A Theory of Mind Approach as Test-Time Mitigation Against Emergent Adversarial Communication

Extended Abstract

Nancirose Piazza
SAIL Lab - University of New Haven
West Haven, CT, United States
npiazza@newhaven.edu

Vahid Behzadan
SAIL Lab - University of New Haven
West Haven, CT, United States
vbehzadan@newhaven.edu

ABSTRACT

Multi-Agent Systems (MAS) is the study of multi-agent interactions in a shared environment. Cooperative Multi-Agent Reinforcement Learning (CoMARL) is a learning framework that leverages cooperative mechanisms or policies that exhibit cooperative behavior. Explicitly, there are works on learning to communicate messages from CoMARL agents; however, non-cooperative agents have been shown to learn sabotage a cooperative team's performance through adversarial communication messages. To address this issue, we propose a technique which leverages local formulations of Theory-of-Mind (ToM) to distinguish exhibited cooperative behavior from non-cooperative behavior before accepting messages from any agent. We demonstrate the efficacy and feasibility of the proposed technique in empirical evaluations in a centralized training, decentralized execution (CTDE) CoMARL benchmark.

KEYWORDS

Adversarial Communication; Theory of Mind; Multi-Agent Reinforcement Learning; Test-time Defense

ACM Reference Format:
Nancirose Piazza and Vahid Behzadan. 2023. A Theory of Mind Approach as Test-Time Mitigation Against Emergent Adversarial Communication: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Multi-agent systems and modeling can be found in many prominent domains including but not limited to the deployment of automation to autonomous mobile vehicles([20]), intelligent transportation systems([12]), and financial trading portfolios([7]). Specifically, we look to Multi-Agent Reinforcement Learning (MARL) where there is long lived interest in theory and application, outlined by a comprehensive survey ([3]) and a recent paper on a selection of theories and algorithms ([19]).

The study of agent-to-agent relationships is often described as cooperative or non-cooperative. Agents that are fully cooperative, often modeled as team games, can have shared rewards. Some value-based works for fully-cooperative MARL tasks (eg. [1], [15],[16], [11], [9]) consider estimating value functions for the shared reward through some mechanism. Communication, as outlined as a pillar for cooperation intelligence, is an important channel for enabling negotiation (eg. emergence of communication through negotiation [4]), transferring information and coordination. Some Cooperative MARL (CoMARL) solutions in-cooperate graph models leverage the CTDE paradigm to learn over a communication channel. However, such mechanisms can be exploited as illustrated by [2], showing that self-interested agents can learn to impair the cooperative team's performance by transmitting adversarial messages. Learning to communicate ([6]) in MARL may be necessary for task completion, but there are challenges that follow such as adversarial communication and its possibility of cascading failure and catastrophic events (eg. impact of cascading failure in complex networks [17].) The contributions of this paper is as followed:

- We present Theory of Mind (ToM) as a cognitive mechanism for defense against adversarial communication by leveraging historic, observable neighboring actions.
- We present an belief-based trust defense test-time mechanism for homogeneous-policy agents or secured and accessible cooperative agents' policies. Such defense requires no additional training and allows agents to form decentralized ToM trust beliefs.
- We present empirical results from an adversarial communication environment, CoverageEnv.

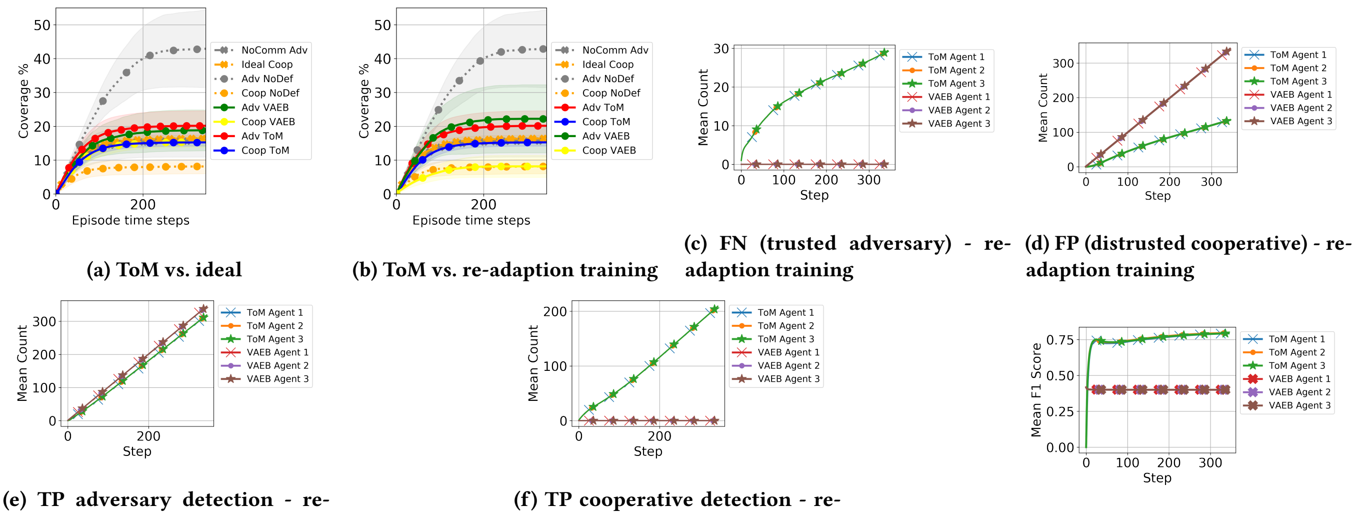
2 RELATED WORK

Modeling trustworthiness is a common direction for defenses (e.g. context aware dynamic trust using hidden markov models by identifying intent [8].) One similar work on leveraging observed past actions of other agents as an initial trust model is from Schillo et al ([13]), but unlike their consensus-based defense, we favor self-evaluation to minimize the need for external input. Message filtering methods that use a variation auto-encoder bayes model ([10]) is similar to our approach; however, instead of directly learning the existing cooperative message distribution, our belief state is dependent on observable behavior during episodic execution. Another relevant work ([5]) uses consensus-based decisions to detect and eliminate adversarial robots in a flocking problem. While their approach is more reactant to adversarial detection and dependent on an uncompromising majority, their method uses centralized trust evaluation, contrast and complimentary to our decentralized method.

3 THEORY OF MIND UNDER TEST-TIME & SELF-REVEALING BEHAVIORS

Theory of Mind (ToM) is the rationalization of other agents to some belief state, often through Bayesian modeling. Bayesian models

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom. © 2023 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.



(a) ToM vs. ideal

(b) ToM vs. re-adaption training

(c) FN (trusted adversary) - re-adaption training

(d) FP (distrusted cooperative) - re-adaption training

(e) TP adversary detection - re-adaption training

(f) TP cooperative detection - re-adaption training

(g) f1 score - re-adaption training

have belief states that approximate probabilistic frequencies for sequences of outcomes. ToM is well discussed in AI; however, the explicit modeling of other agents' belief states can require additional burdensome computational load that may not be feasible in large-scaled systems such as swarm robotics[14]. However, Mean-Field MARL (MF-MARL [18]) can be a substituting framework.

We emphasize that ToM should be influential to a defense's design and can take on many forms. We approach ToM from the reasoning '*if I were them: what message would I transmit to cooperate, and whether their observed behavior is consistent with my expectations?*' Many defenses rely on out-of-distribution detection (e.g. VAEB message filter defense [10].) However, these defenses require additional training outside of test-time. Our ToM belief defense, similar to human credit history, use the history of observed actions to update the trust probability. This defense is directed towards test-time because an agent's belief is only relevant to the current episode and episode's history. Furthermore, we believe defenses should be layered to minimize possible damage from adversarial communication.

We say a communicated message can be described as *self-revealing* or *self-committing / consistent* when the action stated in the message can be observed. A message that is self-revealing reveals a cooperative or non-cooperative behavior. A message is self-committing when it is consistent with the previous transmitted message. Similar to [10] where some empirical belief determines trustworthiness, we consider that if other agents' messages are not self-committing, then communication messages should not impact an agent's decision. We use two algorithms: trust count(1) and consensus update(2) (algorithms details omitted) and can be summarized as the following: all agents are set to probability 1.0 by default and after each environment step, all agents reevaluate their trust in other agents based on observed actions from the previous transmitted message.

4 EXPERIMENTAL RESULTS

We use the adversarial communication repository provided by [2], training under the default settings with deterministic action selection. The team task is to maximize its coverage over an un-visited grid. The agents access a central communication channel and use an Aggregation Graph Neural Network (AGNN) to aggregate their inputs. We trained under default settings with $N = 5$ agents for 6 million (mil) timesteps, 6mil timesteps for the self-interested policy with fixed cooperative policy. The re-adaption training continued with the fixed self-interested policy for 6mil timesteps. The evaluation setup has one self-interested agent (Agent0) and three cooperative agents (Agent1, Agent2, Agent3). Code is publicly available[1].

We present Figure 1a as an ideal baseline comparison (VAEB) with a learning step multiplier of 3.7 and the secondary baseline of no defense (NoDef). Then, we show in Figure 1b, the VAEB is ineffective when cooperative agents re-adapted their messages in the presence of self-interested messages. In support of our hypothesis, we see ToM's performance is not impacted by the re-adaption training. We provide Figure 1d and Figure 1c per agent which is the false negative (FN) and false positive (FP) mean count respectively. We see that according to the true positives (TP) Figures 1f and 1e, our defense by implicit design is prone to distrust. We also include Figure 1g which is the mean f1 score after re-adaption training.

5 CONCLUSION

In this paper, we have introduced Theory of Mind (ToM) as a rationalization of other agents' behaviors through observed actions to determine trust beliefs of other agents' communicated messages. Our defense is for test-time and requires no additional training or retraining, allowing this defense to be easily layered among other defenses. We study a multi-agent environment where adversarial communication emerges and demonstrate the usage of the defense, comparing the average cooperative team's performance with our defense in comparison to the average cooperative team performance without a defense and cooperative team's performance with a variational auto-encoder bayes defense baseline.

6 CITATIONS AND REFERENCES

REFERENCES

[1] Adrian K. Agogino and Kagan Tumer. 2004. Unifying Temporal and Structural Credit Assignment Problems. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 2* (New York, New York) *(AAMAS '04)*. IEEE Computer Society, USA, 980–987.

[1] Piazza, N., Behzadan, V., A Theory of Mind Approach as Test-Time Mitigation Against Emergent Adversarial Communication, (2023), GitHub repository, https://github.com/UNHSAILLab/ToM_Against_AdvComm

[2] Jan Blumenkamp and Amanda Prorok. 2020. The Emergence of Adversarial Communication in Multi-Agent Reinforcement Learning. arXiv:2008.02616 [cs.RO]

[3] Lucian Busoniu, Robert Babuska, and Bart De Schutter. 2008. A Comprehensive Survey of Multiagent Reinforcement Learning. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 38 (04 2008), 156 – 172. https://doi.org/10.1109/TSMCC.2007.913919

[4] Kris Cao, Angeliki Lazaridou, Marc Lanctot, Joel Z Leibo, Karl Tuyls, and Stephen Clark. 2018. Emergent Communication through Negotiation. https://doi.org/10.48550/ARXIV.1804.03980

[5] Mingxi Cheng, Chenzhong Yin, Junyao Zhang, Shahin Nazarian, Jyotirmoy Deshmukh, and Paul Bogdan. 2021. A General Trust Framework for Multi-Agent Systems. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems* (Virtual Event, United Kingdom) *(AAMAS '21)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 332–340.

[6] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to Communicate with Deep Multi-Agent Reinforcement Learning. arXiv:1605.06676 [cs.AI]

[7] Jinho Lee, Raehyun Kim, Seok-Won Yi, and Jaewoo Kang. 2020. MAPS: Multi-Agent reinforcement learning-based Portfolio management System. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (Jul 2020). https://doi.org/10.24963/ijcai.2020/623

[8] Xin Liu and Anwitaman Datta. 2012. Modeling Context Aware Dynamic Trust Using Hidden Markov Model. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (Toronto, Ontario, Canada) *(AAAI'12)*. AAAI Press, 1938–1944.

[9] David Mguni, Joel Jennings, Emilio Sison, Sergio Valcarcel Macua, Sofia Ceppi, and Enrique Munoz de Cote. 2019. Coordinating the Crowd: Inducing Desirable Equilibria in Non-Cooperative Systems. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 386–394.

[10] Rupert Mitchell, Jan Blumenkamp, and Amanda Prorok. 2020. Gaussian Process Based Message Filtering for Robust Multi-Agent Cooperation in the Presence of Adversarial Communication. arXiv:2012.00508 [cs.RO]

[11] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. https://doi.org/10.48550/ARXIV.1803.11485

[12] Sergey Satunin and Eduard Babkin. 2014. A Multi-Agent Approach to Intelligent Transportation Systems Modeling with Combinatorial Auctions. *Expert Syst. Appl.* 41, 15 (Nov. 2014), 6622–6633. https://doi.org/10.1016/j.eswa.2014.05.015

[13] Michael Schillo, Petra Funk, and Michael Rovatsos. 2000. Using trust for detecting deceitful agents in artificial societies. *Applied Artificial Intelligence* 14, 8 (2000), 825–848. https://doi.org/10.1080/08839510050127579 arXiv:https://doi.org/10.1080/08839510050127579

[14] Melanie Schranz, Martina Umlauft, Micha Sende, and Wilfried Elmenreich. 2020. Swarm Robotic Behaviors and Current Applications. *Frontiers in Robotics and AI* 7 (2020).

[15] Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. 2019. QTRAN: Learning to Factorize with Transformation for Cooperative Multi-Agent Reinforcement Learning. https://doi.org/10.48550/ARXIV.1905.05408

[16] 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. 2017. Value-Decomposition Networks For Cooperative Multi-Agent Learning. https://doi.org/10.48550/ARXIV.1706.05296

[17] Lucas D Valdez, Louis Shekhtman, Cristian E La Rocca, Xin Zhang, Sergey V Buldyrev, Paul A Trunfio, Lidia A Braunstein, and Shlomo Havlin. 2020. Cascading failures in complex networks. *Journal of Complex Networks* 8, 2 (05 2020). https://doi.org/10.1093/comnet/cnaa013 arXiv:https://academic.oup.com/comnet/article-pdf/8/2/cnaa013/33582729/cnaa013.pdf cnaa013.

[18] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. 2018. Mean Field Multi-Agent Reinforcement Learning. https://doi.org/10.48550/ARXIV.1802.05438

[19] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. 2019. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. https://doi.org/10.48550/ARXIV.1911.10635

[20] Ming Zhou, Jun Luo, Julian Villella, Yaodong Yang, David Rusu, Jiayu Miao, Weinan Zhang, Montgomery Alban, Iman Fadakar, Zheng Chen, Aurora Chongxi Huang, Ying Wen, Kimia Hassanzadeh, Daniel Graves, Dong Chen, Zhengbang Zhu, Nhat Nguyen, Mohamed Elsayed, Kun Shao, Sanjeevan Ahilan, Baokuan Zhang, Jiannan Wu, Zhengang Fu, Kasra Rezaee, Peyman Yadmellat, Mohsen Rohani, Nicolas Perez Nieves, Yihan Ni, Seyedershad Banijamali, Alexander Cowen Rivers, Zheng Tian, Daniel Palenicek, Haitham bou Ammar, Hongbo Zhang, Wulong Liu, Jianye Hao, and Jun Wang. 2020. SMARTS: Scalable Multi-Agent Reinforcement Learning Training School for Autonomous Driving. arXiv:2010.09776 [cs.MA]