Studying the joint role of partial observability and channel reliability in emergent communication

Valentin Villecroze, Clément Moulin-Frier

Studying the joint role of partial observability and channel reliability in emergent communication

Valentin Villecroze
Flowers Team
Inria
Bordeaux, France
v.villecroze@gmail.com

Clément Moulin-Frier
Flowers Team
Inria
Bordeaux, France
clement.moulin-frier@inria.fr

***Abstract*—Multi-Agent Reinforcement Learning (MARL) provides a powerful conceptual and computational framework for modeling emergent communication as a way to solve complex problems in sequential environments. However, despite the recent advances in this field, there is still a need to better understand the role of heterogeneous factors, e.g. partial observability and channel reliability, in the emergence of communication systems. An important step has recently been done in this direction by proposing new information-theoretic measures of emergent communication. As of yet, very few contributions have taken advantage of these new measures to perform detailed quantitative studies analyzing how different environmental and cognitive factors can foster the emergence of communication systems. This work quantitatively measures the joint role of partial observability and channel reliability in the emergence of communication systems. To this end, we performed experiments in a simulated multi-agent grid-world environment where agents learn how to solve different cooperative tasks through MARL.**

***Index Terms*—Multi-Agent Reinforcement Learning, Emergent communication, Partial observability, Causal influence**

I. Introduction

The ability to use language to express complex thoughts and coordinate with other humans in a wide variety of scenarios is an inherent part of human intelligence. As such, there is a considerable interest in the Artificial Intelligence (AI) community to create artificial agents that efficiently use language. While there have been numerous breakthroughs in natural language processing over the last decade [1] [2], most existing contributions aim at capturing structural aspects and statistical regularities of human language from massive static datasets of text or speech [3]. This approach completely forgoes the interactive and functional nature of language, leading to limited results on interactive tasks such as chatbots or any dynamic machine-human dialogue [4]. Moreover, from the cognitive sciences side, it does not allow the study of how animal communication systems emerge during the species' evolution and the development of individuals.

In contrast, the framework of Multi-Agent Reinforcement Learning (MARL) can help ground artificial agents in concrete cooperative tasks [5], where communication systems can spontaneously emerge as a way to optimize the realization of complex goals. MARL is indeed a subcategory of reinforcement learning (RL) where several artificial agents learn through interactions with each other and their environment to maximize their individual reward. Spearheaded by the recent successes of Deep Reinforcement Learning [6], studying communication in MARL has been getting a lot of attention over the last few years [7] [8] [9] [10] [11] [12]. Agent built out of deep neural networks are being used successfully in elaborate simulations that can foster the emergence of complex behaviors.

However, despite the recent advances in MARL, there is still a need to better understand how various heterogeneous factors influence the emergence of communication systems. An important step has recently been done in this direction by demonstrating the pitfalls of traditional measures of communication and by proposing new measures supposed to be more reliable [13]. As of yet, very few contributions have taken advantage of these new measures to perform detailed quantitative studies analyzing how different environmental and cognitive factors can foster the emergence of communication systems.

This work aims at quantitatively measuring the role of partial observability in the emergence of communication systems. To this end, we perform experiments in a simulated multi-agent grid-world environment where agents learn how to solve different cooperative tasks through MARL. We study how the level of observability provided to the agents influence information-theoretic measures of emergent communication. To our knowledge, this is the first contribution of the field attempting to analyze quantitatively this factor using recent measures of emergent communication. This preliminary work also takes place in a larger scale project of Inria's Flowers team seeking to join the historic studies of the team on language with the recent advances on MARL [14] [15] [16].

We will first briefly review some key related works in Section II, before introducing some notations and formalism in Section III. In section IV, we will then present the setup we use for the experiments explained in Section V. Lastly, we will analyse the contribution of this work and possible future works in Section VI.

II. Related works

When studying language through the scope of Reinforcement Learning (RL), an important distinction has to be made between language-based RL and emergent communication in MARL.

Language-based RL here refers to the the grounding of natural language in an artificial agent's environment, through textual goals for example [17].This aims at giving the artificial agent the generalization and planning abilities that language brings to humans [18] [19]. Language-based RL requires human-annotated data or an already trained "teacher" agent, however, in order to bring natural language in the environment.

When considering emergent communication (see [12] for a comprehensive survey) on the other hand, language emerges through the interactions of artificial agents. The resulting communication system is not influenced by natural language, and instead emerges as a way of solving complex cooperative tasks in a multi-agent environment. This *utilitarian* view of communication is conceptualised in [5].

In most works that follow this paradigm, the multi-agent environment is a cooperative *referential game*, where one agent has some private information about a public observation (for example which of two images is the target image) and must communicate it accurately to the other agent that then has to act accordingly (e.g. select the right image among the two) [20] [21] [22].

However, there is a long-standing theory in cognitive sciences that language understanding is grounded in one's experience of the world and should not be secluded from the perceptual and motor systems ([23]). Therefore, some works introduce agents that can move around simulated worlds and perceive their surroundings with a limited field of view [7] [8] [10] [11]. Such agents are said to be *situated* [24].

We consider situated agents in our work as we want to evaluate the influence of partial observability, which is best illustrated by the limited field of view of such agents.

For more biological plausibility, we also consider our agents to be independent learners, meaning we use fully *decentralized* training (see [25] for a recent survey of MARL algorithms). This is known to perform poorly in practice due to the non-stationnary multiple agent learning simultaneously induces in the environment [26] and the current best performing algorithms use *centralized* training instead [27]. However, recent works such as [10] and [11] have shown successful convergence by adding inductive biases based on information-theoretic measures [13].

III. Backgrounds and methods

A. Markov game

We consider an N-player partially observable Markov game [28], which is a multi-agent extension of Markov decision processes. A Markov game for N players is defined by a set of states $\mathcal{S}$, action sets for each player $\mathcal{A}_1, ..., \mathcal{A}_N$ and a state transition function $\mathcal{T} : \mathcal{S} \times \boldsymbol{\mathcal{A}} \rightarrow \Pi(\mathcal{S})$ where $\boldsymbol{\mathcal{A}} = \mathcal{A}_1 \times ... \times \mathcal{A}_N$, similarly to a MDP, but also by sets of observations $\mathcal{O}_1, ..., \mathcal{O}_N$ for each player.

At each time step t, each agent (or player) i receives an observation o_t^i from the corresponding observation function $\mathbf{o}_i : \mathcal{S} \rightarrow \mathcal{O}_i$. Each agent then samples an action $a_t^i \in \mathcal{A}_i$ from its stochastic action policy $\pi^i : \mathcal{O}_i \times \mathcal{A}_i \rightarrow [0, 1]$. This policy could also be a function of the agent's previous observations $(o_0^i, ..., o_t^i)$ but we use single time step policies here for simplicity.

Our setting is fully-cooperative, meaning that each agent receives the same reward at each time step that we denote r_t, given by the reward function $R : \mathcal{S} \rightarrow \mathbb{R}$. The policies are optimized to maximize the expected discounted total reward J (the sum of the reward at each future time steps multiplied by a discount factor), defined as:

$$J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}, \mathcal{T}} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right] \tag{1}$$

where $\boldsymbol{\pi} = \left\{ \pi^1, \ldots, \pi^N \right\}$ is the joint action policy and $\gamma < 1$ is the discount factor.

B. Training algorithm

In practice, agents and their policies are deep neural networks that we seek to train through reinforcement learning. Each policy π^i thus depends on a set of parameters θ^i (the weights of the neural networks). In order to update those parameters, several algorithms exist in the literature, but most of them use centralized training with parameter sharing [25]. As we want a fully decentralized learning, we chose to train each agent using an *actor-critic* algorithm [29] called *asynchronous advantage actor-critic* (A3C) [30]. Although this algorithm is usually used for single-agent RL, it was used in a similar manner in [10] and [11].

IV. Setup

A. Environment

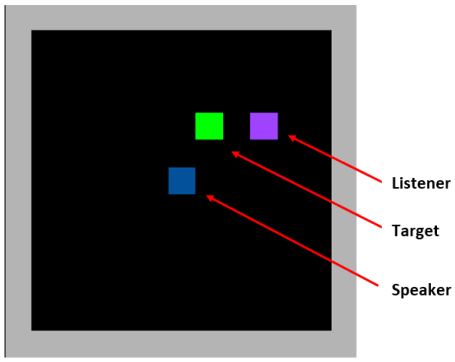
The Markov game we choose to use is a 2-player grid-world game with communication. Two agents are embodied in a 11×11 grid and must find a target spawned randomly inside the grid while communicating with one another (see Figure 1a). Each agent's action a_t^i is thus a couple composed of an environment action $e_t^i \in \mathcal{A}_i^e$ to move around the grid and a communication action (a message) $m_t^i \in \mathcal{A}_i^m$, where $\mathcal{A}_i^e \subset \{$*up, down, left, right, stay*$\}$ and $A_i^m \subset \{0, 1, ..., k_{\textit{vocabulary}} - 1\}$

In order to have communication between the two agents hold a real advantage, we make it so that only one agent can actually pick up the target. This agent, thereafter called *listener*, must pick up the target by moving onto it, while the other agent, called *speaker*, has to communicate valuable information to help him. Once the listener picks up the target, both him and the speaker receives the same positive reward and a new target spawns randomly on the grid. An episode ends after $T = 300$ time steps.

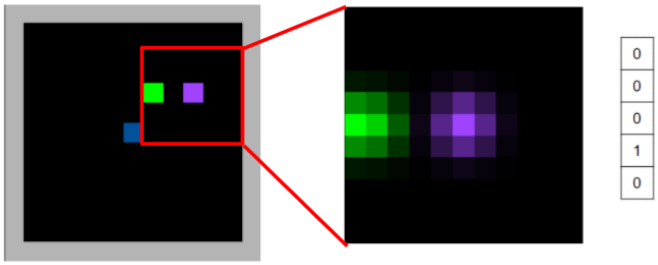
The exact parameterization of the environment will be precised for each individual experiment in Section V.

B. Agents

The agents receives an egocentric view of their surroundings (an RGB matrix of size $v_{\textit{size}}^i \times v_{\textit{size}}^i$) as well as a one-hot encoding of the discrete message sent by the other agent at the previous time-step as their observation o_t^i (see Figure 1b).



(a)



(b)

Fig. 1: (a) Grid world (top view), (b) Visual partial observation received by the listener, and one-hot message sent from the speaker to the listener.

The view size v^i_{size} of each agent is a changeable parameter, as we want to evaluate its impact on the emergence of the communication. In order to have the same input size regardless of the agent's view size, the visual input is resized using linear interpolation to always have the same dimensions (this is the reason the listener's visual observation is blurry in Figure 1b). In most of our experiments, the speaker will have a greater view size than the listener, giving the latter an incentive to listen to him.

The network architecture used to model the listener agent can be seen in figure 2. It is composed of three convolutional layers and a fully connected layer, followed by a couple of linear heads, that outputs the policy' probabilities (the probability to choose each action from the action set) and the value function estimate. Only the visual input is processed by the convolutional layers and the resulting features are concatenated to the one-hot encoding of the message sent at the previous time step before the linear layer.

The speaker's architecture on the other hand changes throughout the experiments and will be explain in more details in the next section.

V. Experiments

In this section, we expose our experiments aiming at quantitatively measuring the joint influence of partial observability, i.e. the agents' view size, and channel reliability, i.e. noise,

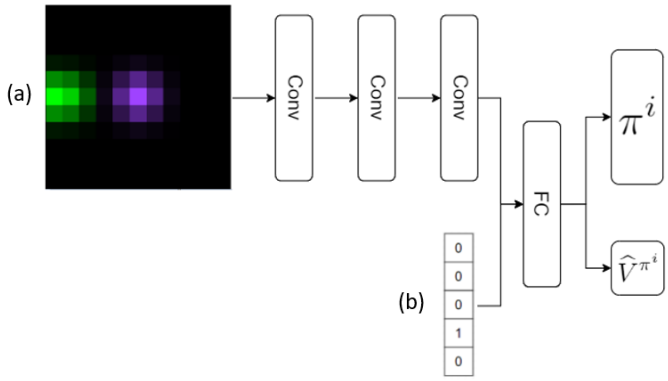


Fig. 2: Standard listener agent architecture, where the visual input (a) goes through convolutional layers and is concatenated with the one-hot encoding of the message (b) before the fully-connected (FC) layer.

on the emergence of communication. We start off with a naive attempt set in a complex setting, before making drastic assumptions to ensure convergence and then progressively relaxing these assumptions.

A. Can communication emerge with no biases?

Our first experiment aims to study whether communication could emerge only from the sparse reward of the environment. We set a speaker that can observe the whole grid ($v^{speaker}_{size} = 11$) but is unable to move ($\mathcal{A}^e_0 = \emptyset$). It can only send messages to a listener that is able to move around the grid to pick up the target but cannot see his surroundings ($v^{listener}_{size} = 1$, he only sees the tile he is standing on, i.e. himself). Communication between the two agents is therefore necessary in order for the agents to solve the task efficiently.

Even after training for $300,000,000$ steps however, i.e $1,000,000$ episodes, no convergence is seen, and the listener still moves around randomly to get to the target. This is due to the challenges of MARL explained in Section II, mainly the non-stationarity as well as the sparse nature of the reward in our setting (the agents only receive a positive reward when stumbling upon a target during their random exploration, an event that is infrequent at best).

B. Fixed, perfect speaker

To study the use of the communication channel in a much simpler setting, we replace the speaker agent by an oracle, that outputs at each turn a message stating the direction in which the listener should go to get closer to the target. We therefore remove the joint learning difficulty of MARL and consider a single-agent RL problem. The listener only has to learn a mapping between the oracle's messages and its own actions.

We observe that with any view size, the listener uses almost exclusively the communication channel and neglects the visual input even when it is able to see the target. This can be seen on Figure 3.a where the gradient values corresponding to the communication part of the input is much higher than its visual

counterpart, meaning that communication influences more the agent's action choice. Such a behavior is to be expected seeing as directly learning from the oracle is easier than learning features through the convolutional layers of the visual channel, and the latter does not bring any useful additional information. The oracle's messages are indeed enough to reach the target in the most efficient way.

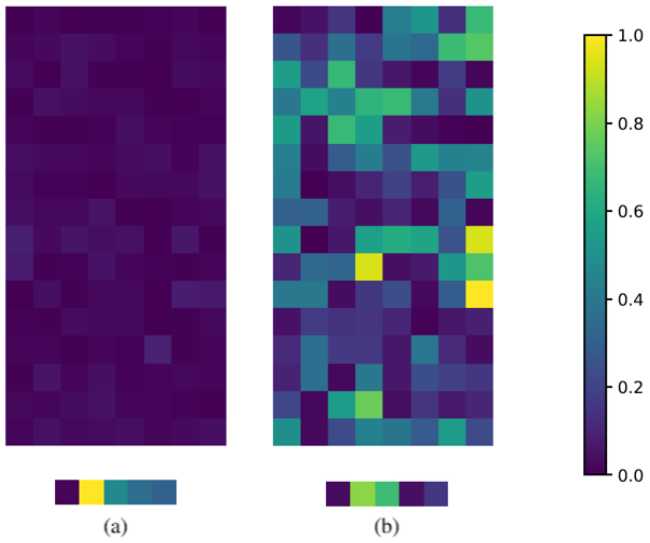


Fig. 3: After feeding the agent's network an observation (visual observation with a view size of 11 and a message), gradient is back-propagated from the agent's action probabilities through the network. The gradient norm before the fully connected layer (see Figure 2) is shown as a heatmap, the large one being the visual part of the input processed by the convolutional layers and the small one being the concatenated message. Gradient norm is much higher for the message. (a) shows the heatmap for a perfect speaker while (b) shows it for a noisy speaker ($\alpha = 0.5$).

C. Noisy speaker

In order to give the listener an incentive to use the visual input, we add a noise to the communication channel. A fraction α of the messages from the oracle are replaced by a random message. We try noise levels of $\alpha = 0.3, 0.5, 0.7$ and 0.9, while also varying the view size of the listener from 0 to 6. We can see in Figure 3.b that the visual input is used a lot more than with the perfect speaker in Figure 3.a.

To measure quantitatively the usage of the communication channel by the listener, we use the *causal influence of communication* measure [13] [10], defined as the mutual information between an agent messages (here the oracle's) and the other agent's actions (those of the listener in this case):

$$CIC = \sum_{a \in \mathcal{A}_1^e} \sum_{m \in \mathcal{A}_0^m} p(a,m) \log \frac{p(a,m)}{p(a)p(m)} \tag{2}$$

where the probabilities of (message, action) co-occurrences are computed using counterfactuals, i.e. manually replacing the messages sent to see how the response from the listener evolves. Intuitively, a high CIC value indicates that messages from the speaker have a high influence on the listener's actions.

We can draw several conclusions out of Figure 4:

- Without noise, CIC is maximal whatever the observability is. This confirms the results of V-B.
- Without observability, the CIC is maximal whatever the noise level is, because the listener can only rely on the speaker messages.
- Increasing the observability or the noise both reduce the CIC, The reason is that observability increases the ability of the listener agent to solve the task by itself, whereas noise reduces the reliability of the speaker messages.

D. Simplified learning speaker

We try once again to tackle a joint learning problem, by adding a simplified adaptive speaker agent, composed of a small multilayer perceptron (MLP) with a single hidden layer. This simplified speaker learns from the previous oracle and not from the pixel input we used in the first experience. It must thus only learn a simple mapping that distinguishes each message from the oracle, seeing as in the previous experiments, the listener could learn efficiently from the oracle's messages. We train both the listener and the speaker using A3C. We can see on Figure 5 that even with this simple task, the joint learning and sparse reward makes it hard for the agents to converge to an efficient policy.

E. Adding positive signaling bias

Seeing as even in a very simplified setting, communication doesn't emerge with an adaptive speaker, we decide to add a bias introduced in [11] to the speaker agent in order to ease the learning. This bias, called positive signalling bias, encourages the speaker to condition its message policy on its observation by adding a loss function to the standard A3C loss computed as follows:

$$L_{ps} = -\mathbb{E}\left(\lambda \mathcal{H}\left(\overline{\pi^{speaker}}\right) - \left(\mathcal{H}\left(m_t^{speaker} \mid o_t^{speaker}\right) - \mathcal{H}_{\text{target}}\right)^2\right) \tag{3}$$

where $\overline{\pi_M^{speaker}}$ is the average policy of the speaker over all trajectories, while the target entropy $\mathcal{H}_{\text{target}}$ and the factor λ are hyperparameters.

The loss function has two components, the first part pushes the speaker to have an average policy with high entropy, while the second part penalizes high entropy of its policy conditioned to its observations. This encourages the speaker to produce messages that uniformly cover the message space overall, while ensuring significant differences between messages produced in different contexts (i.e. different speaker observations).

Adding this loss function helps the speaker to learn a useful communication policy which can in turn be used efficiently by the listener. It allows the simplified speaker and the listened to converge most of the time towards an efficient strategy, although it is still not enough to ensure the emergence of a useful communication system in the settings of Section 5.1.

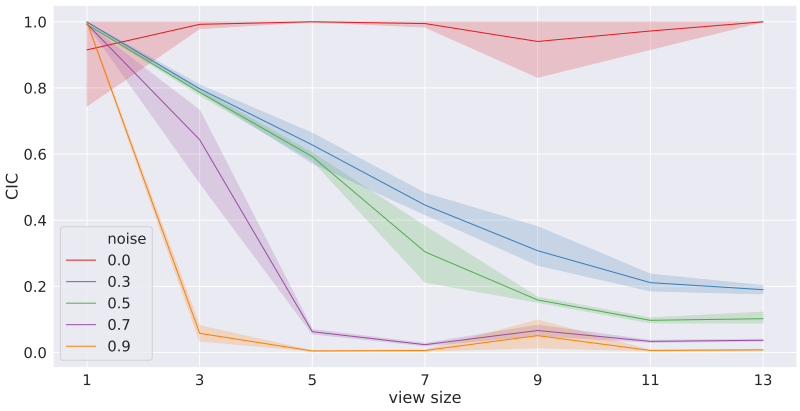


Fig. 4: CIC measure as a function of the listener's view size for an oracle with various noise levels. 95% confidence intervals computed over 10 random seeds are shown.

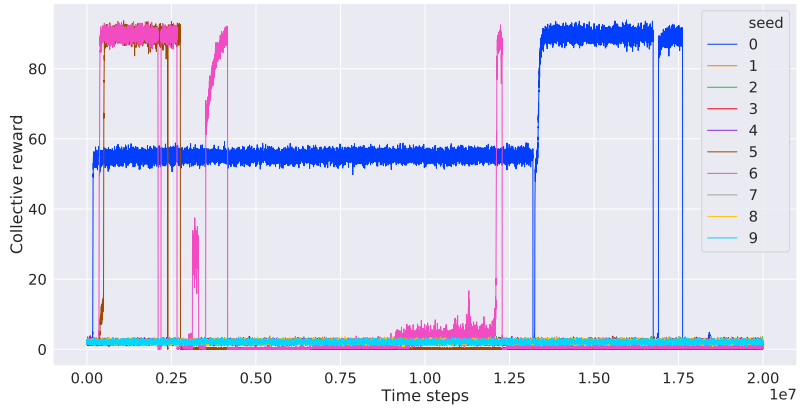


Fig. 5: Mean reward per episode over 20000 time steps with a simplified speaker for 10 random seeds.

Figure 6 shows a similar correlation between the influence of communication on the listener and the size of its visual observation as in Figure 4.

VI. Discussion and future works

We have quantitatively measured the joint influence of partial observability on the emergence of a communication system among artificial agents, through varying the range of the agents' visual pixel input and the channel noise. [7] states that: "Without the combination of multiple agents and partial observability, there is no need to learn a communication protocol". The listener in our experiments does indeed ignore the communication altogether if the range of its visual observation is large enough (Figure 4). The use of backpropagation to evaluate the impact of various part of an agent's input (Figure 3), while being a common technique in computer vision, was never used to our knowledge in the field of emergent communication. The multi-agent aspect of this work however did not yield much results, as the challenges of MARL make it hard for multiple agents to converge to interesting emergent behaviors. Training successfully two or more agents with a similar architecture would be the most straightforward continuation of this work, either by adding other biases from [11] or [10] or forgoing pixel-based input for easier state representation of the environment as in [9].

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *arXiv:1706.03762 [cs]*, Dec. 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805 [cs]*, May 2019.

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," *arXiv:2005.14165 [cs]*, July 2020.

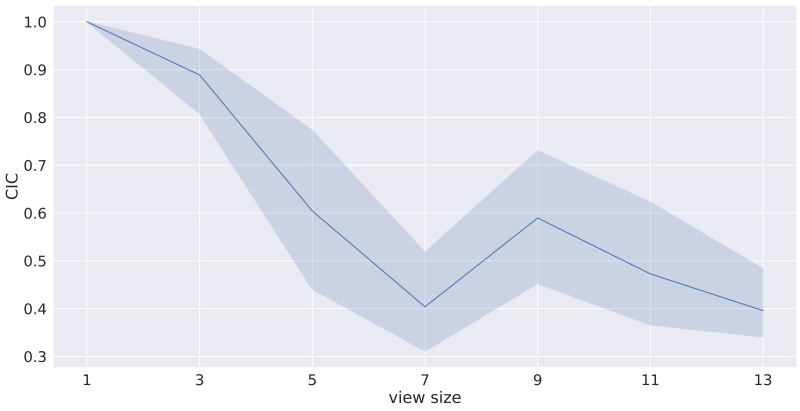


Fig. 6: CIC measure as a function of the listener's view size for an simple adaptive speaker with positive signalling bias. 95% confidence intervals computed over 10 random seeds are shown.

[4] R. Bernardi, G. Boleda, R. Fernández, and D. Paperno, "Distributional Semantics in Use," in *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-Level Semantics*, (Lisbon, Portugal), pp. 95–101, Association for Computational Linguistics, Sept. 2015.

[5] J. Gauthier and I. Mordatch, "A Paradigm for Situated and Goal-Driven Language Learning," in *NIPS 2016 Machine Intelligence Workshop*, Oct. 2016.

[6] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.

[7] J. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Advances in Neural Information Processing Systems*, pp. 2137–2145, 2016.

[8] S. Sukhbaatar, A. Szlam, and R. Fergus, "Learning Multiagent Communication with Backpropagation," *arXiv:1605.07736 [cs]*, Oct. 2016.

[9] I. Mordatch and P. Abbeel, "Emergence of Grounded Compositional Language in Multi-Agent Populations," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2017.

[10] N. Jaques, A. Lazaridou, E. Hughes, C. Gulcehre, P. A. Ortega, D. Strouse, J. Z. Leibo, and N. de Freitas, "Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning," in *Proceedings of the 35 Th International Conference on Machine Learning, Stockholm, Sweden*, Oct. 2019.

[11] T. Eccles, Y. Bachrach, G. Lever, A. Lazaridou, and T. Graepel, "Biases for Emergent Communication in Multi-agent Reinforcement Learning," *arXiv:1912.05676 [cs]*, Dec. 2019.

[12] A. Lazaridou and M. Baroni, "Emergent Multi-Agent Communication in the Deep Learning Era," *arXiv:2006.02419 [cs]*, July 2020.

[13] R. Lowe, J. Foerster, Y.-L. Boureau, J. Pineau, and Y. Dauphin, "On the Pitfalls of Measuring Emergent Communication," in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 693–701, Mar. 2019.

[14] P.-Y. Oudeyer and F. Kaplan, "Discovering communication," *Connection Science*, vol. 18, pp. 189–206, June 2006.

[15] C. Moulin-Frier, S. M. Nguyen, and P.-Y. Oudeyer, "Self-organization of early vocal development in infants and machines: The role of intrinsic motivation," *Frontiers in Psychology*, vol. 4, 2014.

[16] C. Moulin-Frier and P.-Y. Oudeyer, "Multi-Agent Reinforcement Learning as a Computational Tool for Language Evolution Research: Historical Context and Future Challenges," *arXiv:2002.08878 [cs]*, Feb. 2020.

[17] M. Chevalier-Boisvert, D. Bahdanau, S. Lahlou, L. Willems, C. Saharia, T. H. Nguyen, and Y. Bengio, "BabyAI: A Platform to Study the Sample Efficiency of Grounded Language Learning," in *International Conference on Learning Representations*, Sept. 2018.

[18] F. Hill, A. Lampinen, R. Schneider, S. Clark, M. Botvinick, J. L. McClelland, and A. Santoro, "Environmental drivers of systematicity and generalization in a situated agent," *arXiv:1910.00571 [cs]*, Feb. 2020.

[19] C. Colas, A. Akakzia, P.-Y. Oudeyer, M. Chetouani, and O. Sigaud, "Language-Conditioned Goal Generation: A New Approach to Language Grounding in RL," *ICML 2020 Workshop LaReL*, June 2020.

[20] A. Lazaridou, A. Peysakhovich, and M. Baroni, "Multi-Agent Cooperation and the Emergence of (Natural) Language," *arXiv preprint arXiv:1612.07182*, Mar. 2017.

[21] S. Havrylov and I. Titov, "Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols," *arXiv:1705.11192 [cs]*, Nov. 2017.

[22] K. Cao, A. Lazaridou, M. Lanctot, J. Z. Leibo, K. Tuyls, and S. Clark, "Emergent Communication through Negotiation," *arXiv:1804.03980 [cs]*, Apr. 2018.

[23] A. M. Glenberg and M. P. Kaschak, "Grounding language in action," *Psychonomic Bulletin & Review*, vol. 9, pp. 558–565, Sept. 2002.

[24] K. Wagner, J. A. Reggia, J. Uriagereka, and G. S. Wilkinson, "Progress in the Simulation of Emergent Communication and Language," *Adaptive Behavior*, vol. 11, pp. 37–69, Mar. 2003.

[25] P. Hernandez-Leal, B. Kartal, and M. E. Taylor, "A survey and critique of multiagent deep reinforcement learning," *Autonomous Agents and Multi-Agent Systems*, vol. 33, no. 6, pp. 750–797, 2019.

[26] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Independent reinforcement learners in cooperative Markov games: A survey regarding coordination problems.," *Knowledge Engineering Review*, vol. 27, pp. 1–31, Mar. 2012.

[27] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments," *arXiv:1706.02275 [cs]*, 2017.

[28] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Machine Learning Proceedings 1994*, pp. 157–163, Elsevier, 1994.

[29] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning Series, Cambridge, Massachusetts: The MIT Press, second edition ed., 2018.

[30] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous Methods for Deep Reinforcement Learning," *arXiv:1602.01783 [cs]*, June 2016.