Finding the Weakest Link: Adversarial Attack against Multi-Agent Communications

Extended Abstract

Maxwell Standen
The University of Adelaide
Adelaide, Australia
DST Group
Australia
maxwell.standen@adelaide.edu.au
max.standen1@defence.gov.au

Junae Kim
DST Group
Australia
junae.kim@defence.gov.au

Claudia Szabo
The University of Adelaide
Adelaide, Australia
claudia.szabo@adelaide.edu.au

ABSTRACT

Multi-agent systems rely on communication for information sharing and action coordination, which exposes a vulnerability to attacks. We investigate single-victim communication perturbation attacks against Multi-Agent Reinforcement Learning-trained systems and propose methods that use gradient information from the Jacobian to identify which messages, agent, and timesteps are most susceptible to attack and have the greatest impact on the system. We enhance these methods with two proposed adversarial loss functions that trade-off attack success for attack impact which also create more effective perturbations. We empirically demonstrate the effectiveness of our methods against two different multi-agent communication methods in navigation, PredatorPrey, and TrafficJunction environments. Our results show that our novel message selection method achieves a similar or greater impact than random message selection across almost all tested scenarios. Our victim selection, message selection, tempo, and loss functions improve attack effectiveness in half of the thirty scenarios we tested.

KEYWORDS

adversarial attack; adversarial machine learning; multi-agent reinforcement learning; multi-agent communications; robustness; security; LEARN

ACM Reference Format:
Maxwell Standen, Junae Kim, and Claudia Szabo. 2026. Finding the Weakest Link: Adversarial Attack against Multi-Agent Communications: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/GION7107

EXTENDED ABSTRACT

Multi-agent systems have promising applications in a range of important functions such as cyber security [1, 2], but require communication for coordination and information sharing to operate in complex partially-observable environments [3]. Machine-learnt communication protocols have more bandwidth efficiency [4, 5] and noise tolerance [4, 6]. Learnt communication in real-world applications may be vulnerable to adversarial attacks. To understand and mitigate these vulnerabilities, effective attacks are required. However, existing attacks [7–10] inefficiently target messages and timesteps. Our work proposes novel methods that address these key gaps to enhance the effectiveness of attacks and improve our understanding of the vulnerability of multi-agent communications.

Communication perturbation attacks that exploit multi-agent communications by intercepting and perturbing inter-agent messages are a nascent topic of research [7–10]. The unaddressed aspects of these attacks, which may improve their effectiveness, are selecting which messages to perturb, when to attack, which agent to target, and how to craft the perturbation. Previous attacks arbitrarily select which messages will be perturbed [7–9], or target multi-agent systems with few agents so that the selection of which messages to alter is not a consideration [10]. The attack tempo, which determines when an attack occurs, can have a major impact on its effectiveness but has only been explored in attacks against single-agent systems [11–16]. Approaches to attack tempo include *counterfactual tempos* that simulate an attack [13], *learnt tempos* that train a deep RL agent to learn when to attack [13], and *threshold tempos*, that measure certain properties of an agent's logits, and attack when that metric exceeds a hyperparameter threshold [11, 12, 14–16]. The existing communication attacks either *broadcast* perturbed messages to all agents in the system [7, 9, 10] or perturb a subset of messages received by a *single victim* in the system [8]. Single-victim attacks should also consider victim selection. However, this is a gap in the current literature. Communication perturbations may be crafted using gradient-based methods if an adversary has *white-box* knowledge of the victim. The default loss function used in these attacks is called the untargeted loss and aims to minimise the probability that the agent will output the same action as it would with unperturbed input. However, changing an action may not degrade the system because of the diversity of valid solutions to Reinforcement Learning (RL) problems. To overcome this problem, many approaches learn elements of the attack [17–24] and messages [8, 9]. The limitation of these attacks is their reliance on high-compute resources and narrow application to a specific target.

To explore the worst-case vulnerability of multi-agent communication, we assume a strong attacker with white-box knowledge. The attacker aims to maximise the effectiveness of its attack and

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/GION7107

minimise detectability. We explore unaddressed aspects of single-victim communication attacks [8] by modelling the problem using a variant of the Adversarial Partially Observable Stochastic Game (APOSG) [25] that we name Single-Victim Communication Perturbation APOSG (SVCP-APOSG). The SVCP-APOSG is defined by the 14-tuple $\{I, S, \mathcal{M}, \hat{A}, T, \hat{\Omega}, \hat{O}, \hat{R}, v^m, \Theta^m, \Delta^m, \Sigma_a, \Sigma_m, k\}$. The first eight environmental elements are defined as a DEC-POMDP-Comms [9]. The six adversarial elements of the SVCP-APOSG $\{v^m, \Theta^m, \Delta^m, \Sigma_a, \Sigma_m, k\}$ determine the behaviour of a particular attack. The message-permutation function $v^m : \mathcal{M} \to \mathcal{M}$ perturbs messages received by a victim agent. $\Theta^m : S \to [0, 1]$ is the message attack tempo function that determines the probability that a timestep will be attacked. $\Delta^m : \mathbb{R}^+$ is the message attack magnitude, which is an L^2 bound on the perturbation. $\Sigma_a : S \to I$ is the victim selection function. $\Sigma_m : S \to \{I\}^k$ is the message selection function, which determines which of the received messages are perturbed. $k : \mathbb{Z}^+$ is the number of perturbed messages. For each of the adversarial aspects, the adversary aims to maximise attack effectiveness, while minimising detectability by minimising Δ^m, k, and the number of perturbed timesteps as determined by Θ^m.

We only consider agents trained with Q-learning approaches and assume that the agents are well trained and accurately estimate the Q-function. We denote an agent's logits as Q and the Q-value of a specific action a as $Q(a)$. The difference between the maximum logit and the logit of a specific action a is $Q_{\text{diff}}(a) = \max(Q) - Q(a)$.

The untargeted loss function, used in previous attacks [12, 26], uses the cross-entropy loss between an agent's logits and the argmax of the logits to encourage the output of a different action (eq. 1). However, a different action may have a similar outcome to the original action depending on the dynamics of the environment.

$$L_u = -\log\left(\frac{e^{\max(Q)}}{\sum_{a \in A} e^{Q(a)}}\right) \tag{1}$$

To improve the impact of the attack, we propose a loss function to encourage the agent to select the action that was originally considered to be the worst action by maximising the probability of that action (eq. 2), which we call the maximum loss, L_m. If successful, perturbations caused by L_m will cause the maximum impact against a well-trained agent. However, the improved attack impact likely comes with a reduced chance of attack success because actions with higher Q_{diff} values may be harder to induce.

$$L_m = \log\left(\frac{e^{\max(Q_{\text{diff}})}}{\sum_{a \in A} e^{Q(a)}}\right) \tag{2}$$

To balance attack impact and attack success, we propose an alternative loss function, that we call weighted loss, L_w (eq. 3), which maximises the probability of each action relative to its Q_{diff} by using a mean weighted cross-entropy loss across all actions.

$$L_w = \frac{1}{\sum_{a \in A} Q_{\text{diff}}(a)} \sum_{a \in A} Q_{\text{diff}}(a) \log\left(\frac{e^{Q(a)}}{\sum_{b \in A} e^{Q(b)}}\right) \tag{3}$$

To select which message, victim, and timestep to attack, we propose a Jacobian-proxy method, inspired by Jacobian-based saliency methods [27]. As the proxy for attack effectiveness, we take the absolute sum of each element of the Jacobian $J(o_i) = \nabla L(o_i)$, for observation o_i of agent i and loss function L, that corresponds to an element of the received message $o^m_{i,j}$ from agent j, which we denote as $J^m(o_i, j)$ such that our proxy $P : \Omega \times I \to \mathbb{R}^+$ is

$$P(o_i, j) = \sum_k |J^m(o_i, j)_k| \tag{4}$$

Our proposed message selection function ranks each message received by its Jacobian magnitude and selects the top-k messages. Our message selection function for the victim agent i is

$$\Sigma_m(o_i) = \text{top}_k(\{P(o_i, j); j \in I\}) \tag{5}$$

where the function top_k returns the indices of the highest k values. We use the observation function O to get the observation o_i for agent i from the state s.

Our victim selection function selects the agent with the largest total Jacobian magnitude of the top k messages is

$$\Sigma_a(\hat{o}) = \arg\max_{i \in I}\left[\sum_{j \in \Sigma_m(o_i)} P(o_i, j)\right] \tag{6}$$

using the joint observation $\hat{o} = \{o_i\}$ of all agents $i \in I$.

Our tempo function selects the timesteps where the magnitude of the Jacobian of the top k messages received from the selected victim i exceeds the hyperparameter threshold ϕ as shown in

$$\Theta(o_i) = \mathbf{1}\{[\sum_{j \in \Sigma_m(o_i)} P(o_i, j)] > \phi\} \tag{7}$$

We empirically test our attack in five environments, namely, a simple grid world navigation game [28], two variants of Predator-Prey [9] and two variants of TrafficJunction [9, 29], and against two communication methods: full observation sharing (OBS) and RIAL [4]. We measure attack effectiveness with the cumulative reward of an episode. We compare our attacks with other baseline attacks using different tempo methods, namely, CBTS [15], MMR [16], ML [16], NS [16], VL [16], and ST [11], random message selection, and the untargeted loss function. Further details about these experiments and our implementation can be found on GitHub[1].

Our Jacobian-proxy attack achieved the best impact in half of our tested scenarios, with a mean improvement of 42% over baseline attacks in those scenarios. The effectiveness was generally higher in more complex environments, at higher attack rates, and against the full observation sharing system. However, our attack was not the universal best with different baselines achieving better results against some of the tested systems. An ablation test showed that our ranked message selection method improves or maintains the performance of the attack across all systems compared to random message selection. This is particularly demonstrated in the diagonal PredatorPrey environments. An ablation test also showed that one or both of our loss functions were better than untargeted loss in most of the environments, with the exception of the orthogonal PredatorPrey. Our results suggest that our assumption that the system is well trained is inaccurate, because some of the attacked systems achieve a better result than the non-attacked system.

We have demonstrated that targeting vulnerable agents, messages, and timesteps makes attacks on communication are more effective, which highlights the importance of using strong attacks to test the robustness of multi-agent communications and emphasises the importance of developing mitigations against such attacks.

[1] https://github.com/maxstanden/weakest_link

REFERENCES

[1] Z. Tolba, N. E. H. Dehimi, S. Galland, S. Boukelloul, and D. Guassmi, "Multi-Agent Deep Reinforcement Learning Applications in Cybersecurity: Challenges and Perspectives," in *International Conference on Electrical, Computer, Telecommunication and Energy Technologies*, Dec. 2024.

[2] S. Finistrella, S. Mariani, and F. Zambonelli, "Multi-agent reinforcement learning for cybersecurity: Approaches and challenges," in *Workshop "From Objects to Agents"*, pp. 103–118, July 2024.

[3] F. A. Oliehoek and C. Amato, *A Concise Introduction to Decentralized POMDPs*. Springer International Publishing, 2016.

[4] J. Foerster, Y. Assael, N. de Freitas, and S. Whiteson, "Learning to Communicate with Deep Multi-Agent Reinforcement Learning," in *Advances in Neural Information Processing Systems*, pp. 2145–2153, Dec. 2016.

[5] S. Zhang, J. Lin, and Q. Zhang, "Succinct and robust multi-agent communication with temporal message control," in *Advances in Neural Information Processing Systems*, pp. 17271–17282, Dec. 2020.

[6] B. Freed, G. Sartoretti, J. Hu, and H. Choset, "Communication learning via backpropagation in discrete channels with unknown noise," in *AAAI Conference on Artificial Intelligence*, pp. 7160–7168, Apr. 2020.

[7] J. Tu, T. Wang, J. Wang, S. Manivasagam, M. Ren, and R. Urtasun, "Adversarial Attacks On Multi-Agent Communication," in *IEEE/CVF International Conference on Computer Vision*, pp. 7748–7757, Oct. 2021.

[8] Y. Sun, R. Zheng, P. Hassanzadeh, Y. Liang, S. Feizi, S. Ganesh, and F. Huang, "Certifiably Robust Policy Learning against Adversarial Multi-agent Communication," in *International Conference on Learning Representations*, Apr. 2023.

[9] W. Xue, W. Qiu, B. An, Z. Rabinovich, S. Obraztsova, and C. K. Yeo, "Mis-spoke or mis-lead: Achieving Robustness in Multi-Agent Communicative Reinforcement Learning," in *International Conference on Autonomous Agents and Multiagent Systems*, pp. 1418–1426, May 2022.

[10] X. Ma and W.-J. Li, "Grey-box Adversarial Attack on Communication in Multi-agent Reinforcement Learning," in *International Conference on Autonomous Agents and Multiagent Systems*, pp. 2448–2450, May 2023.

[11] Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu, and M. Sun, "Tactics of Adversarial Attack on Deep Reinforcement Learning Agents," in *International Joint Conference on Artificial Intelligence*, pp. 3756–3762, Aug. 2017.

[12] J. Kos and D. Song, "Delving into adversarial attacks on deep policies," in *International Conference on Learning Representations*, Apr. 2017.

[13] J. Sun, T. Zhang, X. Xie, L. Ma, Y. Zheng, K. Chen, and Y. Liu, "Stealthy and Efficient Adversarial Attacks against Deep Reinforcement Learning," in *AAAI Conference on Artificial Intelligence*, pp. 5883–5891, Apr. 2020.

[14] Y. Qiaoben, X. Zhou, C. Ying, and J. Zhu, "Strategically-timed State-Observation Attacks on Deep Reinforcement Learning Agents," in *ICML Workshop on Adversarial Machine Learning*, July 2021.

[15] Y. Zheng, Z. Yan, K. Chen, J. Sun, Y. Xu, and Y. Liu, "Vulnerability Assessment of Deep Reinforcement Learning Models for Power System Topology Optimization," *IEEE Transactions on Smart Grid*, vol. 12, pp. 3613–3623, Mar. 2021.

[16] R. Praveen Kumar, I. Niranjan Kumar, S. Sivasankaran, A. Mohan Vamsi, and V. Vijayaraghavan, "Critical State Detection for Adversarial Attacks in Deep Reinforcement Learning," in *IEEE International Conference on Machine Learning and Applications*, pp. 1761–1766, Dec. 2021.

[17] J. Lin, K. Dzeparoska, S. Q. Zhang, A. Leon-Garcia, and N. Papernot, "On the Robustness of Cooperative Multi-Agent Reinforcement Learning," in *IEEE Security and Privacy Workshops*, pp. 62–68, May 2020.

[18] Y. Sun, R. Zheng, Y. Liang, and F. Huang, "Who Is the Strongest Enemy? Towards Optimal and Efficient Evasion Attacks in Deep RL," in *International Conference on Learning Representations*, Apr. 2022.

[19] X. Wan, L. Zeng, and M. Sun, "Exploring the Vulnerability of Deep Reinforcement Learning-based Emergency Control for Low Carbon Power Systems," in *International Joint Conference on Artificial Intelligence*, pp. 3954–3961, July 2022.

[20] Y. Qiaoben, C. Ying, X. Zhou, H. Su, J. Zhu, and B. Zhang, "Understanding adversarial attacks on observations in deep reinforcement learning," *Science China Information Sciences*, vol. 67, pp. 1869–1919, Apr. 2024.

[21] A. Russo and A. Proutiere, "Towards Optimal Attacks on Reinforcement Learning Policies," in *American Control Conference*, pp. 4561–4567, May 2021.

[22] J. García, R. Majadas, and F. Fernández, "Learning adversarial attack policies through multi-objective reinforcement learning," *Engineering Applications of Artificial Intelligence*, vol. 96, Nov. 2020.

[23] A. Pattanaik, Z. Tang, S. Liu, G. Bommannan, and G. Chowdhary, "Robust Deep Reinforcement Learning with adversarial attacks," in *International Conference on Autonomous Agents and Multiagent Systems*, pp. 2040–2042, July 2018.

[24] H. Zhang, H. Chen, C. Xiao, B. Li, M. Liu, D. Boning, and C.-J. Hsieh, "Robust Deep Reinforcement Learning against Adversarial Perturbations on State Observations," in *Advances in Neural Information Processing Systems*, pp. 21024–21037, Dec. 2020.

[25] M. Standen, J. Kim, and C. Szabo, "Adversarial Machine Learning Attacks and Defences in Multi-Agent Reinforcement Learning," *ACM Comput. Surv.*, vol. 57, pp. 124:1–124:35, Jan. 2025.

[26] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial Attacks on Neural Network Policies," in *International Conference on Learning Representations*, Apr. 2017.

[27] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," in *IEEE European Symposium on Security and Privacy*, pp. 372–387, Mar. 2016.

[28] A. Singh, T. Jain, and S. Sukhbaatar, "Learning when to communicate at scale in multiagent cooperative and competitive tasks," in *International Conference on Learning Representations*, Apr. 2018.

[29] S. Sukhbaatar, A. Szlam, and R. Fergus, "Learning Multiagent Communication with Backpropagation," in *Advances in Neural Information Processing Systems*, pp. 2252–2260, Dec. 2016.