathcal{R}^i\}_{i\in\mathcal{N}}, \gamma \rangle$$

Here, $\mathcal{N}$ denotes the set of N agents, $\mathcal{S}$ is the global state space, $\mathcal{A}^i$ is the action space for agent i, and $\mathcal{O}^i$ is its observation space. The function $P(s'|s, \mathbf{a})$ defines the state transition dynamics, where $\mathbf{a} = \{a^i\}_{i\in\mathcal{N}}$ is the joint action assembled from individual agent actions sampled from their policies, $a^i \sim \pi^i(\cdot|o^i)$. Each agent i receives a local observation o^i and a reward $r^i = \mathcal{R}^i(s, \mathbf{a})$. In fully cooperative settings, agents typically share a common team reward, $r^i = R(s, \mathbf{a})$. The collective goal is to learn policies that maximize the expected discounted return $G = \sum_{t=0}^{T} \gamma^t r_t$ where $\gamma \in [0, 1]$ is the discount factor.

A prevalent paradigm for training MARL agents is Centralized Training with Decentralized Execution (CTDE). During the training phase, CTDE algorithms utilize global information, such as the full state or the actions of all agents, to facilitate learning. However, during execution, each agent must operate solely based on its local observation history. For instance, Value Decomposition Networks (VDN) (Sunehag et al., 2017) learn decentralized policies by decomposing the global team Q-function $Q_{tot}(s, a)$ into a sum of individual agent Q-functions $Q^i(o^i, a^i)$, as shown in Equation 1:

$$Q_{tot}(s, a) = \sum_{i\in\mathcal{N}} Q^i(o^i, a^i) \tag{1}$$

Such methods often employ parameter sharing across agent networks to improve learning efficiency.

To further enhance coordination, particularly under partial observability, agents can utilize explicit communication. Communication channels allow agents to exchange information directly. Typically, agent j generates a message m^j based on its internal state or history h^j via a learned communication policy $m^j \sim C^j(\cdot|h^j)$. Agent i's effective input s^i_{input} for decision-making can then incorporate received messages m^{-i} from other agents alongside its own local observation $s^i_{input} = f(o^i, m^{-i})$. Various architectures facilitate this exchange, such as those employing shared network modules for message processing (e.g., CommNet) or learning selective communication via gating mechanisms (e.g., IC3Net). The structure of communication, dictating which agents can exchange messages, is frequently modeled using graph representations.

3.2 Implicit Communication via Graph Neural Networks

Communication can also be learned implicitly through structured feature aggregation. As explored in Li et al. (2020) and utilized in our experiments, Graph Neural Networks (GNNs) offer a powerful mechanism for this. Instead of learning explicit messages, agents first process their local observations o^i_t to generate feature embeddings $x^i_t \in \mathbb{R}^F$. These are stacked into a matrix $X_t = [x^1_t, \ldots, x^N_t]^T \in \mathbb{R}^{N\times F}$. The GNN operates over a dynamic communication graph represented by an adjacency matrix (or Graph Shift Operator) $S_t \in \mathbb{R}^{N\times N}$, where $[S_t]_{ij} = 1$ if agent j can transmit to agent i at time t. (typically based on proximity).

The core mechanism is a graph convolution layer. For a single layer GNN transforming input features $X_{in} \in \mathbb{R}^{N \times F}$ to output features $X_{out} \in \mathbb{R}^{N \times G}$, the operation can be defined as shown in Equation 2:

$$X_{out} = \sigma\left(\sum_{k=0}^{K-1} S_t^k X_{in} A_k\right) \tag{2}$$

Here, $S_t^k X_{in}$ represents features aggregated from the k-hop neighborhood, effectively requiring k rounds of message passing or communication exchanges. K is the maximum communication hop count (filter size) defining the spatial receptive field of the graph convolution. $A_k \in \mathbb{R}^{F \times G}$ are learnable weight matrices specific to hop k, transforming and combining features across hops and dimensions. $\sigma(\cdot)$ is a non-linear activation function applied element-wise. For multi-layer GNNs (L layers), this operation is cascaded, as described by Equation 3:

$$X_l = \sigma[\mathcal{A}_l(X_{l-1}; S_t)] \quad \text{for } l = 1, \ldots, L \tag{3}$$

where X_0 is the input to the first layer and $\mathcal{A}_l$ denotes the graph convolution operation at layer l.

This GNN architecture learns what information is relevant to share and aggregate from the local neighborhood defined by S_t and K. The resulting aggregated feature vector for agent i, denoted $[X_L]_i \in \mathbb{R}^{G_L}$ (the i-th row of the final layer's output), captures context from its communicative neighbors and serves as the input to its decentralized policy network $\pi^i(a^i|[X_L]_i)$.

3.3 Power

The concept of power refers to the influence one agent has over another agent's decision-making and utility. In shared environments, power dynamics play a critical role in determining how agents interact and coordinate. Li & Dennis (2024) introduced power as a formal measure, redefining the optimization criterion as a combination of expected task return and power utility. By incorporating power regularization into the training process, agents can learn policies that are more resilient to states where power imbalances make them vulnerable. Specifically, power quantifies the expected difference between the current joint policy and a hypothetical joint policy where other agents take adversarial actions over k steps.

Power is closely related to the concept of game security, which evaluates the expected return when facing adversarial opponents. In sequential games, the minimax strategy is often used to select the best action among the worst possible outcomes. When power dynamics naturally emerge or are necessary for completing team tasks, power regularization provides designers with a mechanism to control the autonomous behaviors of agents, ensuring they remain robust to adversarial influences.

The original formulation of power by Li and Dennis, referred to as standard power in this paper, estimates the influence agent j has over agent i as follows (Equation 4):

$$\rho_{i:j}^{\text{standard}}(\pi, s) = Q_i^\pi(s, a^i) - \min_{a^j \in A^j} Q_i^\pi(s, a^j) \tag{4}$$

Here, $Q_i^\pi(s, a^i)$ represents the expected return for agent i under the joint policy π, while $\min_{a^j \in A^j} Q_i^\pi(s, a^j)$ represents the worst-case return if agent j takes an adversarial action. In cooperative settings, the joint policy π differs from adversarial policies, stabilizing the estimation of power.

To incorporate power into the learning process, the state-value function for agent i is modified to include a power regularization term, as shown in Equation 5:

$$V_i(s, a) = V_i^\pi(s, a) + \lambda V_i^{\pi, \rho_{i:j}}(s, a) \tag{5}$$

Here, $V_i^\pi(s, a)$ is the original state-value function of agent i, and $V_i^{\pi, \rho_{i:j}}(s, a)$ represents the power component, which penalizes states where agent j exerts high influence over agent i. This regularization can be viewed as a form of reward shaping, where the power measure is used to guide agents toward policies that are less vulnerable to adversarial influence.

4 Power Regularization Over Communication

Learning communication in cooperative settings can lead to more efficient coordination and strong, mutually dependent relationships among agents. However, misaligned agents can exploit these dependencies through sensory manipulation over the communication medium. Given the potential misuse of the communication medium, it is important to address how much dependency an agent delegates to other agents through the communication channel or protocol. Furthermore, it is imperative to ask how much dependency an agent should delegate to the communication medium, regardless of who uses the communication medium. In our work, we train policies to be more resilient to misaligned communication and miscommunication through adversarial training. Adversarial training is the practice of incorporating a variety of adversarial experiences and adversarial communication into training. We propose Communicative Power Regularization (CPR) to improve robustness against adverse impacts from misaligned communication by incorporating adversarial messages during training. Unlike standard power, which keeps the agent state constant, the communication message affects the perceived agent state and, therefore, can be viewed as a form of state regularization where the proximity of other states consisting of adversarial messages affects its perceived utility, similar to that of stochastic transitions by an environment.

4.1 Communicative Power Regularization (CPR)

We define communicative power as the decomposition of power into two components: standard power and power of communication. Standard power is the power delegated to other agents without leveraging the communication channel or protocol. In contrast, the power of communication is the power delegated to other agents over the communication channel or protocol. The total power ρ_{ij}^{CPR} that agent j has over agent i is defined as the sum of these two components (Equation 6):

$$\rho_{ij}^{\text{CPR}}(\pi, s^i, m^j) = \rho_{ij}^{\text{Standard}}(\pi, s^i) + \rho_{ij}^{\text{Communication}}(\pi, s^i, m^j) \tag{6}$$

Here, $\rho_{ij}^{\text{Standard}}(\pi, s^i)$ represents the standard power, which quantifies the influence agent j has over agent i through actions alone, and $\rho_{ij}^{\text{Communication}}(\pi, s^i, m^j)$ represents the power of communication, which quantifies the influence agent j has over agent i through communication.

The communicative power ρ_{ij}^{CPR} is defined as shown in Equation 7:

$$\rho_{ij}^{\text{CPR}}(\pi, s^i, m^j) = Q_i^\pi(s^i, m^j, a^i, a^j) - \min_{a_{\text{adv}}^j \in A^j, m_{\text{adv}}^j \in M^j} Q_i^\pi(s^i, m_{\text{adv}}^j, a^i, a_{\text{adv}}^j) \tag{7}$$

This measures the difference in agent i's expected return when agent j takes an adversarial action and sends an adversarial message m_{adv}^j compared to when agent j follows the joint policy.

Standard power can directly regulate misaligned communication in scenarios where communications are considered actions in the action space. However, its effectiveness in regularizing state-related misaligned communication assumes that appropriate variance is introduced into training, such as simultaneously learning a communication policy with an environment policy. Communicative power incorporates adversarial messages, whether they are individually sent or aggregated. This is particularly important in cases where individual messages are not misaligned but the aggregation of messages is misaligned.

In the traditional approach, the methodology does not provide direct defense mechanisms against adversarial attacks that specifically exploit model parameterization. Adversaries perform adversarial attacks to craft and inject adversarial samples, which usually target a model's parameterization (e.g., a neural network's decision boundary). However, some state-actions performed by certain agent roles contribute more to the environment's expected utility than others, which finite-budget adversaries often consider. To address these challenges, we propose a framework that explicitly accounts for the influence of communication on power dynamics, ensuring robustness against both misaligned communication and adversarial exploitation of model parameterization.

The subsequent expressions are adapted from Li & Dennis (2024) to align with our proposed setting. To incorporate power into the learning process, the state-value function for agent i is modified to include a power regularization term, as shown in Equation 8:

$$V_i(s^i, m^j) = V_i^{\pi}(s^i, m^j) + \lambda V_i^{\pi,\rho_{ij}}(s^i, m^j) \tag{8}$$

Here, $V_i^{\pi}(s^i, m^j)$ is the original state-value function for the task, and $V_i^{\pi,\rho_{ij}}(s^i, m^j)$ represents the power component, which penalizes states where other agents exert influence over agent i through both actions and communication. The parameter λ is a scalar that controls the degree of power regularization over the expected return.

Another approach involves applying both standard power and power of communication separately to enable individualized penalization of states where other agents exert greater control and penalization for states where there is excessive reliance on communicated messages.

The power regularization term $V_i^{\pi,\rho_{ij}}(s^i, m^j)$ is defined as the sum of power rewards $R_i^{\text{power}}(\pi, s_t^i, m^j)$ over states starting from s reached by unrolling the policy π, as given by Equation 9:

$$V_i^{\pi,\rho_{ij}}(s^i, m^j) = \sum_{t=0}^{T} R_i^{\text{power}}(\pi, s_t^i, m^j) \tag{9}$$

In the 2-agent setting, the power reward $R_i^{\text{power}}(\pi, s^i, m^j)$ is defined as shown in Equation 10:

$$R_i^{\text{power}}(\pi, s^i, m^j) = -\rho_{ij}^{\text{CPR}}(\pi, s^i, m^j) \tag{10}$$

This indicates that the power reward penalizes the influence agent j has over agent i through both actions and communication in the state s. In settings with more than two agents, the power reward captures the strongest individual influence that any other agent j exerts on agent i, as defined in Equation 11:

$$R_i^{\text{power}}(\pi, s^i, m^j) = -\max_{j \neq i} \rho_{ij}^{\text{CPR}}(\pi, s^i, m^j) \tag{11}$$

By applying a maximization, this formulation emphasizes the worst-case dependency, making the regularization more sensitive to the most dominant external influence.

Our definition of communicative power is to further specify how power is allocated in the presence of a communication medium. This is in contrast to standard power, which makes no distinction over how power is distributed over coordinating devices or mechanisms. It is within the designer's discretion whether communication is appropriate for a cooperative task.

5 Experiment Results

We evaluate CPR in the Grid Coverage (Blumenkamp & Prorok, 2020) (GC) environment. This setting allows us to assess CPR's effectiveness in a scenario where cooperative agents must coordinate effectively while being resilient to misaligned communication from adversarial entities. To demonstrate the broader applicability of our method, we provide further evaluations in two additional, distinct environments—Predator-Prey (Section C) and Red-Door-Blue-Door (Section D)—in the Supplementary Materials.

Grid Coverage. To evaluate the effectiveness of Communicative Power Regularization (CPR) in mitigating the effects of adversarial communication, we conducted experiments within the non-Convex Coverage map from the Adversarial Comms repository (Blumenkamp & Prorok, 2020). This environment challenges a team of cooperative agents to maximize area coverage on a grid map while contending with explicit adversarial agents designed to disrupt cooperative performance through the communication channel. Agents operate with limited local observations and communication ranges, utilizing a CNN-GNN-MLP architecture to process environmental input, exchange messages via the GNN, and select actions using the MLP. Detailed experimental configuration parameters for this environment are provided in Appendix A (Table 2).

Table 1: Grid Coverage: Cooperative agents' scores (mean (± std.) over 100 trials), comparing performance with and without CPR across various [adversarial, cooperative] agent compositions. Asterisked (*) configurations denote training and evaluation with the same number of agents; others are scaled in evaluation.

[adversarial, cooperative] # of agents	Cooperative agents' scores		Improvement (%)
	With CPR	Without CPR	
$[1, 5]^*$	**257.74 (± 45.93)**	218.82 (± 66.40)	18
$[2, 4]^*$	**204.92 (± 59.84)**	93.56 (± 77.53)	119
$[3, 3]^*$	**188.85 (± 88.50)**	33.53 (± 39.69)	463
$[6, 30]$	**294.65 (± 31.21)**	285.40 (± 43.90)	3
$[4, 8]$	**242.25 (± 53.18)**	116.27 (± 83.21)	108
$[8, 16]$	**262.29 (± 46.94)**	134.22 (± 86.52)	95
$[12, 24]$	**273.94 (± 39.33)**	142.86 (± 89.00)	92
$[18, 18]$	**232.66 (± 77.85)**	61.89 (± 64.19)	276

Our evaluation directly compares the performance of cooperative MARL agents trained with CPR against baseline agents trained without CPR. Both sets of agents were evaluated over 100 trials in scenarios featuring varying numbers of cooperative and adversarial agents, where all agents, including adversaries, actively communicated throughout the episodes. This setup isolates the impact of CPR on the robustness of the cooperative strategy against communication-based attacks.

The comprehensive results presented in Table 1 quantitatively demonstrate the significant advantage conferred by CPR across various team compositions and evaluation scales. We first established baseline performance in configurations where training and evaluation agent counts were identical ([1,5], [2,4], [3,3]). In these scenarios, cooperative agents employing CPR consistently achieved substantially higher mean scores, often with reduced variance as indicated by the standard deviations, compared to baseline agents without CPR. For instance, with 1 adversary and 5 cooperative agents ([1,5]), CPR-trained agents achieved a mean score of 257.74, while baseline agents scored 218.82. This performance gap widened as the proportion of adversarial agents increased; in the challenging [3,3] scenario, agents with CPR maintained a strong cooperative score of 188.85, whereas the performance of baseline agents severely degraded to 33.53.

To assess the scalability of the learned policies, models trained on these starred configurations were then evaluated on significantly larger teams without retraining. Remarkably, CPR-trained agents consistently maintained robust performance levels even in these scaled-up scenarios. For example, when scaling from the [2,4] training setup, agents with CPR achieved mean scores of 242.25 for [4,8] and 262.29 for [8,16], substantially outperforming baseline agents whose scores were 116.27 and 134.22, respectively. This trend continued with further scaling; for instance, in the [12,24] setup (also scaled from [2,4]), CPR agents scored 273.94 against the baseline's 142.86. Even in the highly scaled [18,18] scenario (from [3,3] training), CPR-enabled agents achieved a mean score of 232.66, a stark contrast to the 61.89 achieved by baseline agents. While the score difference in the [6,30] configuration (294.65 with CPR vs. 285.40 without, scaled from [1,5]) was more modest, likely due to the very low adversary-to-cooperative agent ratio, CPR still provided a clear benefit. The overall trend strongly indicates that CPR facilitates the learning of robust coordination strategies that generalize effectively and preserve a high level of performance when deployed in larger, more complex multi-agent systems.

Figure 1 provides a more granular view, illustrating the average coverage percentage achieved by cooperative agents over episode time steps for [1,5], [2,4], and [3,3] configurations, respectively.

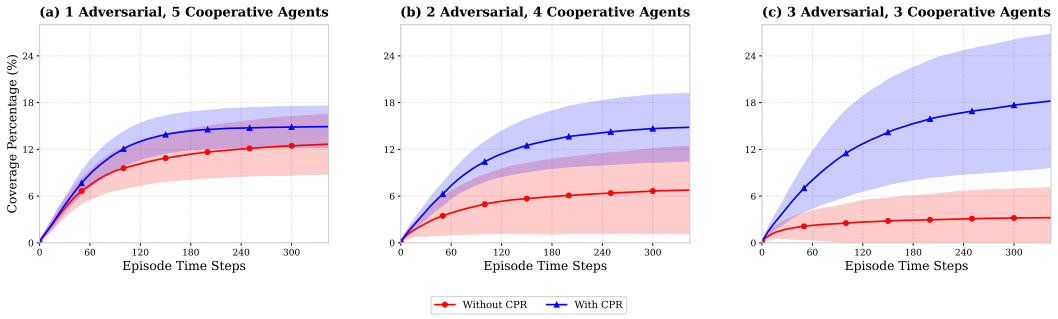

Figure 1: Grid coverage, average cooperative coverage percentage (over 100 trials) across varying team compositions, comparing agents trained with CPR (blue) vs without CPR (red).

In all depicted scenarios, the agents trained with CPR (blue curves) consistently outperform the baseline agents (red curves). They not only reach a higher final coverage percentage but also exhibit faster convergence towards their optimal performance early in the episode. Furthermore, the tighter variance bands (shaded regions) associated with the CPR agents suggest that CPR contributes to more stable and reliable performance across trials, reducing the detrimental impact of adversarial interference.

Synthesizing these findings, both the aggregate scores and the temporal coverage dynamics unequivocally show that CPR enhances the ability of cooperative agents to maintain effective coordination and achieve superior performance in the presence of adversarial communication. By regularizing the power or influence of messages, CPR fosters more robust communication strategies that are less susceptible to manipulation. This validation in the complex Grid Coverage task, featuring decentralized control and explicit adversaries, underscores the practical value of CPR for developing resilient multi-agent systems.

A key concern is that CPR might inadvertently lead agents to avoid communication to prevent penalties. To investigate this, we conducted an ablation study and evaluated 5 CPR-trained cooperative agents, comparing their performance with active communication versus communication explicitly disabled. The results (detailed in Appendix B, Figure 2) clearly refute this concern. Agents utilizing communication (blue triangles) significantly outperform the same agents operating without it (orange circles), achieving faster score accumulation and a higher final score. This confirms that CPR-trained agents do not abandon communication but learn to use it robustly, leveraging it for improved coordination and performance, thus demonstrating that CPR encourages resilient communication strategies rather than avoidance.

6 Conclusion

This work tackled the inherent vulnerability of communicating MARL systems where reliance on information exchange can be exploited by misaligned agents. We argued that prior work on power regularization, focused on action-based influence, inadequately captures the risks associated with delegating control through communication protocols. To address this, we introduced Communicative Power Regularization (CPR), a method to enhance MARL system robustness against misaligned communication by penalizing over-reliance on communication channels and increasing the self-autonomy of agents. Evaluations in the Grid Coverage environment demonstrated CPR's ability to significantly improve cooperative performance and resilience in adversarial communication scenarios, without sacrificing communication when beneficial. While CPR involves a trade-off between robustness and optimal cooperative performance, it provides a practical framework for developing more secure and reliable cooperative MARL systems. Future work could focus on developing adaptive CPR frameworks, for instance by employing an adaptive Power Regularization Factor (λ), and

on addressing heterogeneous trust dynamics. CPR offers a valuable step towards deploying resilient multi-agent systems in challenging real-world environments.

A Grid Coverage: Experimental Configuration Parameters

This appendix details the experimental configuration parameters used for the Grid Coverage environment, as referenced in Section 5.

Table 2: Grid Coverage: Experimental Configuration Parameters

Description	Value
Episode max timestep	345
Communication range	16
Observation range	8
Reward	+1 for visiting a previously uncovered cell; 0 otherwise
Agent actions	up, down, left, right, stay
World shape	24×24
Cooperative training	20 million time steps
Adversarial training	20 million time steps
Power regularization factor (λ)	0.3

B Grid Coverage: Ablation Study on Communication

To address the concern that Communicative Power Regularization (CPR) might inadvertently incentivize agents to cease communication, we conducted an ablation study. This study, referenced in Section 5, compared the performance of five CPR-trained cooperative agents in the Grid Coverage environment under two conditions: (1) with their standard learned communication enabled, and (2) with their ability to send or receive messages explicitly disabled.

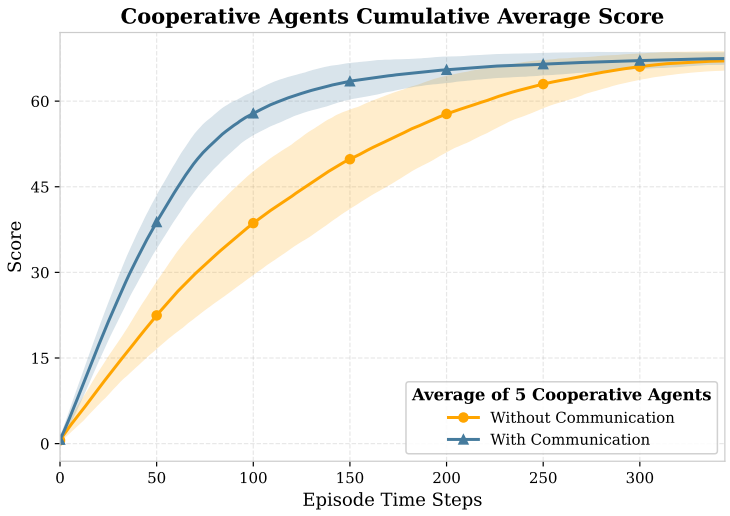

Figure 2: Grid coverage, cumulative average score of CPR-trained agents (5 cooperative, 100 trials), comparing performance with (blue triangles) and without (orange circles) communication.

Figure 2 presents the cumulative average scores over 100 trials for these two conditions. The results demonstrate that agents actively utilizing their learned communication protocols achieve significantly better performance than when communication is unavailable.

References

Jan Blumenkamp and Amanda Prorok. The emergence of adversarial communication in multi-agent reinforcement learning, 2020. URL https://arxiv.org/abs/2008.02616.

Juncheng Dong, Suya Wu, Mohammadreza Sultani, and Vahid Tarokh. Multi-agent adversarial attacks for multi-channel communications, 2022. URL https://arxiv.org/abs/2201.09149.

Yali Du, Lei Han, Meng Fang, Ji Liu, Tianhong Dai, and Dacheng Tao. Liir: Learning individual intrinsic reward in multi-agent reinforcement learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/07a9d3fed4c5ea6b17e80258dee231fa-Paper.pdf.

Martin Figura, Krishna Chaitanya Kosaraju, and Vijay Gupta. Adversarial attacks in consensus-based multi-agent reinforcement learning, 2021. URL https://arxiv.org/abs/2103.06967.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. DOI: 10.1609/aaai.v32i1.11794. URL https://ojs.aaai.org/index.php/AAAI/article/view/11794.

Jun Guo, Yonghong Chen, Yihang Hao, Zixin Yin, Yin Yu, and Simin Li. Towards comprehensive testing on the robustness of cooperative multi-agent reinforcement learning, 2022. URL https://arxiv.org/abs/2204.07932.

Yizheng Hu and Zhihua Zhang. Sparse adversarial attack in multi-agent reinforcement learning, 2022. URL https://arxiv.org/abs/2205.09362.

Aly Ibrahim, Anirudha Jitani, Daoud Piracha, and Doina Precup. Reward redistribution mechanisms in multi-agent reinforcement learning. In *Adaptive Learning Agents Workshop at the International Conference on Autonomous Agents and Multiagent Systems*, 2020.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, Dj Strouse, Joel Z. Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3040–3049. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/jaques19a.html.

Raphael Köster, Kevin R. McKee, Richard Everett, Laura Weidinger, William S. Isaac, Edward Hughes, Edgar A. Duéñez-Guzmán, Thore Graepel, Matthew Botvinick, and Joel Z. Leibo. Model-free conventions in multi-agent reinforcement learning with heterogeneous preferences, 2020. URL https://arxiv.org/abs/2010.09054.

Michelle Li and Michael Dennis. The benefits of power regularization in cooperative reinforcement learning, 2024. URL https://arxiv.org/abs/2406.11240.

Qingbiao Li, Fernando Gama, Alejandro Ribeiro, and Amanda Prorok. Graph neural networks for decentralized multi-robot path planning, 2020. URL https://arxiv.org/abs/1912.06095.

Simin Li, Jun Guo, Jingqiao Xiu, Yuwei Zheng, Pu Feng, Xin Yu, Aishan Liu, Yaodong Yang, Bo An, Wenjun Wu, and Xianglong Liu. Attacking cooperative multi-agent reinforcement learning by adversarial minority influence, 2024. URL https://arxiv.org/abs/2302.03322.

Jieyu Lin, Kristina Dzeparoska, Sai Qian Zhang, Alberto Leon-Garcia, and Nicolas Papernot. On the robustness of cooperative multi-agent reinforcement learning, 2020. URL https://arxiv.org/abs/2003.03722.

Toru Lin, Jacob Huh, Christopher Stauffer, Ser Nam Lim, and Phillip Isola. Learning to ground multi-agent communication with autoencoders. *Advances in Neural Information Processing Systems*, 34:15230–15242, 2021.

Afshin OroojlooyJadid and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning, 2021. URL https://arxiv.org/abs/1908.03963.

Nancirose Piazza and Vahid Behzadan. A theory of mind approach as test-time mitigation against emergent adversarial communication, 2023. URL https://arxiv.org/abs/2302.07176.

Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. Learning when to communicate at scale in multiagent cooperative and competitive tasks, 2018. URL https://arxiv.org/abs/1812.09755.

Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Learning multiagent communication with backpropagation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 2252–2260, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning, 2017. URL https://arxiv.org/abs/1706.05296.

James Tu, Tsunhsuan Wang, Jingkang Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Adversarial attacks on multi-agent communication, 2021. URL https://arxiv.org/abs/2101.06560.

Atsushi Ueshima, Shayegan Omidshafiei, and Hirokazu Shirado. Deconstructing cooperation and ostracism via multi-agent reinforcement learning, 2023. URL https://arxiv.org/abs/2310.04623.

Tianhao Zhang, Yueheng Li, Chen Wang, Guangming Xie, and Zongqing Lu. Fop: Factorizing optimal joint policy of maximum-entropy multi-agent reinforcement learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12491–12500. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zhang21m.html.

Supplementary Materials

The following content was not necessarily subject to peer review.

C Additional Environment: Predator-Prey

The Predator-Prey environment is a grid-world scenario designed to evaluate MARL algorithms across cooperative, competitive, and mixed settings. In this setup, $N-1$ predators (specifically, 3 in our experiments) aim to capture a single prey agent. All agents possess limited local vision. During cooperative interactions, predators utilize communication to coordinate their approach to the prey's location. A time-step penalty encourages predators to find the most efficient path.

Our experimental configuration for Predator-Prey is detailed in Table 3. The reward mechanism incorporates δ_i, indicating whether agent i has reached the prey; n_t, the count of agents at the prey's location at timestep t; and ξ, a parameter distinguishing competitive ($\xi = -1$), mixed ($\xi = 0$), and cooperative ($\xi = 1$) objectives. Each agent's observation space consists of a tensor reflecting its local visual field, employing one-hot encoding for the locations of predators and the prey.

We established a baseline using IC3Net trained under cooperative settings with communication always enabled (referred to as 'IC3Net(always-comm baseline)', see Table 4). This baseline represents controllers reliant on continuous communication during cooperative training. Evaluating this model with communication disabled during testing ('IC3Net(no comm)' in Table 4) revealed a drop in performance (0.84 success rate vs. 1.0 during training), demonstrating that the cooperatively trained agents developed policies overly dependent on communication and lacked self-sufficiency when messages were absent. This finding highlights a potential vulnerability: even in cooperative scenarios without explicit adversaries, the unexpected absence of communication from a teammate can disrupt coordination, mimicking the effect of adversarial interference.

Subsequently, we trained an IC3Net model incorporating communicative power regularization ('IC3Net(CPR)') within the cooperative setting ($\lambda = 0.25$, Table 4). Test results demonstrate that these agents maintained high performance even when communication was disabled ('IC3Net(CPR no comm)' in Table 4, success rate 1.0), indicating improved robustness compared to the baseline.

Table 3: Predator-Prey, Experimental configuration parameters.

Description	Value
# of agents	predator: 3 prey: 1
reward	$r_i(t) = \delta_i * r_{explore}$ $+(1-\delta_i) * n_t^{\xi} * r_{prey} * \lvert\xi\rvert$
maximum timestep	20
predator actions	up, down, left, right, stay
observation space	(2*vision + 1)^2 $\times$(ohe-location, ohe-predator, ohe-prey)
vision	1
grid-size	5×5
training	2,000 epochs
testing	1,000 episodes
power regularization factor (λ)	0.25

We further investigated performance in a competitive setting. The baseline IC3Net model, trained cooperatively, failed entirely when tested in the competitive environment (success rate 0.0), despite communication channels remaining open ('IC3Net(always-comm)' competitive test, Table 4). This

Table 4: Predator-Prey, IC3Net success rates with/without CPR in cooperative/competitive settings and communication ablation tests.

test/train	algorithm	setting	success
train	IC3Net(always-comm baseline)	cooperative	1.0
test	**IC3Net(no comm)**	cooperative	**0.84**
test	**IC3Net(always-comm)**	competitive	**0.0**
train	IC3Net(CPR always-comm)	cooperative	1.0
test	**IC3Net(CPR no comm)**	cooperative	**1.0**
test	**IC3Net(CPR always-comm)**	competitive	**1.0**

underscores the difference between merely tolerating the absence of communication (where agents might learn redundant strategies) and possessing resilience against potentially uncooperative or misleading communication inherent in competitive settings. Policies that simply minimize communication among non-cooperative agents are not necessarily prepared for adversarial communication that exploits naive trust in received messages.

Conversely, the IC3Net model trained with communicative power regularization achieved successful task completion in the competitive setting, even with communication enabled ('IC3Net(CPR always-comm)' competitive test, Table 4, success rate 1.0). This outcome suggests that applying power regularization over the communication channel guides agents towards learning more autonomous policies, equipping them to handle misaligned communication effectively.

More generally, our findings imply that incorporating CPR during training in cooperative settings can foster behaviors typically associated with policies learned under mixed or competitive conditions. The extent to which agent behavior deviates towards non-cooperative resilience depends on factors such as the power regularization factor (λ) and the permissible reduction in expected utility associated with adopting these more robust, autonomous strategies.

D Additional Environment: Red-Door-Blue-Door

To evaluate performance in a scenario demanding strict sequential coordination under potentially misleading communication, we utilize an adapted version of the Red-Door-Blue-Door (RDBD) environment (Lin et al., 2021). This version focuses solely on the coordination and communication aspects without grid-world movement, providing a clear testbed for communication robustness. The environment features three distinct policy roles: a Red agent, a Blue agent, and a Communication agent.

The Red and Blue agents are the primary actors. The Red agent can choose to open the red door or wait. The Blue agent can choose to open the blue door, open the red door, or wait. The cooperative task objective is achieved if the Red door is opened before or concurrently with the Blue door within 50 time steps, yielding a positive reward annealed by completion time. Opening the Blue door first or exceeding the time limit results in task failure and a negative reward. The Red agent possesses a hidden internal state determining whether its objectives are cooperative or adversarial (inducing a zero-sum reward structure between Red and Blue).

Communication is mediated by the dedicated Communication agent, which observes the Red agent's hidden state and generates a message $m \in \{0, 1\}$, intended to signal this state to the Blue agent at the start of each episode. The Blue agent observes the status of both doors and this message m, while the Red agent only observes door statuses. This setup explicitly isolates the communication channel. The challenge lies in the Blue agent learning appropriate reliance on m, as the Communication agent could represent a faulty, noisy, or explicitly adversarial channel sending misleading information. RDBD thus serves as a focused benchmark to assess if our proposed method (CPR) fosters

Table 5: Red-Door-Blue-Door, MAPPO performance (rewards, episode length) with/without CPR under cooperative/adversarial communication.

test/train	algorithm	setting	blue reward	red env reward	comm acc	episode len
train	MAPPO(cooperative baseline)	cooperative	1.000	1.000	1.000	2.000
train	**MAPPO(no adv-comm)**	competitive	0.475	-0.475	0.925	3.01 ($\pm$ 0.005)
train	MAPPO(adv-comm (ideal))	competitive	0.499 ($\pm$ 0.020)	-0.499 ($\pm$ 0.020)	0.411	3.000
train	**MAPPO(CPR)**	cooperative	0.464	-0.464	1.0	**3.000** ($\pm$ 0.014)
test	**MAPPO(no CPR adv-comm)**	competitive	0.368	-0.368	-	**3.64** ($\pm$ 0.933)
test	**MAPPO(CPR adv-comm)**	competitive	0.497	-0.497	-	**3.016** ($\pm$ 0.127)

policies robust against risks associated with the delegation of control via potentially compromised communication. Key configuration details are in Table 6.

Standard power and CPR are defined over Q-values estimations; however, for the implementation, we enact the immediate reward penalty through $k = 1$-step adversarial action and adversarial message. This is defined in Equation 12:

$$r_i(o, m_j, a_i, a_j) = r_i(o, m_j, a_i, a_j) + \lambda r_i(o, m_j^{adv}, a_i, a_j^{adv}) \tag{12}$$

Our experiments in RDBD (Table 5) demonstrate CPR's effectiveness in enhancing robustness. The standard MAPPO baseline trained cooperatively (MAPPO(cooperative baseline)) learns an optimal, fast (ep. length 2.000), but communication-dependent strategy. Consequently, when faced with adversarial communication (MAPPO(no CPR adv-comm)), its performance collapses (blue reward 0.368 vs 1.0), highlighting the fragility of policies over-reliant on communication integrity.

In contrast, MAPPO agents trained with CPR ($\lambda = 0.75$) exhibit significant resilience. While their cooperative strategy (MAPPO(CPR)) is slightly more cautious (ep. length $\approx$3.000, favouring the safer communication-independent sequence), they maintain strong performance under adversarial communication (MAPPO(CPR adv-comm)), achieving a blue reward of 0.497 and an episode length of 3.016. Notably, this performance is nearly identical to the ideal communication-independent strategy (MAPPO(adv-comm (ideal))).

Table 6: Red-Door-Blue-Door, Experimental configuration parameters.

Description	Value
# of agents	blue: 1 red: 1 dummy-comm: 1
max timestep	50
reward	$r(s,a) = \begin{cases} 1/(t-1) & t < 50 \;\&\&\; \text{all doors open} \\ 0 & t < 50 \\ -1 & 50 \leq t \end{cases}$
red reward (competitive)	- blue reward
observation	door-status (red) red-team status (dummy) door-status + message (blue)
training	100,000 timesteps
testing	1,000 episodes
power regularization factor (λ)	0.75

This demonstrates that CPR successfully guides the Blue agent to mitigate risks by reducing dependence on the communication channel. By regularizing the influence exerted via communication

during training, CPR encourages the agent to learn policies that are less susceptible to manipulation, effectively achieving robust, autonomous behavior akin to that learned under adversarial conditions, but within a cooperative training framework.