Towards Situated Communication in Multi-Step Interactions: Time is a Key Pressure in Communication Emergence

Aleksandra Kalinowska (ola@u.northwestern.edu)
Northwestern University, Evanston, IL and DeepMind, Edmonton, AB
Elnaz Davoodi (elnazd@deepmind.com) and **Kory W. Mathewson (korymath@deepmind.com)**
DeepMind, Montreal, QC, Canada
Todd D. Murphey (t-murphey@northwestern.edu)
Northwestern University, Evanston, IL
Patrick Pilarski (ppilarski@deepmind.com)
University of Alberta, Edmonton, AB and DeepMind, Edmonton, AB

Abstract

Enabling efficient communication in artificial agents brings us closer to machines that can cooperate with each other and with human partners. Hand-engineered approaches have substantial limitations, leading to increased interest in methods for communication to emerge autonomously between artificial agents. Most of the research in the field explores *unsituated* communication in one-step referential tasks. The tasks are not temporally interactive and lack time pressures typically present in natural communication and language learning. In these settings, agents can successfully learn *what* to communicate but not *when* or *whether* to communicate. Here, we extend the literature by assessing emergence of communication between reinforcement learning agents in a temporally interactive, cooperative task of navigating a gridworld environment. We show that, through multi-step interactions, agents develop just-in-time messaging protocols that enable them to successfully solve the task. With memory—which provides flexibility around message timing—agent pairs converge to a look-ahead communication protocol, finding an optimal solution to the task more quickly than without memory. Lastly, we explore *situated* communication, enabling the acting agent to choose when and *whether* to communicate. With the opportunity cost of forgoing an action to communicate, the acting agent learns to solicit information sparingly, in line with the Gricean Maxim of quantity. Our results point towards the importance of studying language emergence through situated communication in multi-step interactions.

Keywords: emergent communication; reinforcement learning; artificial agents; cooperative game

Introduction

Communication is a key skill for collaboration and hence largely beneficial in multi-agent settings. As humans, we share well-established communication protocols that have evolved over thousands of generations to suit the needs of our daily tasks and to take advantage of our cognitive and physical capabilities. As an example, natural languages are known to be compositional, making them easier to learn and use (Kirby & Hurford, 2002). Similarly, when we communicate, we are known to try to be as informative as possible, giving only as much information as is needed (Grice, 1975). If future artificial systems are to cooperate with humans, it will be beneficial for their communication protocols to follow these patterns. Studying communication emergence among artificial agents supports the design of machines that will work well with each other and with people (Crandall et al., 2018; Steels, 2003).

With a recent increase in available computational power, the field has seen a lot of progress (Wagner, Reggia, Uriagereka, & Wilkinson, 2003; Lazaridou & Baroni, 2020). Thus far, emergent communication has largely been studied in one-step referential games, such as the Lewis signalling task (Chaabouni, Kharitonov, Dupoux, & Baroni, 2019; Li & Bowling, 2019; Lazaridou, Hermann, Tuyls, & Clark, 2018). This type of learning environment is known to successfully enable language development (Kirby & Hurford, 2002) but does not allow agents to accelerate the learning process through back-and-forth interaction. Simulated communication emergence has also been studied in other game-based environments, some allowing multi-step interaction, such as a 2-player negotiation task (Cao et al., 2018) or a multi-modal referential game (Evtimova, Drozdov, Kiela, & Cho, 2018). Initial results show benefits of multi-step dialogue for communication emergence (Evtimova et al., 2018). Here, we build on this idea and evaluate the consequences of allowing multi-step communication in a multi-step task.

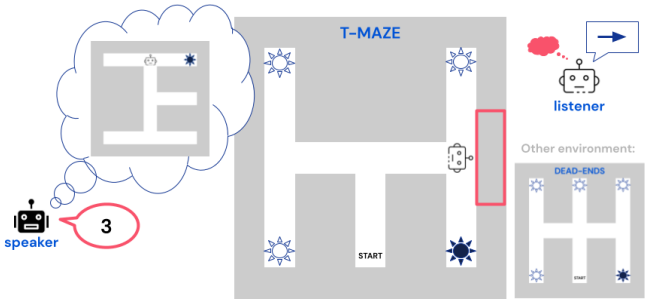

Figure 1: **Experimental setup.** The speaker sees a pixel-based representation of the maze and can broadcast messages; the listener sees either no environmental context or has partial visibility (3 pixels in front as indicated by the colored box) and can navigate the maze to reach the goal. Stars indicate possible goal locations.

In most studies, the emerged language structures are analyzed for shared commonalities with natural languages, such as compositionality or encoding efficiency. Although desired, it is nontrivial for such properties to emerge spontaneously between artificial agents (Kottur, Moura, Lee, & Batra, 2017). For instance, artificial agents tend towards an anti-efficient encoding (Chaabouni et al., 2019). This likely happens because in the Lewis signalling task, as well as in other simulated environments (Cao et al., 2018), agents have no incentive to be concise. The communication is not situated in the task and hence excessive use of communication does not neg-

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

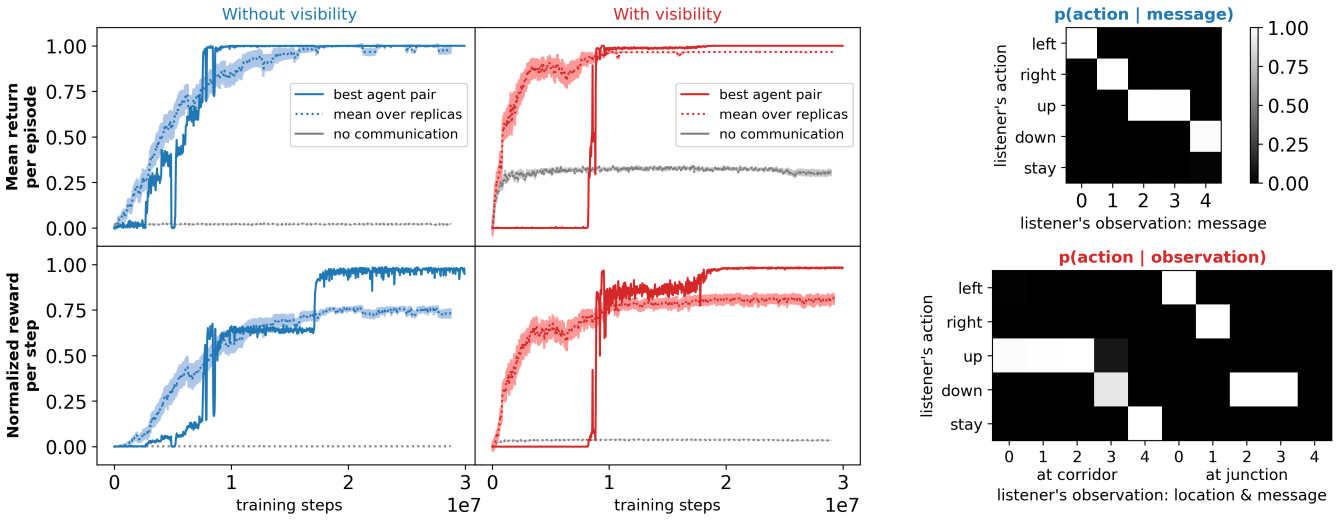

Figure 2: **Agents without memory in a T-maze environment.** Both agents with and without visibility learn to solve the task via the shortest path, as indicated by the normalized reward per step converging to 1. For the best agent pairs there is no ambiguity in the messages—each message corresponds to exactly one action. Some messages have synonyms—the listener interprets multiple messages in the same manner.

atively affect the outcome of the game (or cause agent frustration), as it might in a real-world situation (Steels & Brooks, 1995). In real life, there might be an opportunity cost to language that is not reflected in the learning environment of the Lewis signalling task. Initial work shows that it is possible to incentivise efficient communication by modifying the agents' reward structure, e.g., by adding an internal cost of articulation (Rita, Chaabouni, & Dupoux, 2020). In our approach, we show it is possible to obtain sparse communication by providing the agent with an action-communication trade-off, in line with the idea that *reward is enough to shape language* (Silver, Singh, Precup, & Sutton, 2021). We provide the listener (i.e. acting agent) with agency to reason about the timing of communication and about whether to communicate at all.

In our experiments, we explore the emergence of communication in a cooperative multi-step navigation task. We place agents in a pixel-based gridworld setting, expanding on the work of (Kajić, Aygün, & Precup, 2020). In the task, the speaker observes a maze with a goal location and communicates information while the listener navigates the maze, obtaining a reward for both agents at the goal. Like humans or robots that can only observe a small part of the world in their proximity, the listener has a limited view of its environment and has to rely on the speaker for guidance (Denis, Pazzaglia, Cornoldi, & Bertolo, 1999). In the first set of experiments, we implement *unsituated* communication—the speaker's message gets broadcasted to the listener at each step of the task, similarly to the communication setup in prior work. In the second set of experiments, we *situate* the communication in the environment—we allow the acting agent to actively choose between (i) taking an action to move through the maze and (ii) soliciting information from the speaker.

Our contributions are three-fold: (1) we study the emergence of *unsituated* communication through multi-step interaction, (2) we explore the effect of agents having *memory*—and the ability to reason about message timing—on the emerged communication protocol and agents' ability to converge to a collaborative solution, and (3) we investigate how *situating* communication in the task affects the communication protocol and overall task performance.

Experimental Setup

The environment. We define a cooperative navigation task as a Markov Decision Process (MDP). The environment is set up as a pixel-based gridworld. Each gridworld is 7 by 7 cells with different maze patterns, as illustrated in Figure 1. We refer to the two tested gridworld patterns as a T-maze and a dead-ends environment. Features of the world are represented with colors: walls are black, the maze is white, the agent is green, and the target is blue, as visualized in Figure 4. The features are encoded with binary vectors.

The agents. There are two agents, a speaker and a listener (i.e. acting agent). The listener is embedded inside the gridworld and can take actions to move between cells. The action space of the listener spans 5 actions [move up, move down, move right, move left, stay in place]. The listener's observation consists of the environmental view (if any) concatenated with the message from the speaker. We test the listener under two conditions: (1) with no visibility, where the listener's observation consists solely of the speaker's message, and (2) with partial visibility, where the listener can see the 3 pixels directly in front of them. The second variant gives the listener environmental context to take actions without needing to rely solely on communication.

The speaker does not reside within the gridworld and cannot take environmental actions (i.e. navigate the maze) but instead can communicate information to the listener. The message space of the speaker spans 5 symbols $[0, 1, ..., 4]$. At each timestep, the speaker can see the entire gridworld, including the location of the agent and the location of the goal. The

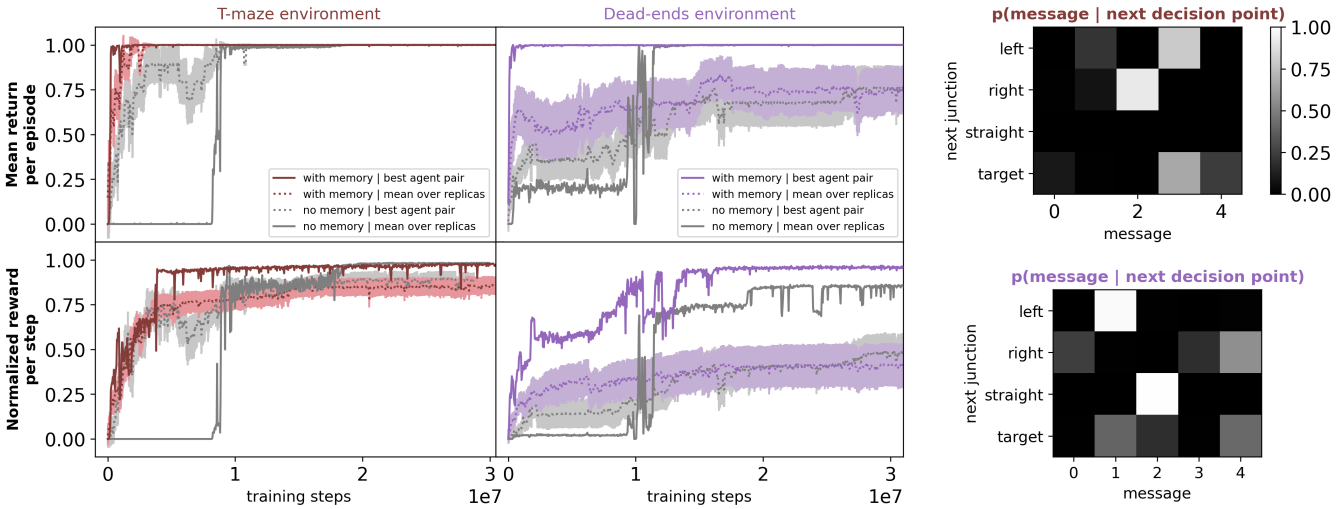

Figure 3: **Comparison of agents with and without memory in two environments; listener has partial visibility.** Memory plays a role in shaping the learned language. (1) Instead of using just-in-time communication, agents gravitate towards using a look-ahead communication protocol. Already before the listener reaches a junction, the speaker broadcasts a consistent message, signalling the correct action at the next junction. (2) Memory improves convergence. Note that the grey lines converge slower and plateau at lower values.

speaker's view of the world map is rotated to align with the direction that the listener is facing.

In our experiments, we test agents with and without memory. Agents without memory have to rely only on their current observations to generate messages or pick actions. Agents with memory have an internal representation of the history of an episode—they can use accumulated knowledge from prior timesteps to make decisions in the current timestep.

Agent architectures. The speaker and the listener are designed as two independent reinforcement learning (RL) agents. Both agents have the same architecture without sharing weights or gradient values. They both have a 2-layer Convolutional Neural Network (CNN) that generates a $8-32$ bit representation of the environment. In the case of the listener, this representation of the environment gets concatenated with the message received from the speaker. In both cases, the vector gets passed into a fully connected layer that generates the agent's action. Agents with memory have an additional single-layer LSTM (Hochreiter & Schmidhuber, 1997) after their fully connected layer.

We train the agents using neural fitted Q learning (Riedmiller, 2005), with an Adam optimizer (Kingma & Ba, 2015) and $Q_t(\lambda)$ with $\lambda = 0.9$. The Q values are updated using temporal difference (TD) error where the bootstrapped $Q_t(\lambda)$ is defined as follows:

$$Q_t(\lambda) = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}Q_t^{(n)}$$

During training, agents use an ε-greedy policy with the exploration rate set as $\varepsilon = 0.01$.

The task. The goal of the agents is to cooperate so that the listener reaches the target. In each experimental episode, both agents receive a reward $R = 1$ if the listener reaches the target before the episode terminates. Episode timeout is set to 100 steps. The goal locations are randomly assigned to one of 4 corners in the T-maze or one of 5 corridor ends in the dead-ends maze, as indicated with stars in Figure 1. In each episode, the listener agent starts from the bottom middle cell.

Communication modes. We compare two modes of communication: (1) real-time messaging with a passive listener, and (2) real-time messaging with an active listener. In mode 1, the speaker generates a 1-token message at every timestep and the message gets broadcasted to the listener before they choose an action. The speaker has to reason about both the content and timing of their message, deciding both *what* and *when* to communicate. In mode 2, we implement real-time messaging with an active listener. Here, the message is only broadcasted to the listener after they ask for information. The active listener can solicit to receive information in the next timestep by choosing to stay in place at the current timestep. The listener has to learn *whether* to communicate at all. In our implementation, the speaker still generates a message at every timestep, even though it might not be shared with the active listener.

We define the communication in mode 1 as *unsituated*—it is free and guaranteed to the agent at every timestep. There is no opportunity cost to communication. The communication in mode 2 is *situated*—we allow the acting agent to actively choose between (i) taking an environmental action and (ii) soliciting information from the speaker. As a result, the active listener experiences an opportunity cost to communication. They have to forego a move in the environment (that could bring them closer to the target) in order to obtain information from the speaker and make an informed decision.

Evaluation metrics. We evaluate agent performance using 3 metrics: (1) task success (via the mean return per episode), (2) optimality of task solution (via the normalized reward per step), and (3) communication sparsity (via the number of asks

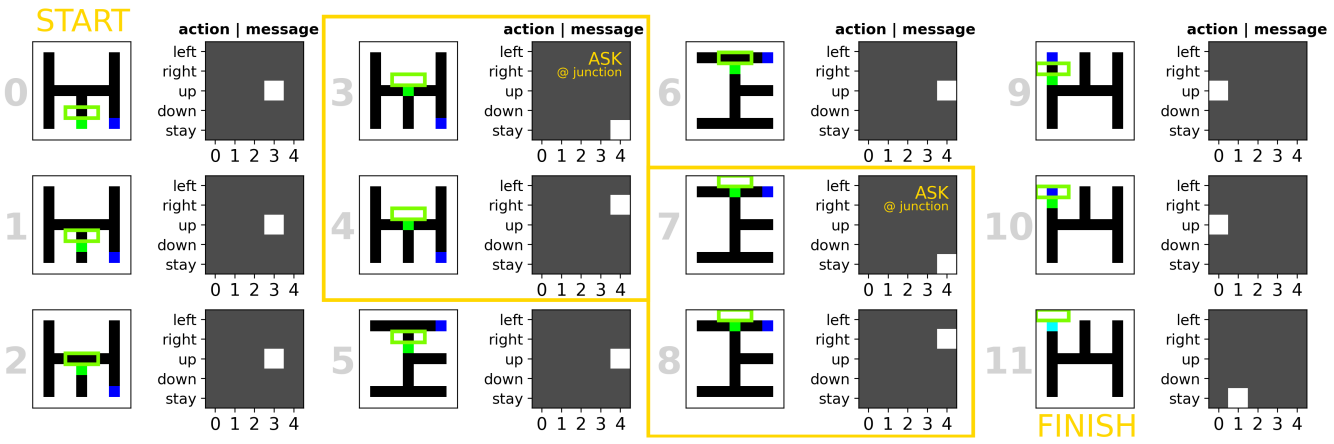

Figure 4: **An episode walk-through of an active listener with partial visibility.** The listener learns to solve the task optimally, deciding to stay and ask for information when at a junction (twice during the episode). At each of the 11 timesteps, we visualize (left) the speaker's view of the board with an overlaid green box indicating the listener's view, and (right) the speaker's message and listener's action at that step.

per episode). The metric of task success represents the likelihood of the agents succeeding at reaching the target before episode timeout. At each training step, this metric is calculated as the average per-episode reward (0 or 1) over all completed episodes until that step. When ≈ 1, the agents are reliably reaching the target in each episode. The second metric quantifies the optimality of the path taken to solve the task. If the task is solved in the optimal number of steps ($n_{opt\ steps} = 9$ in the T-maze, $n_{opt\ steps} = 8.4$ in the dead-ends maze), the agents obtain a per-step reward of 1. The value is calculated as $n_{opt\ steps} * (R_{current\ episode}/n_{episode\ steps})$. Finally, the metric of communication sparsity quantifies the efficiency of information exchanged between the agents when communication is situated. The metric represents the average per-episode number of messages that the active listener requests from the speaker. Depending on the listener's characteristics: partial or no visibility, the optimal number of asks in the T-maze environment is equal to 2 or 9 asks per episode for agents without memory, and 1 ask per episode for agents with memory.

Hyperparameters. After an initial exploration, we use a reward discount $\gamma = 0.99$. For each experiment, we run a hyperparameter sweep over learning rates of the speaker and listener $\alpha = [10^{-5}, 10^{-6}, 10^{-7}]$ and over the size of the environmental representation $s = [4, 8, 16, 32]$. We run the simulation with each hyperparameter setting 10 times with different random seeds. In the results for each experiment, we present the best performing agent pair from our hyperparameter sweep and/or the mean over the 10 replicas with the same hyperparameters as the best performing pair. When we plot metric means, we include the standard error of the mean.

Results

Task validation. We start by generating a baseline for the task, experimentally validating that communication is required to solve it. In the T-maze environment, we compare agents with the communication channel open and inactive. We find that without communication agents are unable to reliably solve the task. When the listener has no visibility (see left panel in Figure 2), agents with no communication are unable to solve the task at all. Under partial visibility (middle panel), agents without communication can succeed in the task with a mean return of ≈ 0.25 per episode, taking close to 100 steps per episode to reach the target.

With memory, baseline performance improves. The listener (i.e. acting agent) is able to reliably traverse the maze in search of the target, particularly when equipped with partial visibility. However, due to the random location of the target, the listener cannot consistently solve the task in an optimal number of steps. The solution optimality of the best agent pairs with memory and no communication converges to a normalized reward per step of ≈ 0.45. When allowed to communicate, all agents in the T-maze environment learn to solve the task and best agent pairs find an optimal solution.

Agents learn timely communication through multi-step interaction. We evaluate communicating agents in the T-maze environment under two visibility conditions. Given no memory, agents learn a just-in-time communication protocol, as visible in the heatmaps in Figure 2. Agents agree on unambiguous meanings of messages and, in some cases, learn synonyms to signal the same environmental action. Under partial visibility, the meaning of messages depends on the environmental context (e.g. message 1 at the corridor is consistently interpreted by the listener as 'move up' but at the junction as 'move right'). Importantly, successful agents converge to a just-in-time protocol, where at each time step the listener can unambiguously interpret the speaker's message.

Memory improves communication emergence and influences timing in the established communication protocols. To investigate how memory can affect communication, we trained agents in the two environments. We find that memory improves time to convergence. Note in Figure 3 how the agents with memory converge faster than agents without memory. Memory can also improve the agents' ability to find

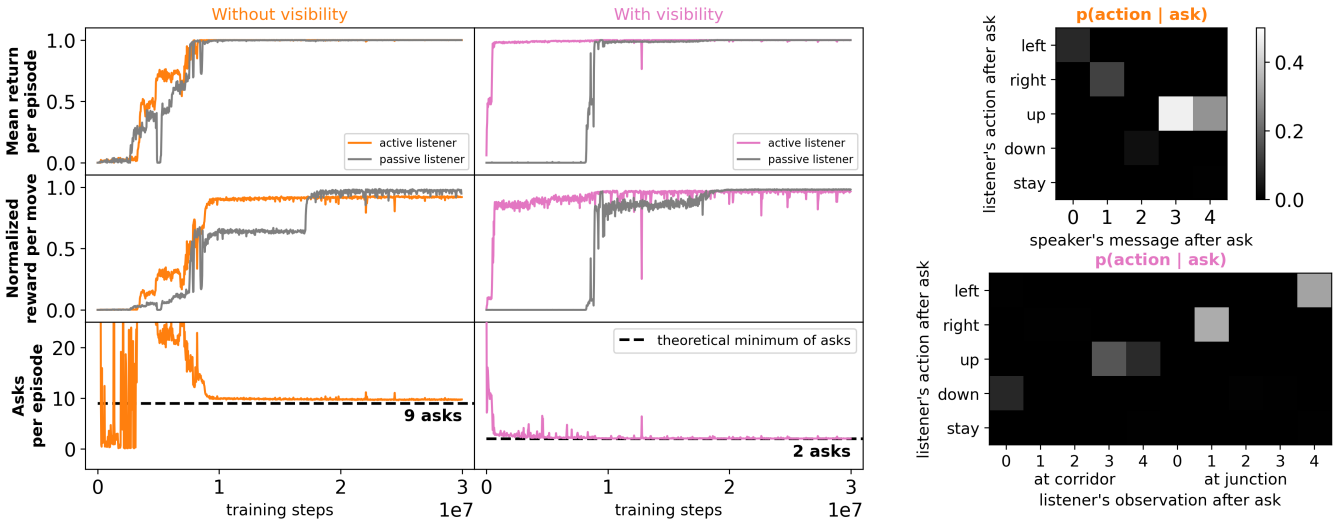

Figure 5: **Best performing pair of agents without memory with an active listener in a T-maze environment.** Both agents with and without visibility learn to solve the task via the shortest path. The listeners learn to query the speakers in the optimal number of asks (once per step when the listener has no visibility and once per junction when the listener sees environmental context). Note that the listeners ask for information frequently at the beginning of the interaction and gradually less over time.

an optimal solution. We observe this in the more difficult dead-ends environment, where even the best agent-pair without memory does not converge to an optimal solution.

What is more, the emerged protocol is different. The best-performing agents with memory learn a look-ahead communication protocol instead of deferring to communicating just-in-time like before. An example of this is that the speaker with memory might broadcast the same message for the first 4 steps of a T-maze episode, alerting the listener to make a left or right turn at the junction. In contrast, the successful agents with no memory would broadcast a 'go straight' message for the first few steps of an episode and a unique message for turning at the junction. As a result, with memory we no longer observe a one-to-one mapping between the messages broadcasted at each step by the speaker and the actions taken by the listener at that step. However, we still observe a consistent communication pattern and unambiguous messaging, as visible in the heatmaps in Figure 3. Empowering agents with memory results in them incoporating message timing into their newly developed communication protocol.

The pressure of time in a multi-step interaction can incentivise sparse communication. In the last set of experiments, we evaluate the impact of situated communication on language emergence. We train agent pairs with an active listener to solve the navigation task in a T-maze environment under two visibility conditions. Figure 4 shows a step-by-step example episode for a listener with partial visibility. The heatmaps in Figure 5 illustrate the communication protocol of the best agent pairs. Under the no visibility condition, the listener queries the speaker for 'move up', 'turn left', and 'turn right' actions, proportionally to their frequency in the task solution. Under the partial visibility condition, information solicitation takes place mostly at the junctions, where the acting agent has a choice between two viable environmental actions. The active listener can learn to near optimally solicit information, asking ≈ 9.76 and ≈ 2.06 times per episode under the two visibility conditions, respectively.

In Figure 5 on the left, we illustrate the learning curves of the best performing agent pairs. Note that the active listeners ask for information frequently at the beginning of the interaction and gradually less over time. This suggests that agents initially have opportunities to align on a protocol. Over time, listeners learn *when* and *whether* to solicit information as communication comes with a cost. We also observe that the best performing agent pairs with an active listener converge to an optimal solution faster than the best performing agent pairs with a passive listener. The results suggest that situated communication not only allows the agents to learn a sparse communication protocol, in line with the Gricean Maxim of quantity, but also has a positive impact on convergence speed.

The active listener exhibits a preference for just-in-time communication. Interestingly, when we test situated communication between agents with memory, agents continue to ask for information at the junctions, as visible in the heatmaps in Figure 6. This is non-obvious—given memory, the active listener could ask for information at any point in the maze. In fact, if the agent were to be optimally sparse, they could (1) ask for information only once at the beginning of an episode, (2) receive a message encoding the address of the target, and (3) follow the relevant policy from memory. Instead, the active listener with memory learns sparse communication relative to a passive listener but they do not achieve the theoretically maximum sparsity. An active listener with memory persists to ask for information at the junctions when it is immediately actionable. This speaks to the importance of allowing agents to learn the timing of communication. This result suggests that it is easier for agents to succeed at the task when they exchange information when it is immediately actionable.

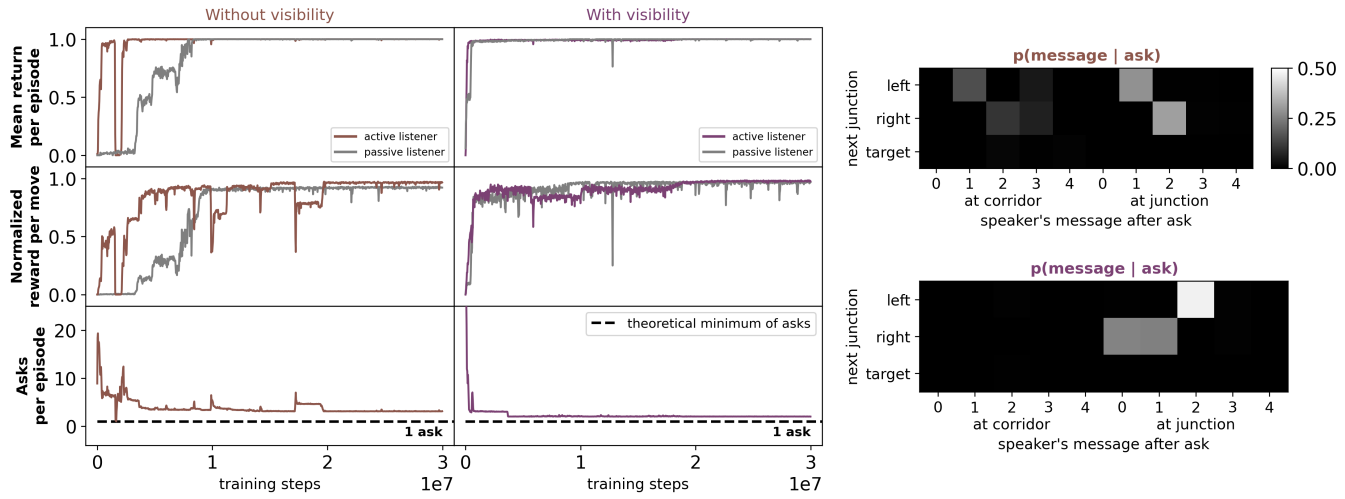

Figure 6: **Best performing pair of agents with memory with an active listener in a T-maze environment.** Both agents with and without visibility learn to solve the task via the shortest path. The listeners learn to query the speaker sparsely compared to a passive listener, but they do not converge to the theoretically minimal number of asks, persisting to ask for information when it is immediately actionable.

Conclusion & Discussion

Our results point towards the importance of studying emergent communication in multi-step interactions. The interactive aspect of communicating over time enables agents to learn both *what* and *when* to communicate. Secondly, we find that there is value in situating the communication in the task and giving the listener agency to choose *whether* to communicate at all. In this way, we improve convergence and allow the reward to shape the emergent communication protocol to exhibit properties of natural languages, such as sparsity. Our ongoing work will expand this idea and situate both the speaker and listener in the environment, allowing both agents to communicate and take actions in the gridworld environment.

References

Cao, K., Lazaridou, A., Lanctot, M., Leibo, J. Z., Tuyls, K., & Clark, S. (2018). Emergent communication through negotiation. *Int. Conf. on Learning Representations*.

Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2019). Anti-efficient encoding in emergent communication. *Advances in Neural Information Processing Systems*.

Crandall, J. W., Oudah, M., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., ... others (2018). Cooperating with machines. *Nature Communications*.

Denis, M., Pazzaglia, F., Cornoldi, C., & Bertolo, L. (1999). Spatial discourse and navigation: An analysis of route directions in the city of Venice. *Applied Cog. Psychology*.

Evtimova, K., Drozdov, A., Kiela, D., & Cho, K. (2018). Emergent communication in a multi-modal, multi-step referential game. *Int. Conf. on Learning Representations*.

Grice, H. P. (1975). Logic and conversation. *Speech Acts*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*.

Kajić, I., Aygün, E., & Precup, D. (2020). Learning to cooperate: Emergent communication in multi-agent navigation. *Meeting of the Cognitive Science Society*.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *Int. Conf. on Learning Representations*.

Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. *Simulating the Evolution of Language*.

Kottur, S., Moura, J. M., Lee, S., & Batra, D. (2017). Natural language does not emerge 'naturally' in multi-agent dialog. *Conf. on Empirical Methods in NLP*.

Lazaridou, A., & Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*.

Lazaridou, A., Hermann, K. M., Tuyls, K., & Clark, S. (2018). Emergence of linguistic communication from referential games with symbolic and pixel input. *Int. Conf. on Learning Representations*.

Li, F., & Bowling, M. (2019). Ease-of-teaching and language structure from emergent communication. *Advances in Neural Information Processing Systems*.

Riedmiller, M. (2005). Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method. *European Conf. on Machine Learning*.

Rita, M., Chaabouni, R., & Dupoux, E. (2020). "LazImpa": Lazy and impatient neural agents learn to communicate efficiently. *Conf. on Computational Natural Language Learning*.

Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*.

Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Sciences*.

Steels, L., & Brooks, R. (1995). *The artificial life route to artificial intelligence: Building embodied, situated agents.*

Wagner, K., Reggia, J. A., Uriagereka, J., & Wilkinson, G. S. (2003). Progress in the simulation of emergent communication and language. *Adaptive Behavior*.