Reinforcement Communication Learning in Different Social Network Structures

Anonymous Authors[1]

Abstract

Social network structure is one of the organizing aspects of human language evolution. Previous work has shown that the network of social interactions shapes a multi-agent decentralized coordination process, leading to different kinds of communicative conventions to emerge. We examined the effects of social network organization on the properties of communication systems emerging in decentralized, multi-agent reinforcement learning communities. We found that that global connectivity of a social network drives the convergence of populations on shared and symmetric communication systems, preventing the agents from forming many local "dialects". Moreover, the average degree of an agent is inversely related to the consistency of their use of communicative conventions. These results show the importance of the basic properties of social network structure on reinforcement communication learning and suggest a new interpretation of findings on human convergence on word conventions.

1. Introduction

Human languages evolve as complex adaptive systems, driven by micro-level processes and constraints (such as individual learning mechanisms and perceptual biases), macro-level factors (such as a topology of social interactions), and the history of their development (Steels, 2000; "Five Graces Group" et al., 2009).

Linguistic communication depends on the shared knowledge of word-to-meaning mapping conventions (Lewis, 1975), upon which the population converges through the local interactions between agents, often with no central controller available (Baronchelli, 2018). Empirical studies of human learning demonstrate that groups quickly converge on new communicative conventions in "decentralized" settings, even when prevented from using natural languages (Garrod & Doherty, 1994; Selten & Warglien, 2007).

Multi-agent reinforcement learning to communicate (MARLC), however, faces instability challenges if no central optimization is introduced (Bernstein et al., 2002; Laurent et al., 2011; Matignon et al., 2012). This influences the ability of groups consisting of reinforcement learners to converge on efficient and stable communication systems which are shared by all their members. Therefore, different methods of centralized control and optimization have been proposed to stabilize multi-agent reinforcement learning to communicate (Sukhbaatar et al., 2016; Foerster et al., 2016; Lowe et al., 2017; Pesce & Montana, 2020; Foerster et al., 2018). Central optimization makes the simulations more brittle and less flexibly adapted, and, potentially, less promising in developing communication systems as freely expressive and well-optimized for their users (Gibson et al., 2019) as natural languages.

We argue that empirical evidence on individual- and population-level factors that drive decentralized learning in human groups can guide simulations of language evolution in MARLC settings. In this work, we explore whether the social network organization shapes the properties of communication systems that arise through decentralized MARLC in simplified settings.

1.1. Human Learning in Different Social Network Structures

Convergence of human groups on word conventions is dramatically affected by the social topology that determines the

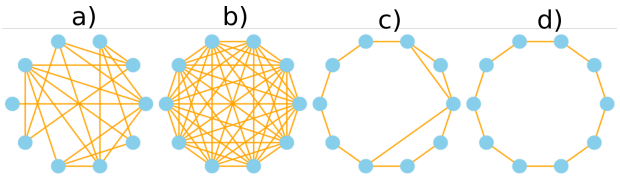

Figure 1. Types of social topologies tested in our experiments. a) Random (Erdos, 1959), b) Fully-connected (clique), c) Small-world (Newman & Watts, 1999), d) Ring

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

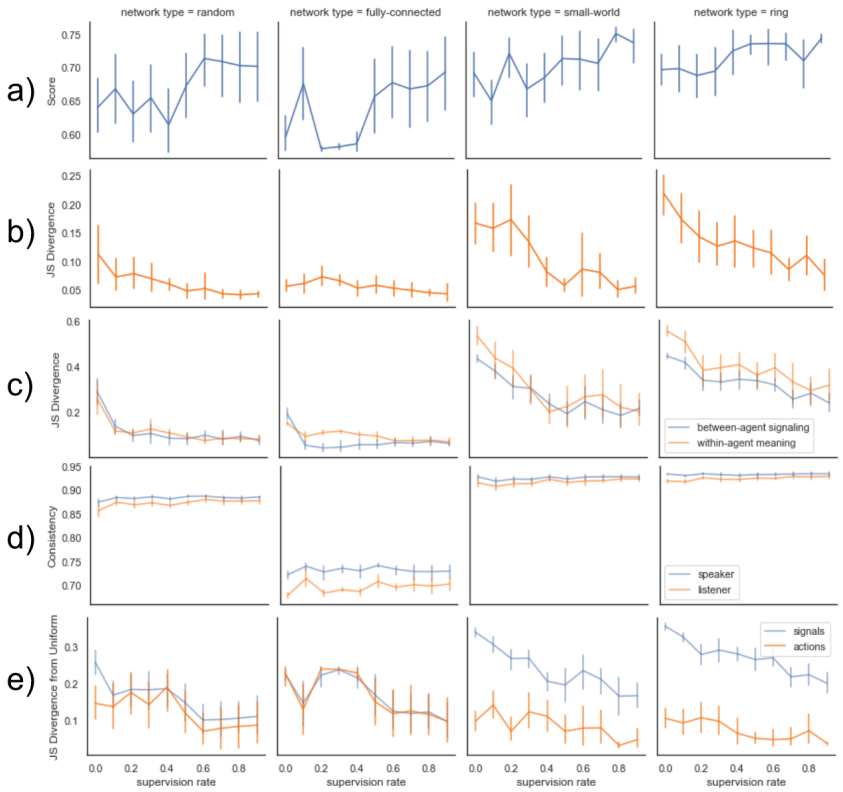

Figure 2. The Communication Analysis Metrics with 95% Confidence Intervals for Experiment 1. a) Average rewards b) Between-agent signal-action mapping divergence c) Signaling divergence (blue) and within-agent signal-action mapping divergence (orange) d) Average speaking (blue) and listening (orange) consistencies e) Average predictability of agents' signals (blue) and actions (orange).

possible interactions between participants, as demonstrated by the naming game experiment on large groups of participants (Centola & Baronchelli, 2015). In particular, when arranged in a social network with many local connections (e.g. "ring" topology (Fig. 1d)) or a randomly-connected network (Fig. 1a), groups converge on many local word conventions, reaching no global consensus. However, if each person is equally likely to interact with any other person in the group ("clique" (Fig. 1b)), global consensus is easily achieved with no centralization. Other studies on decentralized problem solving in human groups demonstrated that social network organization shapes the multi-agent optimization process, with different network types being beneficial for different types of optimization landscapes. High local connectivity supports independent local exploration, whereas high global connectivity helps groups to converge on a shared solution, choosing among the local ones (Fang et al., 2010; Mason et al., 2008; Mason & Watts, 2012; Lazer & Friedman, 2007; Wisdom & Goldstone, 2011).

In this study, we looked at how the type of social network organization, its average degree, and local connectivity affect the results of communication learning in groups of deep reinforcement learning agents.

2. Method

2.1. Coordination Game

Every game episode involves two agents, randomly assigned to speaker and listener roles. The speaker produces a message and an action, while the listener receives the message and outputs an action[1]. If the actions match, the agents receive a reward. This setting encapsulates the most basic form of a language game that benefits from convention formation. We implement an additional penalty for overusing any specific action to avoid degenerate solutions that ignore the communication channel altogether. Please see supplementary materials for more detail.

2.2. Types of Social Network Organization

Social network type determines how the agents are sampled to play games with one another. For each game round, one agent is selected randomly, and then its partner is selected from its neighbors. Therefore, an agent can only play a game with one of its immediate neighbors in the social network. In all the simulations, the social network size was set to

[1]Action and signaling sizes were set to 4 in all the simulations.

10. The networks were undirected, and the self-connections were not allowed. We tested 4 types of social networks in our experiments (Fig. 1):

1. Random (ER). Random graph is generated by connecting the nodes with equal probability p (Erdos, 1959).

2. Fully-connected (clique). In clique, all the nodes are connected to each other.

3. Ring. In the ring network, all the nodes have exactly two neighbors, and connections form a single continuous path through all the nodes.

4. Small-world. The small-world network is generated by adding new, "global", connections to the ring network with constant probability p (Newman & Watts, 1999).

2.3. Agents

We used simple feed-forward neural networks to represent the agents. The networks were trained using a vanilla deep Q-learning algorithm (Mnih et al., 2015) with an added proportion of bottom-up driven "supervising" feedback (see supplementary materials). We tried to avoid any centralized optimization, popular in MARLC settings, to study the ability of different networks to self-organize conventions.

2.4. Metrics

We used a number of information-theoretic metrics, developed in Lowe et al. (2019) and Author & Author (2020), for a comprehensive evaluation of communication protocols: speaker & listener consistency, between- and within-agent signal-action mapping divergence, signaling divergence, and behavioral predictability (see supplementary materials).

3. Results

3.1. Experiment 1: types of social network organization

We simulated multi-agent learning in 4 types of social networks: **ring** (avg degree=2, var=0), **random** (avg degree=2, var=1.6, p(connection)=0.2), **small-world** (avg degree=2.2, p(add global connection)=0.2), and **clique** (avg degree=9, var=0). We also varied the supervision rate (from 0.0 to 0.9 with a step of 0.1). MARLC in each combination of these two conditions was simulated 10 times to get statistical estimates of the communication metrics. In all the experiments, each simulation consisted of 120000 game rounds.

The average speaker and listener consistencies were highest in the ring and small-world networks, and the lowest in the fully-connected structure. Consistency scores did not vary with the supervision rate (Fig. 2d). These patterns indicated the potential dependency of speaker and listener consistency on a single factor: agents' degrees. The fully-connected network had the highest possible average degree. This hypothesis is tested in Experiment 2.

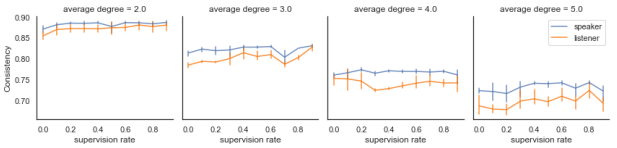

Figure 3. Speaker and listener consistency estimates with 95% Confidence Intervals and average degree of agents in random social network (Experiment 2)

The agents in random and fully-connected social networks developed almost perfectly symmetric and homogeneous communication systems according to all three communication asymmetry metrics (Fig. 2b and 2c). Small-world and ring social networks lead the populations to develop asymmetrical and local communication patterns. Moreover, supervised feedback helped the agents in all topology conditions to develop more shared and symmetric communication systems. We hypothesized that the effect of network type on asymmetry scores is driven by local connectivity of the populations: fully-connected and random networks do not form local communities, whereas the ring-shaped and small-world networks mainly consist of local connections in our simulations. We tested this hypothesis in Experiment 3.

Lastly, agents in fully-connected and random networks produced less diversified actions, but their action and signaling distributions were much more coordinated than in the ring-shaped and small-world network structures (see Fig. 2). It suggests that agents in the last two conditions overfitted to the "action diversity" part of the reward function, while failing to coordinate using the communication channel.

3.2. Experiment 2: average degree

To test the effect of network's average degree on speaker and listener consistencies, we focused on the random (ER) network, which produces variation in the average degrees of the nodes, while keeping most of the other network properties constant. We varied the probability of connecting nodes in the network from 0.2 to 0.9, and supervision rate from 0 to 0.9 with a step of 0.1. This resulted in 80 conditions total, each of which was simulated 5 times to obtain consistency estimates with confidence intervals.

We found that the average degree negatively affected listener ($p < 0.001$) and speaker ($p < 0.001$) consistencies (Fig. 3). Consistency scores for the average degrees higher than 5.0 stayed similar to the estimate we obtained for the fully-connected topology in Experiment 1.

3.3. Experiment 3: global and local connections

We examined the effect of locally-connected groups in a social network on homogeneity and symmetry of the developed communication systems. For this, we tested MARLC

in the small-world network with different probabilities of adding a "global" connection to the initial ring-shaped structure. We classify a link as "global" if it connects agents further than one link apart in the ring network. We varied both the probability of global connection and the supervision rate from 0 to 0.9 with a step of 0.1. Each combination of these two factors was simulated 5 times.

Global connection probability was inversely related to the signaling divergence ($p < 0.001$), between- ($p < 0.001$) and within-agent signal-action mapping divergence ($p < 0.001$) of the developed communication systems (Fig. 3). Higher proportion of global connections led the groups to converge on communication systems that are shared by all their agents. Supervised feedback also helped the agents to develop homogeneous and symmetric communication patterns. The cumulative effect of both high supervision rate and probability of global connections lead to communication systems that are shared by all the agents in a group.

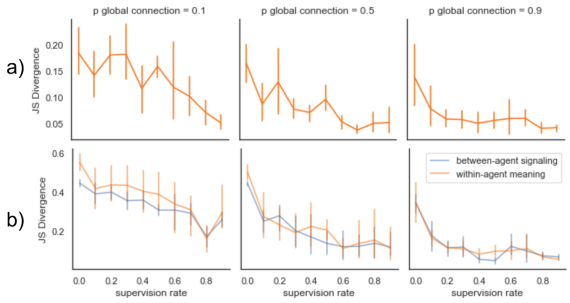

Figure 4. The Communication Analysis Metrics with 95% Confidence Intervals for the Experiment 3. a) Between-agent signal-action mapping divergence b) Signaling divergence (blue) and within-agent signal-action mapping divergence (orange) and the probability of global connection in the small-world network

4. Discussion and Conclusion

We conducted three experiments to explore and isolate the factors of social network organization that drive the effectiveness, homogeneity, and symmetry of communication systems developed in MARLC settings. Following the suggestions of Lowe et al. (2019), we used a set of information-theoretic metrics to evaluate our results. This allowed us to determine which particular properties of communication systems are affected by our interventions.

The results of our first experiment partially replicated the effects of social network type on learning word conventions by human groups (Centola & Baronchelli, 2015). This suggests that the simple reinforcement learning model can capture core regularities found in human convergence on linguistic conventions. Our further experiments provide a more detailed account of particular network properties that are responsible for these patterns in communication learning.

The second experiment demonstrated that the average degree of a social network is a central factor affecting how deterministically agents use the communication channel. The more communication partners an agent needs to adapt to, the more variable are its signaling and listening patterns. Key finding of the third experiment is that the proportion of global connections is the primary determining factor for three variables: homogeneity, within- and between-agent symmetry of the developed communication system. In particular, local social connections led to the emergence of local communication patterns ("dialects") within the population, whereas adding more global connections forced the agents to find global consensus.

At present, social network effects on MARLC have been largely under-investigated. To the best of our knowledge, there is only one work in this direction; L. Graesser and her colleagues (2019) looked at how communication systems developed within communities change if one community is introduced to another, as a function of their inter- and intra-connectivity. Unfortunately, the insights from that work are only applicable for the specific scenario of community merging. In our study, we focused on the population-level factors that are applicable to almost any MARLC setting.

The results of our analysis corroborate Lowe et al.'s (2019) claim that overall scores do not necessarily reflect whether the agents use the communication channel to solve the task, and how efficiently they do so (see Fig. 2). For example, the performance scores of the agents learning to coordinate in a ring-shaped social network were the highest among all the network types. However, more detailed analysis determined that agents in this condition developed many local communication patterns, which varied across agents and their roles. This, again, illustrates the importance of thorough analysis of communication protocols learned in MARLC settings.

Our approach has a number of limitations that are important to mention. Firstly, while we aimed to test a broad spectrum of social network factors, our experiments are by no means comprehensive. There are other important aspects of social network structure that may have a key role in determining the properties of emerging communication protocols. We believe that studying the effects of variance of network's degree distribution and its modularity on MARLC is especially promising. Secondly, we used a very simplified setting of vanilla Q-learning in an "amodal" coordination game to minimize the number of assumptions that might make our results less representative to the MARLC problem in general. We suggest testing these results on more advanced reinforcement learning models and realistically perceptually grounded game settings.

References

Author, X. and Author, Y. Suppressed for anonymity (in press). In *Supplementary Materials*, 2020.

Baronchelli, A. The emergence of consensus: a primer. *Royal Society open science*, 5(2):172189, 2018.

Bernstein, D. S., Givan, R., Immerman, N., and Zilberstein, S. The complexity of decentralized control of markov decision processes. *Mathematics of operations research*, 27(4):819–840, 2002.

Centola, D. and Baronchelli, A. The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences*, 112(7):1989–1994, 2015.

Erdos, P. On random graphs. *Publicationes mathematicae*, 6:290–297, 1959.

Fang, C., Lee, J., and Schilling, M. A. Balancing exploration and exploitation through structural design: The isolation of subgroups and organizational learning. *Organization Science*, 21(3):625–642, 2010.

Foerster, J., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pp. 122–130. International Foundation for Autonomous Agents and Multiagent Systems, Jul 2018.

Foerster, J. N., Assael, Y. M., de Freitas, N., and Whiteson, S. Learning to communicate to solve riddles with deep distributed recurrent q-networks. *arXiv:1602.02672 [cs]*, Feb 2016. URL http://arxiv.org/abs/1602.02672. arXiv: 1602.02672.

Garrod, S. and Doherty, G. Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*, 53(3):181–215, 1994.

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., and Levy, R. How efficiency shapes human language. *Trends in cognitive sciences*, 2019.

Laurent, G. J., Matignon, L., Fort-Piat, L., et al. The world of independent learners is not markovian. *International Journal of Knowledge-based and Intelligent Engineering Systems*, 15(1):55–64, 2011.

Lazer, D. and Friedman, A. The network structure of exploration and exploitation. *Administrative science quarterly*, 52(4):667–694, 2007.

Lewis, D. Languages and language. 1975.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, O. P., and Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, pp. 6379–6390, 2017.

Lowe, R., Foerster, J., Boureau, Y.-L., Pineau, J., and Dauphin, Y. On the pitfalls of measuring emergent communication. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 693–701. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

Mason, W. and Watts, D. J. Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, 109(3):764–769, 2012.

Mason, W. A., Jones, A., and Goldstone, R. L. Propagation of innovations in networked groups. *Journal of Experimental Psychology: General*, 137(3):422, 2008.

Matignon, L., Laurent, G. J., and Le Fort-Piat, N. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *The Knowledge Engineering Review*, 27(1):1–31, 2012.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015.

Newman, M. E. and Watts, D. J. Renormalization group analysis of the small-world network model. *Physics Letters A*, 263(4-6):341–346, 1999.

Pesce, E. and Montana, G. Improving coordination in small-scale multi-agent deep reinforcement learning through memory-driven communication. *Machine Learning*, Jan 2020. ISSN 1573-0565. doi: 10.1007/s10994-019-05864-5. URL https://doi.org/10.1007/s10994-019-05864-5.

Selten, R. and Warglien, M. The emergence of simple languages in an experimental coordination game. *Proceedings of the National Academy of Sciences*, 104(18): 7361–7366, 2007.

Steels, L. Language as a complex adaptive system. In *International Conference on Parallel Problem Solving from Nature*, pp. 17–26. Springer, 2000.

Sukhbaatar, S., Fergus, R., et al. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems*, pp. 2244–2252, 2016.

Wisdom, T. N. and Goldstone, R. L. Innovation, imitation, and problem solving in a networked group. *Nonlinear Dynamics-Psychology and Life Sciences*, 15(2):229, 2011.

"Five Graces Group", Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., et al. Language is a complex adaptive system: Position paper. *Language learning*, 59: 1–26, 2009.

A. Coordination game environment

Each game consists of two-step episodes. Every episode of the game involves two agents: a speaker and a listener who need to coordinate their actions.

On the first step, the speaker produces an action and a message. On the second step, the listener receives the message and outputs an action. If the actions match, agents receive a fixed "coordination" reward R_c. There are four unique actions and messages that the agents choose from.

In order to avoid trivial solutions when agents converge on a single action and ignore the communication channel, we introduced a penalty for repeating an action too often. Specifically, an additional reward of $min(0, 1/4 - \hat{p}_A)$ is received when an agent chooses action A. Here $\hat{p}_A$ denotes an empirical proportion of selecting action A during the last H steps. We used a fixed length history $H = 100$. Current $\hat{p}_A$ are added to the agent's input state spaces to aid convergence.

On every episode, each agent can be assigned to play either the speaker or the listener role. The current role type is provided as a binary input.

Overall, the agents receive, as inputs:

1. A binary role indicator
2. Random input (for speakers, in order to allow for non-deterministic policies) or speaker's message (for listeners)
3. Proportions of different actions in agent's recent history

The agents have two output layers both stemming from the last hidden layer:

1. Action output layer (for neurons, and the action is determined as argmax)
2. Message output layer which has the same structure as the action output layer. Message outputs are ignored if the agents plays the listener role

B. Agents

Agents are implemented using simple feed-forward neural networks with two hidden layers (hidden sizes of 25 and 15) and ReLU activation function.

We used vanilla deep Q-learning algorithm with additional supervised updates to train the agents. Supervised feedback corresponds to "peeking" the other agent's action and storing it in memory (with the signaling experience from that game) as a "correct" response. In this case, the action-signal mapping is not superimposed, and the "correct" answers in the supervising trials are bottom-up driven and reflect the dynamics of agents themselves. Supervised feedback was implemented by changing a certain proportion of negative ("miscoordinated") experiences to the "supervising" ones. We decided to control for supervision in our experiments, since partial supervising feedback in different forms is often available in realistic language learning situations. By amplifying the reinforcement signal, supervising feedback may help to overcome the "sparcity of rewards" problem in multi-agent reinforcement learning.

C. Metrics

We follow the same set of metrics as used in (Author & Author, 2020).

1. **Speaking Consistency and Listening Consistency**.

   This metric provides a quantitative measure of whether the actions that the agents perform are related to the signals that they send or receive. We follow Lowe et al. (2019) who proposed to use the normalized mutual information between the distributions over messages and actions induced by an agent. The metric can be formally described as follows:

$$C = \sum_{a \in A_l} \sum_{m \in A_c} p_{a,m}(a,m) log \frac{p_{a,m}(a,m)}{p_a(a) p_m(m)} / Z \quad (1)$$

   Here, Z is the average entropy of the two marginal distributions: $Z = \frac{(H(p_a)+H(p_m))}{2}$. A_l and A_c denote the set of available actions and the set of available messages respectively.

   This metric is computed twice for every agent, conditioned on the role played by the agent (speaker or listener). We average these metrics across agents in every simulation to obtain the speaking and listening consistency metrics that we report.

2. **Communication Asymmetry metrics**:
   a. Between-agent signal-action mapping divergence. While the previous metric aimed to measure

whether an agent is using communication channel (i.e. whether its actions correspond to the signals in any way), the second metric aims to measure whether the communication patterns differ between agents.

We compute the average Jensen-Shannon pairwise divergence between distributions of agents' actions following a specific signal (averaged over all signals and pairs of agents).

For a pair of agents, the metric is defined as follows

$$\sum_{m \in A_c} JSD(p_{a_1|m}, p_{a_2|m})/|A_c| \tag{2}$$

Here, $p_{(a_1|m)}$ are action distributions of agent 1 conditional on the message (received or sent) being equal to m.

b. Within-agent signal-action mapping divergence.

This metric aims to capture internal inconsistencies in agent's behaviours when they switch between roles: whether the agent's behaviours differ depending on whether they receive or send a specific message.

For that, we use the average Jensen-Shannon divergence between the distributions over agent's actions (conditioned on receiving or sending a specific message) when the agent plays the speaker and the listener role. Formally, we use the same definition as in Equation 2, but now $p_{a_1|m}$ and $p_{a_2|m}$ correspond to the same agent's distributions when this agent plays different roles (as opposed to distributions of a pair of different agents playing the same role). We average the metric scores across all agents to obtain the final measure that we report.

c. Signaling divergence.

This metric aims to measure the difference in individual agents' messaging preferences. We define *signaling divergence* as an average pairwise Jensen-Shannon divergence of marginal signaling distributions of different agents.

3. **Behavioral Predictability**. This last metric is created to assess the general diversity in agent's actions. When the agents' actions are less diverse (and hence, more predictable), it is easier to achieve successful multi-agent coordination without using the communication channel. To look at whether the diversity of actions corresponds to the diversity of signals, we also compute the predictability of agents' signaling patterns. We define *behavioural action/message predictability* as Jensen-Shannon divergence between marginal distributions of agent's actions/messages and the uniform distribution.

Many of these metrics require knowing probability distributions. We estimate all such distributions empirically.

As a short summary, if we see the learned language as a simple probabilistic dictionary that maps messages to actions, the metrics can be summarized as follows (note that every agent defines two such dictionaries: one for the speaker and one for the listener role):

1. **Speaking Consistency and Listening Consistency**. Are the dictionaries reliable? I.e. if we look up a specific message, do we consistently get the same action, or is there a lot of randomness?

2. **Communication Asymmetry**:
   a. Between-agent signal-action mapping divergence. Are the dictionaries similar for different agents?

   b. Within-agent signal-action mapping divergence. How different, on average, are the "speaker" and "listener" dictionaries that agents have?

   c. Talking divergence. Do different agents show different patterns in their dictionary lookups?

3. **Behavioral Predictability**. How uniformly do agents look up different words in the dictionary (speaking predictability)? How uniform are the results of their queries (behavioral predictability)?

D. Statistical analysis

All hypotheses were tested using a linear regression model with robust covariance estimation, controlling for supervision rate. Excluding the supervision rate did not qualitatively change the results, however.

Anonymized Reference (in press): Effects of Supervision, Population Size, and Self-Play on Multi-Agent Reinforcement Learning to Communicate

Author X. and Author Y.

Abstract

Learning to communicate in adaptive multi-agent populations introduces instability challenges at the individual and population levels. To develop an effective communication system, a population must converge on a shared and sufficiently stable vocabulary. We explore the factors that affect the symmetry and effectiveness of the communication protocols developed by deep reinforcement learning agents playing a coordination game. We looked at the effects of supervising feedback, bigger agent population size, and self-play ("inner speech") on the properties of the developed communication systems. We found that all these factors greatly benefit the learning outcomes of the adaptive agents. The populations with more than 2 agents or with a self-play learning mode converge on more shared and symmetric communication protocols than the 2-agent (no self-play) groups. Supervision, in turn, augments the learning results of all groups, helping the agents learning in bigger populations or with self-play to coordinate and converge on maximally shared and symmetric communication systems. We discuss the implications of our results for future work on modeling language evolution with multi-agent reinforcement learning.

Introduction

Human shared communication systems (languages) evolved through complex interaction of cultural and individual adaptation processes. There are many factors that determine the trajectory of language evolution, such as developmental, morphological, and cognitive constraints of the communicating agents; the dynamics of their learning processes; and population-level factors, such as geographical constraints, social structure, and many others (Christiansen and Kirby, 2003; Oudeyer, 2006; The Five Graces Group et al., 2009; Hashimoto, 2002).

Even though the mapping of words to their meanings is arbitrary, individuals in a society must converge on some fixed set of shared mappings to have an effective communication system. It is not surprising that even in animals, a wide array of strategies are used to ensure consistency of such mappings within agent communities. For example, *goldfinch* birds alter their calls to match those of the most dominant bird in the group (Snowdon, 1990). In human communities, the role of communication is, arguably, more diverse, and a broad range of tools is employed, helping individuals learn a consistent and shared mapping from signals to meaning. The existence of various cultural transmission mechanisms, asymmetric adaptation patterns in parent-child dyads, and more generally, the education system as a whole, can all be seen as examples of tools helping us to ensure the consistency of our communication system.

In our work, we study communication learning in a simplified setting, and investigate the factors that drive symmetry and consistency in the communication patterns of artificial agent populations.

Related work

Existing language evolution models include *genetic algorithms* (Nowak et al., 2002; Nowak and Krakauer, 1999; Nowak et al., 1999; Reali and Griffiths, 2010), *iterative learning* (the model of cultural evolution) (Kirby et al., 2014; Griffiths and Kalish, 2007), *multi-agent models* (Gong, 2009, 2011; Hashimoto and Ikegami, 1996), *robotic simulations* (Steels, 2015; Spranger and Steels, 2012; Oudeyer, 2005), as well as *hybrid models* (i.e. combining genetic algorithms and neural networks) (Batali, 1994; Cangelosi and Parisi, 1998), and others (Kirby, 2002). While language change is induced by both individual-level and population-level mechanisms, these models often focus on formalizing the adaptation and transmission processes on only one of these levels. For instance, genetic, iterative learning, and most multi-agent models formalize the transmission on the population-level, sacrificing the details of individual learning processes. Robotic simulations, in turn, focus more closely on individual and dyadic learning mechanisms and dynamics, and are often difficult to scale to bigger populations.

In this context, neural network models provide a good compromise. Their expressiveness allows us to model rich individual learning dynamics. At the same time, with enough computational resources, such models can be scaled to model moderate-size population dynamics, as opposed to, for example, robotic simulations, which necessarily run

in real time. It is not surprising, therefore, that rise of deep learning research prompted a new wave of language evolution studies. A number of works have demonstrated the promise of this approach, showing that the deep learning agents can learn communication protocols in various game scenarios (such as Lazaridou et al. (2016); Evtimova et al. (2018); Lowe et al. (2017); Foerster et al. (2016); Sukhbaatar et al. (2016); Mordatch and Abbeel (2018), and many others).

Instability of Multi-Agent Reinforcement Learning

The work on training deep neural network agents to communicate demonstrates the main challenge of this approach: the instability of independent multi-agent reinforcement learning. In this case, the individual agents adapt simultaneously and determine the reward and training data for the other agents, making the optimization problem for each of them non-stationary and non-Markov (Bernstein et al., 2002; Laurent et al., 2011; Matignon et al., 2012). Such a population is not guaranteed to converge on shared and stable strategies, such as a consistent mutual communication system.

To overcome this issue, the aforementioned works have restricted the learning situations to the fully-supervised or semi-supervised interactions of 2 agents with fixed or flexible communication roles (speaker and listener) (Lazaridou et al., 2016; Havrylov and Titov, 2017; Chaabouni et al., 2019; Lazaridou et al., 2018; Evtimova et al., 2018; Graesser et al., 2020); centralized the optimization problem by fitting a shared controller for all the agents (Sukhbaatar et al., 2016); adopted inter-agent weight sharing (Foerster et al., 2016); used centralized critic in the actor-critic algorithm (Lowe et al., 2017), and allowed shared collective memory or access of agents to each other agent's exact weights (Pesce and Montana, 2020; Foerster et al., 2018). Some of the latter works report a comparison of the chosen centralization method with the independent learners' scenario, demonstrating that the alternative model performs poorly and fails to converge. Moreover, the multi-agent learning instability dramatically increases with the population size, often restricting the "decentralized" (independent) learning simulations to 2-4 agents.

The challenges with decentralized optimization of multiple deep learning agents raise the question of how humans, flexible and independent in their learning, have successfully converged on shared communication systems.

Stabilizing naive multi-agent communication learning with realistic interventions

We hypothesize that human learners have access to various sources for supervising feedback on their communication, which amplifies and disambiguates the reinforcement signals. For example, humans can infer potentially rewarding actions or symbol meanings even after an unsuccessful interaction, by observing the action of their partner. In our simulations, we test how the generic supervising feedback helps to overcome the "sparsity of rewards" problem in multi-agent reinforcement learning case. We adapted the supervising feedback for the Q-learning model, and tested how different rates of supervision affect the population learning results.

Secondly, a small number of adaptive agents playing a communication game encounter no pressure that would prevent them from partitioning the signaling space and developing an *asymmetric* communication system (Evtimova et al., 2018). It is particularly true for the 2-agent setup when the agents do not experience any randomness in who they are going to play with. To address this issue, we are looking at two potential factors enforcing the development of symmetric communication protocols: population size and self-play ("inner speech"). We hypothesize that human languages have originally emerged and developed in groups of more than 2 individuals. If such a case is applied to the modeling setup, each agent tries to communicate with different partners, and, potentially, with new ones. This makes specialization of agents' vocabularies non-effective, especially when the partner's identity could not be reliably inferred. The "inner speech" manipulation, in turn, also constrains the learning problem towards developing symmetric meanings. In this case, the agents communicate with the other agents as well as with themselves and try to coordinate their responses in these two cases. Therefore, the self-play setting forces them to coordinate their own responses to communicative signals with the other agent's responses to them.

A couple of other works have tested some realistic factors stabilizing independent communication learning. In particular, Eccles et al. (2019) explored the effects of the biases for positive signalling and positive listening and Bogin et al. (2019) looked at the effect of the bias for speaker consistency on the overall scores of the communication game learning.

We view our work as a continuation of this exciting line of research, and extensively test whether the population size, self-play, and supervising feedback improve the results of multi-agent communication learning. To analyze the results of our simulations, we develop a set of metrics to isolate the properties of communication that are affected by these manipulations.

Problem Formulation

We consider a number of multi-agent communication settings in which two or more agents have to coordinate their actions in order to receive reward. In our problem setting, we aimed to study a minimal example representing the challenges of multi-agent communication learning.

In this section, we formalize the notion of a communication game, and formally describe the settings that we considered.

One-way communication

Before we describe the multi-way multi-agent settings that constitute the main body of our work, let's first consider the simpler situation of one-way two-agent communication.

In this simplified scenario, there are only two agents: the speaker and the listener. The speaker chooses an action and a message to communicate, the listener receives the message and chooses an action. If the actions match, both agents receive a reward.

The setting for every agent is close to that of one-step Markov Decision Process (MDP), which is a four-tuple: (S, A, P, R). We will use a standard notation, denoting the set of states as S, the set of actions as A, the transition probabilities as P, and the reward function as R.

These processes are different for the speakers and the listeners. The action space is $A_l = \{1...K\} \subset N$ for the listeners. In other words, the listeners pick a unique action among K available options. For the speakers $A_s = A_l \times A_c$, where $A_c = \{1...T\} \subset N$. That is, in addition to picking an action from the same state as the listeners, speakers also select a message to communicate among T available options. If the chosen actions coincide, the agents receive a fixed reward for successful cooperation.

Agents also receive an additional penalty if the action they chose was overused in the past. The penalty is asymmetric: agents receive no bonus for using rare actions. The penalty is determined by $\beta(1/K - \hat{p}_k)$, where K is the number of possible actions and $\hat{p}_k$ is the proportion of times this action was selected in during the last H episodes, where H is a hyperparameter [1]. This reward is utilize to make sure that agents learn to cooperate using a communicative channel, instead of relying on behavioral predictability of the other agent.

The state spaces are different for speakers and listeners as well. For the speaker, the state space includes a random input (in order to allow for non-deterministic policies), as well as the average frequencies of different actions. The listener receives only the speaker's message and the frequency history of its actions.

Despite the interface similarity between MDPs and the setting describe above, the problems that the agents are facing do not fall under the MDP framework. The problem is that the reward depends on the policies (and, hence, internal states) of other agents, which are not observed. This makes the environment only partially observable. Moreover, the actions that the agents pick affect what the other agents are going to learn, which, together with the fact that the states of other agents are not observed, makes the environment non-Markov.

[1] H was set to 100 in all simulations

Two-way communication

One-way communication setting described above introduces unrealistic restrictions: one agent is always speaking, while another is always listening. This presents a sharp contrast with most naturalistic interactions.

We, therefore, use an extension of the problem posed above, with only one key difference: on each game episode, the speaker and listener roles are assigned randomly.

To account for that difference, the state space is modified accordingly. More specifically, $S = S_r \times S_i \times S_h$. Where $S_r = \{0, 1\}$ is a binary indicator of the current role. $S_i = R^K$ is a input vector, that either contains random noise (for speakers) or the speaker utterance (for listeners). Lastly, $S_m = [0, 1]^K \subset R^K$, represents the memory of the agent, specifically, the action choice proportions over the last H episodes.

The action state is now the same for both agents and coincides with A_s in the one-way communication setting. The communication part of the action, however, is not used if the agent is assigned the listener role.

Multi-agent two-way communication

In most experiments we consider a direct generalization of the two-way communication setting. Instead of two agents, we have a pool of agents P. In every episode, two agents are randomly selected, and then the episode proceeds as described in the two-way-communication section.

Model

Our choice of the model and its training procedure is also motivated by the considerations of simplicity and representativeness to the main multi-agent communication learning challenges.

By focusing on a simple problem, we can reasonably hope that any difficulties in agent's learning are caused by the dynamics of multi-agent interaction, and not by limitations of the underlying RL-algorithm or network architecture. In general, the policies that must be learned to solve our task are very simple, the only challenge is that all agents need to learn *compatible* policies (e.g. Q-functions that would allow them to behave in a coordinated manner).

Therefore, we opted for using a vanilla Deep Q-learning algorithm as the main workhorse for training the agent networks. All agents are represented by feedforward neural networks with two hidden layers containing 25 and 15 neurons respectively. We used the ReLU (Rectified Linear Unit) activation function.

Metrics

Following Lowe et al. (2019), we used a number of specific metrics in addition to the reward scores to uncover the properties of learned communication systems.

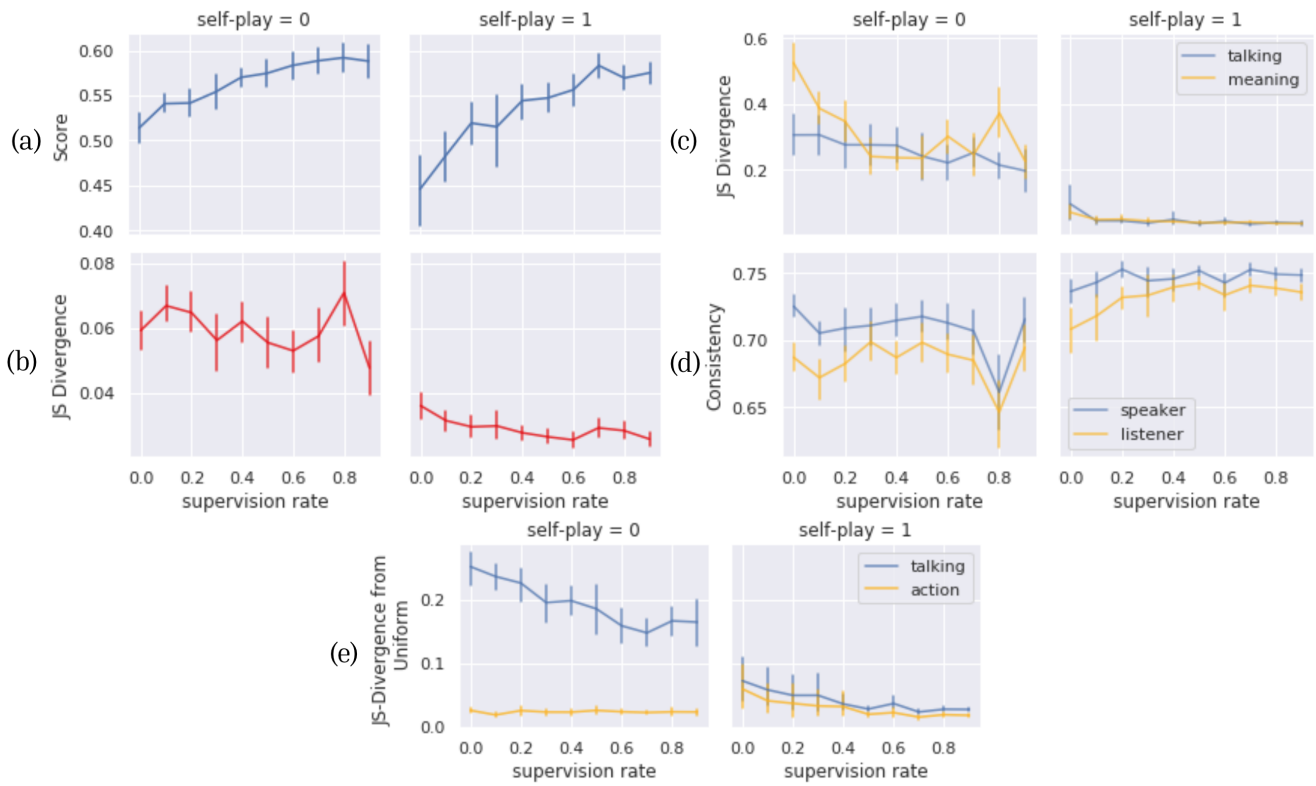

Figure 1: The Communication Analysis Metrics with 95% Confidence Intervals for the Experiment 1. a) Average rewards b) Between-agent signal-action mapping divergence c) Talking divergence (blue) and Within-agent signal-action mapping divergence (yellow) d) Average speaking (blue) and listening (yellow) consistencies e) Average predictability of agents' signals (blue) and actions (yellow).

Most of the metrics need a number of probability distributions to be computed. We estimate all such probability distributions empirically.

1. **Speaking Consistency and Listening Consistency**.

   First of all, it is important to establish whether the actions that the agents perform are in any way related to the signals that they send or receive. Lowe et al. (2019) proposed to use the normalized mutual information between the distribution over messages and actions that an agent defines. Formally, the metric is defined as

$$S/LC = \sum_{a \in A_l} \sum_{m \in A_c} p_{a,m}(a,m) log \frac{p_{a,m}(a,m)}{p_a(a) p_m(m)} / Z \tag{1}$$

   Here, Z is the average entropy of the two marginal distributions: $Z = \frac{(H(p_a)+H(p_m))}{2}$. As in the previous sections, A_l and A_c denote the set of available actions and the set of available messages respectively.

   Note that this metric is computed twice for every agent (separately for the speaker and the listener roles). To obtain the final metric characterising the behaviour of an agent population, we average individual metrics over all agents and roles.

2. **Communication Asymmetry**:

   a. Between-agent signal-action mapping divergence. Even if the messages and actions are coordinated in each agent, it may still be the case that that different agents learn different "languages".

   To quantify the extent of such language difference, we propose to use the average Jensen-Shannon pairwise divergence between distributions of agents' actions following a specific signal (averaged over all signals and pairs of agents).

   For a pair of agents 1 and 2, the metric is defined as follows

$$\sum_{m \in A_c} JSD(p_{a_1|m}, p_{a_2|m}) / |A_c| \tag{2}$$

   Where $p_{(a_1|m)}$ are action distributions of agent 1 conditional on the message (received or sent) being equal to m. We report this metric separately for all pairs of speakers and for all pairs of listeners.

   c. Within-agent signal-action mapping divergence.

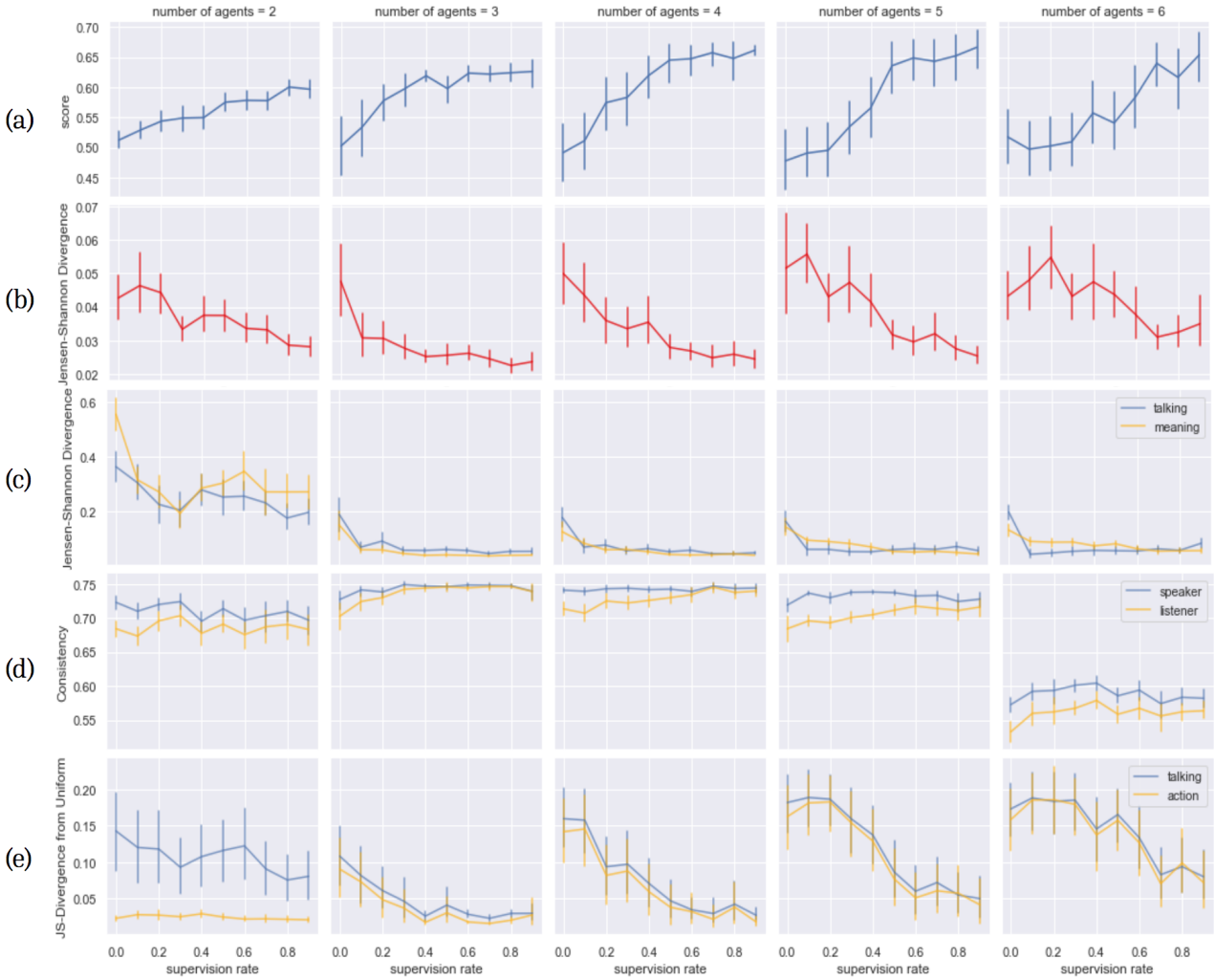

Figure 2: The Communication Analysis Metrics with 95% Confidence Intervals for the Experiment 1. a) Average rewards b) Between-agent signal-action mapping divergence c) Talking divergence (blue) and Within-agent signal-action mapping divergence (yellow) d) Average speaking (blue) and listening (yellow) consistencies e) Average predictability of agents' signals (blue) and actions (yellow).

Another important aspect of the learnt communication is whether an agent acts similarly when it produces and receives a specific message.

To quantify this effect, we use the average Jensen-Shannon divergence between the conditional distributions (conditioned on receiving or sending a specific message) over agent's actions when the agent plays the speaker and the listener role. To capture this effect, we use the same formula as in Equation 2, with the only difference that $p_{a_1|m}$ and $p_{a_2|m}$ now correspond to the same agent's distributions when this agent plays different roles (as opposed to distributions of a pair of different agents playing the same role). The metric is averaged over all agents.

b. Talking divergence.

To measure the variation in individual agents' messaging preferences, we report the average pairwise Jensen-Shannon divergence of marginal signaling distributions of different agents.

3. **Behavioral Predictability**. Lastly, to assess whether agents diversify their actions and messages, we report the Jensen-Shannon divergences between marginal distributions of agent's actions and messages and the uniform distribution.

Experiments

For agents' optimization, we used the Adam method with learning rate 0.0001 and momentum (β) 0.3. The memory and replay sizes were set to 10; the exploration decay and

the minimal exploration rate were set to 0.999 and 0.1 respectively. For each simulation, the agent population played 60000 coordination games total, and the agents updated their weights after each game they played. Each agent's weights were updated independently from the other agents.

Preliminary experiments: model capacity assessment and calibration

Before evaluating the effects of the main variables we wanted to study (self-play, supervision rate, and population size), we performed a number of calibration experiments, in order to make sure that the proposed model architecture and learning algorithms are adequate for the task at hand. For these preliminary experiments, we used a one-way communication problem setting, and fixed one of the agents (either the speaker or the listener) to an arbitrary fixed mapping.

Experiment 1: self-play and supervision rate

We simulated 2-agent coordination game learning with and without fictitious self-play for supervision rate varying from 0 to 0.9 (20 simulations per condition). In the self-play condition, the agents were randomly assigned to either play with each other or with their own decisions in the previous game. The agents played 60000 games, which includes the games with themselves for the self-play condition.

We excluded the self-play games' data from the subsequent analysis. We computed the average agents' scores for the last 2000 games, and the average between- and within-agent signal-mapping divergence, "talking" divergence, speaker and listener consistencies, as well as the average divergence of agents' signaling and acting distributions from the uniform rates for the last 4000 games that the agents played with each other (Figure 1).

Experiment 2: population size and supervision rate

We simulated the coordination game learning of 2 to 6 agents with supervision rate varying from 0 to 0.9 (20 samples per condition). For each condition, the agents played 60000 games total, meaning that the number of training interactions and weight updates per agent decreased with the increasing population size ($\frac{60000}{n}$).

We analyzed the data from this experiment in the same way as for the experiment 1. All the metric scores were averaged for each agent/pair of agents and for each signal (when necessary).

Results

Preliminary experiments

Preliminary experiments demonstrated that the agents quickly reach ceiling performance when one of them is exhibiting arbitrary deterministic performance. The type of function that the model needs to represent in order to solve the problem in multi-agent independent learning setting is the same as in our preliminary experiments. Therefore, we can be confident that any difficulties that agents may encounter in other experiments can be seen as a consequence of the multi-agent interactions, and not the model's representative power.

Supervision rate

Supervising feedback improved the overall performance of all the groups in our experiments (Figure 1a and 2a). There was a negative effect of supervision rate on the asymmetry scores: JS-divergence of between-agent signal-action mappings (Figure 1b and 2b), within-agent signal-action mappings, and talking distributions of different agents (Figure 1c and 2c). Supervision rate negatively affected behavioral predictability exhibited by all the groups in our experiments (Figure 1e and 2e).

Self-play

The self-play condition made the task more challenging, which is indicated by the lower reward scores than in the control, no self-play, condition (Figure 1a). However, agents in the self-play group demonstrated much lower between- and within-agent signal-action mapping divergence (Figure 1b and 1c), as well as lower divergence of talking distributions (Figure 1c) than the control group. Agents in self-play condition had more consistent responses to signals in both listener and speaker roles than the agents learning with no self-play (Figure 1d). Agents in both self-play and the alternative conditions achieved low behavioral (action) predictability. In the self-play group, however, the signaling distribution was coordinated with the action distribution, whereas for the alternative group these distributions were highly disparate (Figure 1e).

Population size

Learning in bigger populations was more challenging, but it was greatly leveraged by the supervising feedback (Figure 2a). All communication asymmetry scores were much lower for populations of 3-6 agents than for the dyad learning condition (Figure 2b and 2c). The 3-5 agents' groups exhibited higher speaker and listener consistencies than the 2 and 6 agents' groups (Figure 2d), the listener consistency was lower than the speaker consistency in all the groups. We explain the decreased consistency scores of the 6 agents' population as a result of smaller amount of training updates per agent (10000 in this case).

Agents learning in 2-agent condition demonstrated low action predictability, which was not aligned with the signaling distribution. The agents learning in bigger groups exhibited higher action predictability solutions, which decreased with the supervision rate. The signaling and acting predictability scores were coordinated in the 3-6 agents' populations (Figure 2e).

Discussion

First of all, our simulations replicate the result that naive reinforcement learning feedback is not sufficient to ensure that independently adapting communicative agents converge on effective strategies, even in a simple coordination game with 4 actions.

Motivated by examples of naturalistic communication learning, we investigated whether such variables as population size, limited supervision, and self-play (inner speech), can foster the development of efficient communication protocols. To isolate such effects, we followed the general principles advocated for in (Lowe et al., 2019), significantly expanding, however, the set of proposed metrics.

We found that independent learning in a group of two agents leads to the development of internally and externally asymmetric communication protocols. These between- and within-agent asymmetries are ameliorated by either introducing self-play or increasing the agent population. Adding supervision positively affected agents' performance across all metrics, but was not as efficient in solving the asymmetry issues as introducing self-play or increasing the population size.

We believe that one of the main merits of our contribution is proposing a number of metrics that allowed us to isolate different components of communication protocols learned by agents. In line with (Lowe et al., 2019), we found that using the average reward metric alone does not provide enough information about the success of communication learning (even in our simple case), and that it can even be misleading.

For example, the agents in no self-play condition achieve higher reward scores than the agents in the alternative, self-play, condition (Figure 1). However, further examination reveals that this result is mainly associated with a more uniform action distribution exhibited by the agents with no self-play (and hence, a smaller diversity penalty), and not by a development of a more symmetric and consistent communication system. While the non-uniform penalty is specific to our setup, most contemporary reinforcement learning studies include additional "shaping rewards" of similar nature to avoid trivial solutions. This further supports the idea of using a set of more specific metrics in addition to the average reward.

Our extended set of metrics also allowed us to achieve a more nuanced understanding of the population size effects. The connection between the population size and the success of communication learning was recently reported in Graesser et al. (2020), but the authors only examined the "self-play success" of the agents as a measure of communication symmetry. Our results provide a new insight on this improvement: perhaps surprisingly, increasing the population size mostly affects the within-agent symmetry (similar to "self-play success" used in Graesser et al. (2020)), while the between-agent mappings may still stay asymmetric (see Figure 2 b and c).

The relative simplicity of our experimental setup presents a sharp contrast with a number of recent studies. Our results demonstrate, however, that such a setting is sufficiently rich to capture some of the key challenging aspects of multi-agent reinforcement learning. We believe that it motivates a partial shift in research practices from necessarily aiming to expand the set of currently solvable problems to focusing on a more nuanced analysis of limitations and properties of already available solutions. In our view, such a shift may yield numerous insights for the studies of language evolution and we hope to inspire more works in this direction.

References

Batali, J. (1994). Innate biases and critical periods: Combining evolution and learning in the acquisition of syntax. In *Artificial life IV*, pages 160–171.

Bernstein, D. S., Givan, R., Immerman, N., and Zilberstein, S. (2002). The complexity of decentralized control of markov decision processes. *Mathematics of operations research*, 27(4):819–840.

Bogin, B., Geva, M., and Berant, J. (2019). Emergence of communication in an interactive world with consistent speakers. *arXiv:1809.00549 [cs]*. arXiv: 1809.00549.

Cangelosi, A. and Parisi, D. (1998). The emergence of a'language'in an evolving population of neural networks. *Connection Science*, 10(2):83–97.

Chaabouni, R., Kharitonov, E., Dupoux, E., and Baroni, M. (2019). *Anti-efficient encoding in emergent communication*, page 6293–6303. Curran Associates, Inc.

Christiansen, M. H. and Kirby, S. (2003). *Language evolution*. OUP Oxford.

Eccles, T., Bachrach, Y., Lever, G., Lazaridou, A., and Graepel, T. (2019). *Biases for Emergent Communication in Multi-agent Reinforcement Learning*, page 13111–13121. Curran Associates, Inc.

Evtimova, K., Drozdov, A., Kiela, D., and Cho, K. (2018). Emergent communication in a multi-modal, multi-step referential game. *arXiv:1705.10369 [cs, math]*. arXiv: 1705.10369.

Foerster, J., Assael, I. A., de Freitas, N., and Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2137–2145.

Foerster, J., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. (2018). Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, page 122–130. International Foundation for Autonomous Agents and Multiagent Systems.

Gong, T. (2009). Computational simulation in evolutionary linguistics: A study on language emergence.

Gong, T. (2011). Simulating the coevolution of compositionality and word order regularity. *Interaction Studies*, 12(1):63–106.

Graesser, L., Cho, K., and Kiela, D. (2020). Emergent linguistic phenomena in multi-agent communication games. *arXiv:1901.08706 [cs]*. arXiv: 1901.08706.

Griffiths, T. L. and Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, 31(3):441–480.

Hashimoto, T. (2002). *The Constructive Approach to the Dynamic View of Language*, page 307–324. Springer London.

Hashimoto, T. and Ikegami, T. (1996). Emergence of net-grammar in communicating agents. *Biosystems*, 38(1):1–14.

Havrylov, S. and Titov, I. (2017). *Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols*, page 2149–2159. Curran Associates, Inc.

Kirby, S. (2002). Natural language from artificial life. *Artificial Life*, 8(2):185–215.

Kirby, S., Griffiths, T., and Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, 28:108–114.

Laurent, G. J., Matignon, L., Fort-Piat, L., et al. (2011). The world of independent learners is not markovian. *International Journal of Knowledge-based and Intelligent Engineering Systems*, 15(1):55–64.

Lazaridou, A., Hermann, K. M., Tuyls, K., and Clark, S. (2018). Emergence of linguistic communication from referential games with symbolic and pixel input. *arXiv:1804.03984 [cs]*. arXiv: 1804.03984.

Lazaridou, A., Pham, N. T., and Baroni, M. (2016). Towards multi-agent communication-based language learning. *arXiv:1605.07133 [cs]*. arXiv: 1605.07133.

Lowe, R., Foerster, J., Boureau, Y.-L., Pineau, J., and Dauphin, Y. (2019). On the pitfalls of measuring emergent communication. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 693–701. International Foundation for Autonomous Agents and Multiagent Systems.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, O. P., and Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, pages 6379–6390.

Matignon, L., Laurent, G. J., and Le Fort-Piat, N. (2012). Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *The Knowledge Engineering Review*, 27(1):1–31.

Mordatch, I. and Abbeel, P. (2018). Emergence of grounded compositional language in multi-agent populations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Nowak, M. A., Komarova, N. L., and Niyogi, P. (2002). Computational and evolutionary aspects of language. *Nature*, 417(6889):611–617.

Nowak, M. A. and Krakauer, D. C. (1999). The evolution of language. *Proceedings of the National Academy of Sciences*, 96(14):8028–8033.

Nowak, M. A., Plotkin, J. B., and Krakauer, D. C. (1999). The evolutionary language game. *Journal of Theoretical Biology*, 200(2):147–162.

Oudeyer, P.-Y. (2005). The self-organization of speech sounds. *Journal of Theoretical Biology*, 233(3):435–449.

Oudeyer, P.-Y. (2006). *Self-organization in the evolution of speech*, volume 6. OUP Oxford.

Pesce, E. and Montana, G. (2020). Improving coordination in small-scale multi-agent deep reinforcement learning through memory-driven communication. *Machine Learning*.

Reali, F. and Griffiths, T. L. (2010). Words as alleles: connecting language evolution with bayesian learners to models of genetic drift. *Proceedings of the Royal Society B: Biological Sciences*, 277(1680):429–436.

Snowdon, C. T. (1990). Language capacities of nonhuman animals. *American Journal of Physical Anthropology*, 33(S11):215–243.

Spranger, M. and Steels, L. (2012). *Emergent functional grammar for space*, volume 3, page 207–232. John Benjamins Publishing Company.

Steels, L. (2015). *The talking heads experiment*. Best of Publishing.

Sukhbaatar, S., Fergus, R., et al. (2016). Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems*, pages 2244–2252.

The Five Graces Group, Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., and et al. (2009). Language is a complex adaptive system: Position paper. *Language Learning*, 59:1–26.