The Power in Communication: Power Regularization of Communication for Autonomy in Cooperative Multi-Agent Reinforcement Learning

Nancirose Piazza
npiazza@newhaven.edu
University of New Haven

Vahid Behzadan
vbehzadan@newhaven.edu
University of New Haven

Stefan Sarkadi
stefan.sarkadi@kcl.ac.uk
King's College London

Abstract

Communication plays a vital role for coordination in Multi-Agent Reinforcement Learning (MARL) systems. However, misaligned agents can exploit other agents' trust and delegated power to the communication medium. In this paper, we propose power regularization as a method to limit the adverse effects of communication by misaligned agents. Specifically, we focus on communication which impairs the performance of cooperative agents. Power is a measure of the influence one agent's actions have over another agent's policy. By introducing power regularization over communication, we aim to allow designers to control or reduce an agent's dependency on communication when appropriate. With this capability, we aim to train agent policies with resilience to performance deterioration caused by misuses of the communication channel or communication protocol. We investigate several environments in which power regularization over communication can be valuable to regularizing the power dynamics among agents delegated over the communication medium.

1 Introduction

Coordination in Multi-Agent Reinforcement Learning is the pursuit of action synchronization among agents in a shared environment but with individual objectives, often to avoid worst-outcomes. In this context, communication is seen through the lens of information theory and control. Hence, communication is defined as the exchange of information among agents across an established channel which is often used to facilitate coordination. Cooperative Multi-Agent Reinforcement Learning (CoMARL) primarily focuses on parameter-sharing, team training efficiency, and the development of cooperative mechanisms for addressing team challenges. While many of these algorithms optimize training through parameter-sharing, the resulting joint policies can exhibit undesirable behaviors. Examples of undesirable behaviors include individual free-riding (Ueshima et al., 2023), over-reliance on irrelevant features, learning irrelevant conventions (Köster et al., 2020), and lacking experience to respond to acts by misaligned agents including acts seen in adversarial settings. In settings of misaligned agents, public communication channels can be misused and abused to sabotage cooperative agents that have learned to signal over communication channels.

Objective misalignment describes non-cooperative agents in a shared environment where agent goals or objectives are not aligned. Misaligned agents may pursue self-interested objectives indifferent or at the cost of other agents. A crucial property of MAS communication in real world settings where teams of cooperative, autonomous agents are deployed, is that of *resilience* to objective misalignment. To achieve resilient MAS communication it is important to evaluate objective misalignment in both individual and team settings which are not traditionally addressed in out-of-the-box CoMARL algorithms.

Communication is a prominent research direction for collaborative and cooperative settings (Oroojlooy & Hajinezhad, 2023). Communication in MARL literature has primarily been modeled as a dedicated communication protocol controller which propagates information across agents and evaluates the joint policy as seen with CommNet (Sukhbaatar et al., 2016) and IC3Net (Singh et al., 2018). In settings where agents simultaneously learn a communication policy and environment policy, agents may inherently learn to regularize against misaligned communication whether those are (1) mistakes from the co-learning process of other agents or (2) are intentionally learned by agents with misaligned objectives. However, not all settings may see benefit to self-learnt communication, enabling easier means for adversaries to exploit naive learnt behaviors through the usage of adversarial communication.

We define *misaligned communication* as messages transmitted by misaligned agents over an established communication channel that adversely impact the performance of a recipient agent. In the context of MARL communication, if we introduce resilience to misaligned communication in cooperative settings, then cooperative agents will learn policies more suitable for facing communication seen in competitive or mixed settings. This would be in contrast to the context of *adversarial attacks*, which are performed by an explicit adversary with zero-sum objectives against the targeted model or team. Adversarial attacks often deliver their payload to their targets by exploiting vulnerable components of the system. Adversarial attacks on multi-agent communication (Tu et al., 2021) show that targeting the communication channel can drastically deteriorate the performance of agents with communication dependence. This also includes MARL-based adversarial attacks over multi-channel communications (Dong et al., 2022) which furthers exasperates the difficulties of training large-scale MARL to communicate. *Power*, the influence of one agent over another agent's decision-making, is a concept that associates amounts of an agent's expected utility with the actions of other agents. Although in most settings, agents are not explicitly optimizing to maximize their power over other agents, reconstructing the utility as additive components of different types of utility is a step towards better control over agent behaviors. Specifically, this improves the robustness of agent policies from allowing power-seeking behaviors. Similar to the notion of intrinsic reward (Du et al., 2019) in MARL, forward optima reconstruction may be an effective alternative to attempting to deconstruct an existing blackbox policy. In this paper we use power regularization over communication to serve as a means to learn a communication policy which balances dependencies on agents with powerful roles and place value on an agent's autonomy.

The main contributions of this work are as follows:

- We propose modified power regularization as a method to mitigate negative effects of misaligned communication in CoMARL systems. This provides designers with the ability to train more resilient policies to misuses of communication through limiting delegated power over communication channels or protocols.

- We investigate the effectiveness of power regularization over communication in two benchmark environments: Red-Door-Blue-Door and Predator-Prey. We demonstrate its effectiveness in mitigating the impact of misaligned communication and compare it against cooperative baselines.

We have structured the paper in the following manner. First, we present a background section on related works, and introduction of communication in CoMARL frameworks. We introduce the definition of power as part of the optima criterion. Then, we propose power regularization over communication as a variant of the original power formulation. We demonstrate with experiments in two environments: Red-Door-Blue-Door and Predator-Prey. Finally, we address the limitations of this approach and discuss future work.

2 Background

Adversarial communication in MARL settings is often highlighted by its emergence in non-cooperative settings (Blumenkamp & Prorok, 2021) as 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., 2023). Adversarial training, an approach to mitigating against adversarial interests, is an umbrella-term for incorporating adversarial interactions into training. This hardens and produces more resilient policies against adversarial opponents. In support, there are works on the robustness of CoMARL (Lin et al., 2020; Guo et al., 2022). There are many 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 (Ibrahim et al., 2020) was proposed to portion out the team reward based on individual contributions in order to address free-riders. Foerster et al. (2018) proposed Counterfactual Multi-Agent Policy Gradients (COMA), which marginalizes out single agent actions. Additionally, the investigative work on the casual relationship among agents in MARL through counterfactual reasoning by Jaques et al. (2019) was promoted for more efficient and meaningful communication and coordination. However, this can be used as motivation for quantifying the contribution of other agents to a specific agent's return.

Alternatively, the works on empowerment can also shape policy learning. Applied to multi-agent simulations (Guckelsberger et al., 2016), Empowerment is described as capturing quantitatively how much an agent is in control of the world. Transfer empowerment, the potential causal influence one agent has on another, can be used to quantify collaborative and coordinating behaviors (Salge & Polani, 2017). Social empowerment, another variation of transfer empowerment, has been applied to robust MARL for coordination and communication (van der Heiden et al., 2020). The closest existing work to ours is on quantifying adversarial power (Li & Dennis, 2023) in MARL. Adversarial power refers to power associated with an adversarial opponent. The authors discuss various fine-tune parameters for implementing power and measuring power in multi-opponent settings. Our work explicitly investigates power in settings with communication channels and communication protocols.

2.1 Communicative MARL

A communicative multi-agent reinforcement learning (MARL) can be modeled by a standard MARL framework:

$$< N, \{A^i\}_{i\in N}, \{S^i\}_{i\in N}, \{R^i\}_{i\in N}, T, \gamma, \{M^i\}_{i\in N} >$$

where in a N-multi-agent system the joint-action $a := \prod a^i$, $a^i \sim \pi(*|s^i)$. T is the transition matrix $T : s, a \to s$ for system state s. γ is a discount factor $\in (0, 1]$, message $m^i \in M^i$ is communicated by an agent i.

For cooperative settings, methods such as Value Decomposition Network (VDN) or QMIX formulate the global team Q-value as factors of individual Q-values. We express this in the following equation:

$$Q(a|s) := \phi(\bigtimes_{i\in N} Q^i(a^i|s^i))$$

ϕ is a set operator where in VDN, this operator is linear summation and in QMIX it is a mixing function that takes in individual Q-values. $\times$ is the symbol for the cartesian product.

Optimal joint-policy π^* can be extracted from greedy, locally optimal individual policies by the following expression:

$$\pi^* := \max_{\pi} Q_\pi(s, a) = \bigtimes_{i\in N} \max_{\pi^i} Q^i_{\pi^i}(s^i, a^i)$$

These methods can be performed with individualized Q-Learning agents (IQL) or share parameters as seen with many centralized critic methods.

Extending to simple communication with a dedicated communication network $C^i : h^i \to (0, 1]^R$, at every timestep t, agents participate in a communication protocol and exchange messages $m^i \sim C^i(h^i)$ for some i agent. Each agent i's local hidden state h^i is derived from $h^i := (s^i, m^{-i})$ given messages m^{-i} from all other $-i$ agents and agent i's local state s^i. Many communication models leverage graph representation for communication message transfer between senders and receivers.

CommNet & IC3Net. Communication controllers or communication protocols outline the topology of the communication network. This would include which agents communicate and how their messages are aggregated together with an agent's local state. Some shared-communication MARL architectures such as CommNet (Sukhbaatar et al., 2016) are communication controller models that are trained by some Reinforcement Learning algorithm such as REINFORCE. CommNet uses continuous communication cycles to determine joint actions over multi-rounds of communication. IC3Net (Singh et al., 2018) is an individualized continuous controller communication model oriented around individualized agent rewards. There are other variants of communication architectures but we outline these two as motivation.

2.2 Power

The concept of power, which refers to the influence and control that one agent has over another agent's decision-making and utility, holds significant importance in shared environments. Li & Dennis (2023) introduced power as a measure and proposed a redefinition of the criterion optima as the combination of the expected return from a task and power utility. Consequently, employing power regularization through the power measure can serve as an effective approach to learning policies that exhibit stronger resilience towards power-vulnerable states. Specifically, power represents the anticipated disparity between the current joint policy and a joint policy where other agents take adversarial k-step actions. In scenarios where power dynamics among agents either emerge naturally or are necessary for achieving team task completion, the application of power regularization can provide designers with greater control over the autonomous behaviors of agents.

We denote the original power from Li & Dennis (2023) as *standard power* to be the value estimation that an agent j has over an agent i in the following Equation 2.2:

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

where $r_i(*)$ is first evaluated under agent i's action taken by joint-policy π^i, π^j and compared to agent i's reward given agent j performs an adversarial action a_j taken under a joint-policy with the presence of an adversarial agent for 1-step. In cooperative settings, the joint policy composed of individualized π^i and π^j will differentiate in probability distribution than those of adversarial objectives, stabilizing the power estimation. Power regularization penalizes an agent i's state value based on whether other agents have higher power over agent i in a given state.

Furthermore, the Q-value of an agent i is modified to be an interpolation between two expected return value estimation distributions, the original Q-value for the original task and maximizing expected return of power reward over other agents. This is presented in the following equation:

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

3 Power Regularization Over Communication

Learning with communication in cooperative settings can result in more efficient coordination and strong, 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 should an agent delegate over the communication medium regardless of whoever uses the communication medium. In our work, we train policies to be more resilient to the presence of misaligned communication through adversarial training. Adversarial training is the practice of incorporating a variety of adversarial experiences and adversarial communication into training. We propose modified power regularization which incorporates adversarial messages. This is to improve robustness against adverse impacts from misaligned communication. Alternatively, this can be viewed as state regularization over possible adversarial messages receivable at a given state. This is comparable to training with stochastic environment transitions.

Power of Communication. We define power over communication as the decomposition of power into two further components: communicative power and implicit power. Communicative power is the power delegated to other agents over the communication channel or protocol. Implicit power is the power delegated to other agents without leveraging the communication channel or protocol. λ is a scalar value that reduces the effect of power regularization over the expected return. This is presented in the following Equation 3.

$$Q_i(s,a) = Q_i^{\pi}(s,a) + \lambda \overbrace{(\underbrace{Q_i^{\pi,\rho_{i:j}}(s, m=\emptyset, a)}_{implicit\ power} + \underbrace{Q_i^{\pi,\rho_{i:j}}(s,m,a))}^{power}}^{power}_{communicative\ power}$$

Given agent i's local hidden state $h^i := (s^i, m^j)$ where s^i is agent i's local state and m^j is a message sent from agent j, we define the *modified power* agent j has over agent i through message m^j:

$$\rho_{i:j}^{modified}(\pi^i, \pi^j, s^i, m^j) = Q_i^{\pi^i,\pi^j}(s, m^j, a^i, s^{i'}, m^{j'}) - \min_{a^j \in A^j} \left(Q_i^{\pi^i,\pi^j}(s^i, m^j_{adv}, a^i, s^{i'}, m^{j'})\right)$$

The standard power formulation evaluates an agent i's on-policy action expected performance against hypothetical adversarial opponents. It quantifies the portion of the return associated with opponent j's adversarial actions. In scenarios where communication is modeled with the action space, standard power can directly regulate misaligned communication. However, standard power will not necessarily regularize misaligned communication shared by a communication network protocol. Modified power incorporates adversarial messages shared over communication network protocols to regularize power delegated to other agents over communication protocols. These adversarial messages can be sent individually or aggregated together. This is important in cases where individual messages are not misaligned, but the aggregation of messages together is misaligned. Settings with limited, shared communication bandwidth without scheduling is an example of where even aligned agents can negatively impact each other e.g., multiple speakers result in less effective messaging.

Together, we regularize the Q-value function by the modified power term as seen in the following equation:

$$Q_i(s,a) = Q_i^{\pi}(s,a) + \lambda_1 Q_i^{\pi,\rho_{i:j}^{modified}}(s,a)$$

Where λ_1 is a scalar that controls the degree of power regularization over the expected return. Our definition of power over communication is to further specify how power is allocated over 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 the designer's decision on whether if it is appropriate to regularize power over communication in particular settings.

4 Experiment Results

We have two environments which we evaluate modified power over communication: (1) A variant of Red-Door-Blue-Door (RDBD) and (2) Predator-Prey (Singh et al., 2018) (PP). We chose these settings because for power regularization to be feasible in changing behaviors in the environment, we must have multiple, acceptable solutions that may be seen in cooperative, competitive or mixed settings. In addition, RDBD demonstrates communication over the action space while PP demonstrates the usage of a communication controller. Results are presented in Table 2. We do not use expected return as a performance measure given power regularization modifies the resulting expected return. In addition, we evaluate misaligned communication under adversarial settings. Adversarial communication can be implemented by external adversarial attacks, but we emphasize in our settings, the source of conflict is due to misaligned agents.

Red-Door-Blue-Door. The first environment is an adaption to the environment Red-Door-Blue-Door (Lin et al., 2021) with eliminated grid-world components. In RDBD, there are two environment policies and one communication policy. The environment policies are two agents designated as red agent and blue agent. The communication policy is treated as a third agent to separate environment actions from communication actions. The environment agents are placed in a room where every agent is assigned a door. 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 red agent's observation space is a one-hot encoded representation of each door's open status. The blue agent's observation is the one-hot encoded representation of each door's status and a message vector sent by the communication agent. In fully cooperative settings between the red and blue agents, team reward is given when the red door is opened before or concurrently to the blue door with reward of 1. The team reward is annealed based on the number of timesteps it takes for the episode to end. The max number of timesteps is 50 excluding the first timestep dedicated for communication exchange. If the maximum timesteps is reached, the reward is defaulted -1 to avoid annealing the penalty. The red agent has a hidden status flag which represents whether it is an adversarial policy or not. When the red agent's flag is adversarial, then the red and blue agents' reward is zero-sum. Based on its flag status, this determines whether the current game is cooperative or competitive. The communication agent is able to observe the red agent's team status (adversarial or not) and communicate its prediction $m \in \{0, 1\}$ of the red agent's team status to the blue agent at the first timestep. The cooperative, competitive, or mixed nature of the communication agent represents (1) the use and misuse of the communication channel when the blue agent delegates power over the communication channel and (2) the separation of communication actions from environment actions. Environment configurations are presented in Table 1b. We include Figure 1 as visualization of RBDB under cooperative communication agent over the first several timesteps.

description	value
# of agents	predator:$N-1$ prey:1
reward	$r_i(t) = \delta_i * r_{explore}$ $+(1-\delta_i) * n_t^{\lambda} * r_{prey} * \lvert\lambda\rvert$
maximum timestep	20
predator actions	up, down, left, right, stay
observation space	(2*vision + 1)^2 ×(ohe-location, ohe-predator, ohe-prey)
vision	1
Grid-size	5×5

(a) Predator-Prey Configurations

description	value
# of agents	blue:1 red:1 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 space	door-status (red) adversarial status (comm) door-status + message (blue)
action space	open red door, wait (red) open red door, open blue door, wait (blue) {0,1} (comm)

(b) Red-Door-Blue-Door Configuration

Standard power and modified power are defined over Q-values estimations, however for the implementation, we enact the immediate reward penalty through $k = 1$-step adversarial action and message. This is defined in Equation 4:

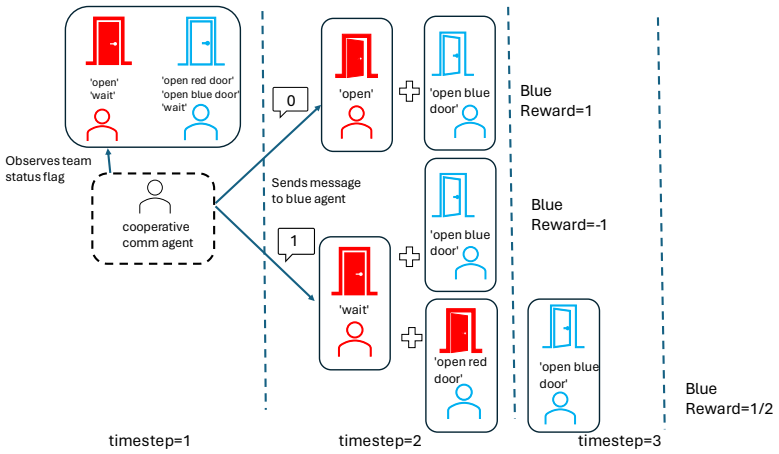

Figure 1: Red-Door-Blue-Door

$$r_i(o, m_j, a_i, a_j) = r_i(o, m_j, a_i, a_j) + \lambda_1 r_i(o, m_j^{adv}, a_i, a_j^{adv})$$

test/train	algorithm	λ_1	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(∓ 0.005)
train	MAPPO(adv-comm (ideal))	-	competitive	0.499(∓ 0.020)	-0.499(∓ 0.020)	0.411	3.000
train	**MAPPO(mod-power)**	0.75	cooperative	0.464	-0.464	1.0	**3.000**(∓ 0.014)
test	**MAPPO(no mod-power adv-comm)**	-	competitive	0.368	-0.368	-	**3.64**(∓ 0.933)
test	**MAPPO(mod-power adv-comm)**	-	competitive	0.497	-0.497	-	**3.016**(∓ 0.127)

(a) Red-Door-Blue-Door (1 red versus 1 blue, train:100,000 timesteps, test:1,000 episodes)

test/train	algorithm	λ_1	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(mod-power always-comm)	0.25	cooperative	1.0
test	**IC3Net(mod-power no comm)**	-	cooperative	**1.0**
test	**IC3Net(mod-power always-comm)**	-	competitive	**1.0**

(b) Predator-Prey (3 predators versus 1 prey, train:2,000 epochs, test: 1,000 episodes)

Table 2: Experiments: Red-Door-Blue-Door & Predator-Prey

In Table 2a, we show the cooperative baseline, demonstrating the expected and ideal behaviors under cooperative settings. However, this solution exhibits powerful behaviors. Specifically, the communication agent has power over the blue agent's behavior through its communicated message. This delegation of power to the communication agent is risky in non-cooperative settings. We highlight the length of the episode as a metric to represent the immediate and rash behavior of the cooperative solution. We also report the performance of the blue agent's policy under competitive settings without adversarial communication. This demonstrates that the blue agent can perform even when there are non-adversarial messages contributing to its observation space. We expect the ideal behavior for the blue agent when faced with an adversarial opponent and therefore adversarial communication is for it to have little dependency on the message contribution of its observation space. This is demonstrated with MAPPO (adv-comm (ideal)). We achieve similar performance to the ideal behavior when training with modified power regularization over communication under cooperative settings (MAPPO (mod-power)). We highlight that without modified power regularization in the presence of adversarial communication (MAPPO - no mod-power, adv-comm), agents can perform worse at test-time in comparison to training with modified power (MAPPO - mod-power, adv-comm). With modified power regularization, the blue agent learned a policy less dependent on communication

and replicate behaviors it would have learned in competitive settings. Therefore, modified power regularization provides designers with the capability of controlling how much sensitivity an agent has to communicated messages. This is in contrast to approaches that minimize communication frequency among non-cooperative agents. Traditional methods of regularization in training such as stochastic dropping of message can inherently regularize over-dependency as well. However, our work distinguishes the need for adversarial message training.

Predator-Prey. Predator-Prey is a grid-world environment that evaluates cooperative, competitive and mixed $N-1$ (3 predators) predators against one prey. Agents are populated onto the grid and the predators must navigate to find the prey limited by their local vision. In cooperative settings, predators communicate with each other at every timestep to arrive at the prey location. There is a timestep penalty which incentives predators to find the shortest path to the prey. In our setup we have the IC3Net baseline in Table (2b) that has communication always enabled, environment configurations are presented in Table 1a. δ_i denotes whether agent i found the prey, n_t is the number of agents in the prey at timestep t, λ represents competitive (λ =-1), mixed (λ =0) and cooperative (λ =1) settings. The observation space is tensor of dimension sized to the local vision by the one-hot encoded location, predator, and prey. IC3Net is used to represent controllers that always communicate during cooperative training settings. We evaluate this policy under disabled communication, demonstrating that cooperative agents were over-dependent on communication and lacked self-sufficiency without the presence of communication. This is problematic in cooperative settings where the lack of communication has similar impact to adversarial intervention. We then trained an IC3Net model with modified power regularization (IC3Net - mod-power always-comm) under cooperative settings. We evaluated in test settings, demonstrating that agents trained with modified power regularization were able to achieve viable performance without the presence of communication. Beyond using the lack of communication presence for adversarial interests, modified power regularization also address adversarial message communication precedent to mixed or competitive settings. We show agents that are trained to always communicate under cooperative settings may not be able to complete tasks under competitive settings (success rate 0.0) with communication enabled (2b). This highlights the difference between our approach that desensitizes agents from communication versus minimizing communication among non-cooperative agents. The latter approach does not enable resilient behaviors in the presence of adversarial communication.

5 Conclusion

Communication can be an effective protocol in cooperative settings for achieving coordination in Multi-Agent Reinforcement Learning. Agents often learn to delegate power of the communication medium. However, misaligned agents can misuse the communication channel or communication protocol to exploit agents have delegated power over communication. Power, a quantitative measures of how much one agent's action affects another agent's decision-making, can be used to regularize policies from learning power-vulnerable behaviors and dynamics. The original power formulation addresses adversarial power over the action space. However, this does not address misaligned communication shared by communication controllers or protocols. We propose power regularization over communication and communication controllers as a means to regulate power-seeking behaviors and power-allowing behaviors through the communication medium. We explicitly introduce power regularization over communication as adversarial training in the presence of adversarial communication. This enables designers to train more resilient policies to both misaligned agents and the presence of misaligned communication. Regulating with power over the communication channel faces similar set of challenges as outlined in the work on original power. Examples of challenges include fine-tuning for distinguishable behaviors, training with varying $k > 1$ adversarial steps, and choice of power aggregation given multiple opponents. Future work may investigate settings of extending power regularization over communication to standard instances of adversarial attacks and defining various power aggregations for communication. In this work we investigate and demonstrated that CoMARL frameworks that use communication can benefit from power regularization over communication by limiting how much power is allowed to be delegated to the communication medium.

References

Jan Blumenkamp and Amanda Prorok. The emergence of adversarial communication in multi-agent reinforcement learning. In *Conference on Robot Learning*, pp. 1394–1414. PMLR, 2021.

Juncheng Dong, Suya Wu, Mohammadreza Sultani, and Vahid Tarokh. Multi-agent adversarial attacks for multi-channel communications. *arXiv preprint arXiv:2201.09149*, 2022.

Yali Du, Lei Han, Meng Fang, Ji Liu, Tianhong Dai, and Dacheng Tao. Liir: Learning individual intrinsic reward in multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Martin Figura, Krishna Chaitanya Kosaraju, and Vijay Gupta. Adversarial attacks in consensus-based multi-agent reinforcement learning. In *2021 American Control Conference (ACC)*, pp. 3050–3055. IEEE, 2021.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

Christian Guckelsberger, Christoph Salge, and Simon Colton. Intrinsically motivated general companion npcs via coupled empowerment maximisation. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, pp. 1–8, 2016. doi: 10.1109/CIG.2016.7860406.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 115–122, 2022.

Yizheng Hu and Zhihua Zhang. Sparse adversarial attack in multi-agent reinforcement learning. *arXiv preprint arXiv:2205.09362*, 2022.

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. *arXiv preprint arXiv:2010.09054*, 2020.

Michelle Li and Michael Dennis. The benefits of power regularization in cooperative reinforcement learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 457–465, 2023.

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, 2023.

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, 2020. doi: 10.1109/SPW50608.2020.00027.

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 Oroojlooy and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722, 2023.

Nancirose Piazza and Vahid Behzadan. A theory of mind approach as test-time mitigation against emergent adversarial communication. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 2842–2844, 2023.

Christoph Salge and Daniel Polani. Empowerment as replacement for the three laws of robotics. *Frontiers in Robotics and AI*, 4, 06 2017. doi: 10.3389/frobt.2017.00025.

Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. Learning when to communicate at scale in multiagent cooperative and competitive tasks. *arXiv preprint arXiv:1812.09755*, 2018.

Sainbayar Sukhbaatar, Rob Fergus, et al. Learning multiagent communication with backpropagation. *Advances in neural information processing systems*, 29, 2016.

James Tu, Tsunhsuan Wang, Jingkang Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Adversarial attacks on multi-agent communication. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7768–7777, 2021.

Atsushi Ueshima, Shayegan Omidshafiei, and Hirokazu Shirado. Deconstructing cooperation and ostracism via multi-agent reinforcement learning, 2023.

T van der Heiden, C Salge, E Gavves, and H van Hoof. Robust multi-agent reinforcement learning with social empowerment for coordination and communication. *arXiv e-prints*, pp. arXiv–2012, 2020.