What are you saying? Explaining communication in multi-agent reinforcement learning

Daniele Meli*[1], Cristian Morasso*[1], Alberto Castellini[1] and Alessandro Farinelli[1]

Abstract

Communication in Multi-Agent Reinforcement Learning (MARL) has the potential to improve the performance of cooperating agents, especially in complex robotic domains under partial observability. However, a transparent interpretation of the learned communication policy is crucial for trustability and safety. In this paper, we use tools from explainable artificial intelligence to investigate the impact of communication in a benchmark MARL setting, involving collision avoidance among multiple agents. Our preliminary tests show that the role of communication cannot be evidenced solely by looking at the state-action policy map; instead, causal discovery on the state and communication spaces highlights the *latent behavioural* impact of messages passed among agents, indirectly affecting the actual actions for more efficient collision avoidance.

Keywords

Multi-Agent Reinforcement Learning, Communication in MARL, Explainable AI, Causal Discovery

1. Introduction

Reinforcement Learning (RL) is an established methodology to achieve agent autonomy in complex scenarios, including robotics [1]. Indeed, given the model of interaction with the environment (the *transition map*) and the *reward* attained in consequence of executing specific *actions* in particular conditions (*states*), a RL algorithm automatically learns the best *policy*, i.e., state-action map, to fulfill the task at the highest cumulative reward (*return*). The advent of Deep Neural Networks (DNNs) has enhanced the learning of complex policies for the most challenging tasks, shifting towards Deep RL (DRL). This has also paved the way towards DRL applied in multi-agent settings (*Multi-Agent RL, MARL*) [2, 3], where the best task strategy does not solely depend on the individual policies, but rather on the *inter-agent coordination*. Inspired from biology and human behaviour, an emerging problem in MARL is the *inter-agent communication* [4, 5], i.e., learning and deploying an efficient mechanism for information sharing among agents, with the goal to enhance coordination and improve the individual and global task performance. While several approaches have been studied and compared, one fundamental question rises when deploying communicating MARL agents in the real world, e.g., on real robots interacting with humans: *what is the meaning of the learned communication policy?* Answering this question is fundamental for the transparency and interpretability of the MARL application, which in

11th Italian Workshop on Artificial Intelligence and Robotics (AIRO 2024).

**Authors contributed equally.*

*Corresponding author.

daniele.meli@univr.it (D. Meli*); cristian.morasso@univr.it (C. Morasso*); alberto.castellini@univr.it (A. Castellini); alessandro.farinelli@univr.it (A. Farinelli)

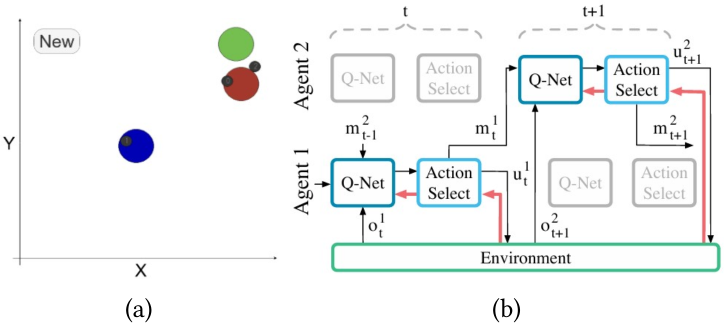


Figure 1: a) The simple spread domain, with agents 0, 1 and 2 moving towards black dot targets; b) the RIAL architecture for communication in MARL (in [4], u and o are actions and state, resp.).

turn are essential for trustability and social acceptance, as well for proper monitoring of the autonomous systems [6].

In this paper, we address the problem of *explaining MARL communication.* We consider a benchmark domain for MARL, *simple spread*[1], where 3 robotic agents must coordinate to reach 3 separate targets (Figure 1a). We design different communication protocols, both hardcoded and learned in the MARL pipeline. We then investigate the impact of different communication strategies on MARL performance, both from a quantitative perspective (i.e., evaluating the achieved return) and exploiting *eXplainable Artificial Intelligence (XAI)* techniques, including relevant feature analysis via Integrated Gradients (IG) and causal discovery, already employed for complex system explanation and monitoring [7, 8]. In this way, we want to analyze the meaning of messages passed among agents, and how specific parts of information affect the overall performance observed by standard RL metrics, as the return.

2. Background

We now provide the relevant background about MARL and related communication strategies, and XAI methods adopted in this paper, i.e., IG and causal discovery.

2.1. Multi-Agent Reinforcement Learning

We frame the problem of single-agent RL as a Markov Decision Process (MDP) $\langle S, A, T, R, \gamma \rangle$, where S is the state space; A is the action space; $T : S \times A \rightarrow S$ is the transition function mapping state s_t and action a_t at time t to the state at $t+1$ (assuming a discretization of the time dimension); $\gamma \in \mathbb{R}$ is the discount factor; $R : S \times A \times S \rightarrow \mathbb{R}$ is the reward map, assigning a real number to incentivize / penalize the agent for executing a_t at s_t, with a corresponding next state determined from T. The goal of RL is to compute a policy map $\pi : S \rightarrow A$, prescribing the best a_t to be performed at s_t, in order to maximize the expected value of the return $\sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t, s_t')$.

In the MARL setting with N agents, we assume that each agent i has access to a state space S^i such that $\langle S^1, \dots, S^N \rangle = A$; similarly, each agent can pick an action from A^i such that

[1] https://pettingzoo.farama.org/environments/mpe/simple_spread/

$\langle A^1, \dots, A^N \rangle = A$; R, γ, T remain unchanged. In this way, each agent has *partial observabiliy* of the environment, but still all agents should coordinate to compute the best global policy towards the maximization of the cumulative shared reward. It is then fundamental for the agents to communicate; however, it is highly domain-dependent, and in general far from trivial, to design the *messages and methodologies* for effective communication [4]. An interesting approach is then to *learn* the best *communication policy* $\pi_m : S \rightarrow M$, M being the set of available communication actions m_t at time t.

2.2. Explainable AI

XAI aims at providing explanations about AI algorithms to a targeted audience, according to their needs and knowledge in relation to a specific domain of application [9]. In this paper, we focus on two main XAI methodologies: *causal discovery* from time series and *Integrated Gradients (IG)* to explain the input / output relations in DNNs.

2.2.1. Causal discovery

Consider a multi-variate time series $\boldsymbol{X} = \{X^j\}_{j=1,\dots N}$ composed of N time series, and denote as $X^j = \{x^j_1, \dots x^j_T\}$ the sequence of observations of variable X^j for T time steps. The goal of causal discovery is to identify *directed causal links* between variables in $\boldsymbol{X}$. More specifically, causal links are determined according to the measure of *Conditional Mutual Information (CMI)*, which is defined for random variables X, Y, Z as:

$$I(X;Y|Z) = \iiint p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} dxdydz$$

where $p(\cdot|\cdot)$ and $p(\cdot,\cdot)$ denote the conditional and joint probability distributions, respectively. From the above definiton, it can be easily shown that variables X and Y are *conditionally independent* under Z, denoted as $X \perp\!\!\!\perp Y|Z$, iff $I(X;Y|Z) = 0$. In other words, X and Y *have no mutual causal influence*, assuming that Z holds. On the other hand, X and Y *may conditionally depend* on Z.

2.2.2. Integrated Gradients (IG)

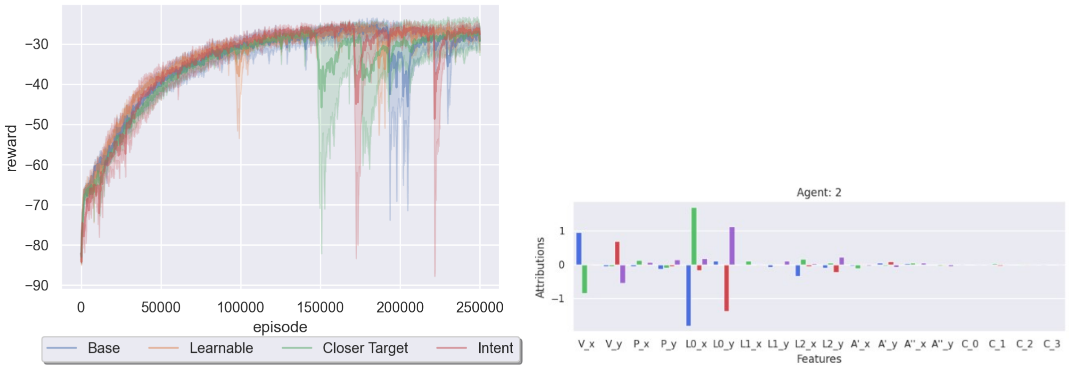
Consider a DNN $f_\theta : \mathbb{R}^n \rightarrow [0,1]^m$, θ being the set of parameters, n, m the input and output dimensions, respectively. Let $x \in \mathbb{R}^n$ be a generic input to f_θ, and $x' \in \mathbb{R}^n$ be a *baseline input*, s.t. $f_\theta(x') = \frac{1}{m} \cdot \mathbf{1}^m$ (i.e., a neutral input for the DNN).

IG [10] is defined as:

$$IG(x) = \left(x - x^{'}\right) \cdot \sum_{k=1}^{q} \frac{\partial f_\theta\left(x' + \frac{k}{q} \cdot (x - x')\right)}{\partial x} \cdot \frac{1}{q}$$

which is the path integral[2] of the gradients of f_θ along the straight line (in $\mathbb{R}^n$) from x to x'. For each input dimension $i < n$, IG measures its *attribution* to $f_\theta(x)$, i.e., the contribution x_i in determining $f_\theta(x)$.

[2] Approximated via q discretization steps.



(a) (b)

Figure 2: a) The return under different communication protocols; b) IG results with *Learnable* protocol for actions left, right, down, up.

3. Methodology

We consider a MARL setting based on deep deterministic policy gradient [11]. The simple spread domain (Figure 1a) is described as a MDP, where each agent observes the following continuous state-variables: *i)* $x - y$ velocities $P_v = \langle P_{vx}, P_{vy} \rangle$; *ii)* coordinates $P = \langle P_x, P_y \rangle$; *iii)* landmark (target) coordinates $L = \langle l_x, l_y \rangle$; *iv)* coordinates of other agents $A = \langle a_x, a_y \rangle$. The continuous action space, for each agent, consists of 4 directional forces in $[0, 1]$, resulting in up / left / down / right motions.

We assume the Reinforced Inter-Agent Learning (RIAL) [4] communication protocol is employed (Figure 1b). In RIAL, the communication action m_t^i from agent i at time t is passed to all other agents as an additional observation (state) input. Hence, we can define the MDP $\langle \bar{S}, \bar{A}, T, R, \gamma \rangle$, where $\bar{S} = S \bigcup M$ and $\bar{A} = A \bigcup M$.

Our goal is to investigate the meaning of the communication policy π_m, i.e., the impact of the communication actions on MARL performance. To this aim, we first apply IG to the trained policy network π, in order to identify the impact of m_t^i on $a_t^j, j \neq i$. This quantifies the *direct impact* of the communication strategy on MARL.

Then, we discover causal relations between time series from $\bar{S}$, generated applying the trained policy π in inference. This study evidences the *latent impact* of the communication strategy, i.e., how π_m influences the general behaviour of the agent (e.g., its intentions), rather than merely its actions. We adopt state-of-the-art PCMCI+ algorithm [12] for causal discovery, which is sound and complete under the assumptions of causal Markovianity and sufficiency, and faithfulness.

4. Experiments

We consider 3 different communication policies: i) *Closest Target (CT)*, where m_t^i is the closest landmark to i-th agent; ii) *Intent*, where m_t^i is the action selected by agent i; iii) *Learnable*, where π_m is trained together with π, resulting in $m_t^i \in \mathbb{R}^{N-1}$ (in our case $\mathbb{R}^{N-1} = \mathbb{R}^2$).

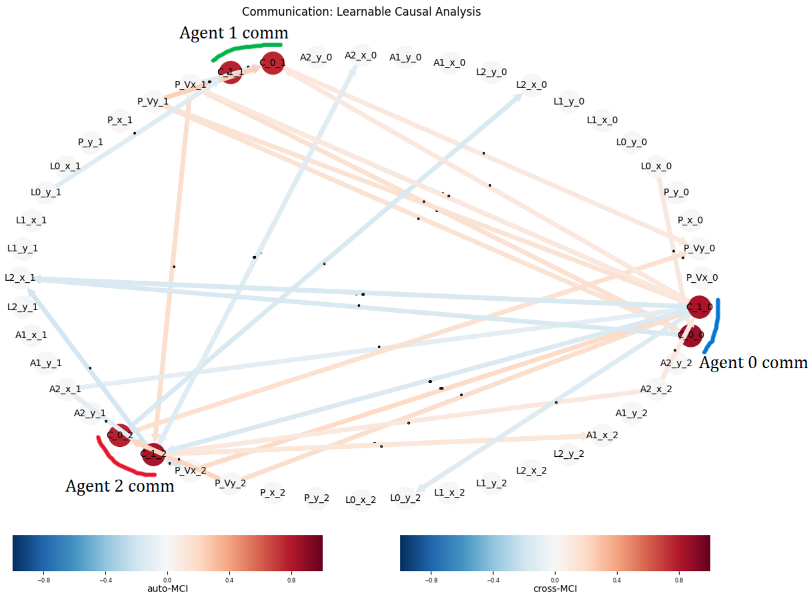


Figure 3: Causal graph for the *Learnable* communication protocol (only links involving m_t are reported for simplicity).

We first report the training performance (over 5 random seeds) with the different communication strategies in Figure 2a, where *Base* denotes no communication, i.e., $\pi_m \equiv 0$. We notice that all MARL policies have large negative drops after the stabilization of the training process. This derives from the non-stationarity of MARL, under the assumption of partial observability from each agent. However, the *Learnable* protocol results in the smallest negative peak in the return trend. This suggests that each agent can learn to communicate useful messages to the others, resulting in more robust performance of MARL.

We first try to understand the role of communication in the *Learnable* protocol via IG analysis. Figure 2b shows the attributions of state features for all actions (see the legend in the caption; we only report one agent for compactness). It is evident that the communication actions ($C_{1,\dots 3}$) do not have significant attribution on the actions chosen by the agent, which in turn depend mostly on its velocity $V_{x,y}$ and the position of target L_0.

We then employ causal discovery with the *Learnable* protocol to show latent connections between variables in $\bar{S}$. Figure 3 shows the causal graph derived from PCMCI+[3], where nodes are variables, edges denote their causal relations, and the color map represents the corresponding CMI value. We observe that a causal link is identified among communication variables of different agents, denoting the tight interaction strategy between agents. Interestingly, the communications between agents 0 and 1 affect each other's position and velocity, as it is visible in Figure 1a, which shows that the two agents decide to reach two targets close to each other, hence they learn to communicate to safely avoid collisions.

[3] We employ the implementation from https://github.com/jakobrunge/tigramite

5. Conclusion

In this paper, we exploited different XAI strategies, particularly integrated gradients and causal discovery, to explain the role of communication in MARL with DNNs. We studied a benchmark multi-robot navigation problem, the simple-spread domain. Among different communication protocols, including pre-defined messages based on prior task knowledge, the agents achieve the best performance when they can *learn* the communication protocol, reducing the negative impact of non-stationarity in MARL. Under the *Learnable* communication protocol, IG detects state-action relations in the policy network, but does not highlight an impact of communication messages. On the contrary, causal discovery evidences the role of communication among close agents in the map, in order to exchange mutual position and velocity information and avoid collisions. In the future, we will extend our study to more complex and real-world robotic MARL domains.

References

[1] H. Jiang, H. Wang, W.-Y. Yau, K.-W. Wan, A brief survey: Deep reinforcement learning in mobile robot navigation, in: 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), IEEE, 2020, pp. 592–597.

[2] S. V. Albrecht, F. Christianos, L. Schäfer, Multi-Agent Reinforcement Learning: Foundations and Modern Approaches, MIT Press, 2024. URL: https://www.marl-book.com.

[3] F. Bianchi, E. Zorzi, A. Castellini, T. D. Simão, M. T. J. Spaan, A. Farinelli, Scalable safe policy improvement for factored multi-agent mdps, in: Proceedings of the 41st International Conference on Machine Learning (ICML 2024), PMLR 235, PMLR, 2024, pp. 3952–3973.

[4] J. Foerster, I. A. Assael, N. De Freitas, S. Whiteson, Learning to communicate with deep multi-agent reinforcement learning, Advances in neural information processing systems 29 (2016).

[5] J. Sheng, X. Wang, B. Jin, J. Yan, W. Li, T.-H. Chang, J. Wang, H. Zha, Learning structured communication for multi-agent reinforcement learning, Autonomous Agents and Multi-Agent Systems 36 (2022) 50.

[6] G. A. Vouros, Explainable deep reinforcement learning: state of the art and challenges, ACM Computing Surveys 55 (2022) 1–39.

[7] J. Runge, Causal network reconstruction from time series: From theoretical assumptions to practical estimation, Chaos: An Interdisciplinary Journal of Nonlinear Science 28 (2018).

[8] D. Meli, Explainable online unsupervised anomaly detection for cyber-physical systems via causal discovery from time series*, in: 2024 IEEE 20th International Conference on Automation Science and Engineering (CASE), 2024, pp. 4120–4125.

[9] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, et al., Explainable ai (xai): Core ideas, techniques, and solutions, ACM Computing Surveys 55 (2023) 1–33.

[10] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.

[11] T. Lillicrap, Continuous control with deep reinforcement learning, arXiv preprint arXiv:1509.02971 (2015).

[12] J. Runge, Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets, in: Conference on Uncertainty in Artificial Intelligence, Pmlr, 2020, pp. 1388–1397.