Symmetric Machine Theory of Mind

Melanie Sclar [1 *] **Graham Neubig** [2] **Yonatan Bisk** [2]

Abstract

Theory of mind, the ability to model others' thoughts and desires, is a cornerstone of human social intelligence. This makes it an important challenge for the machine learning community, but previous works mainly attempt to design agents that model the "mental state" of others as passive observers or in specific predefined roles, such as in speaker-listener scenarios. In contrast, we propose to model machine theory of mind in a more general *symmetric* scenario. We introduce a multi-agent environment SymmToM where, like in real life, all agents can speak, listen, see other agents, and move freely through the world. Effective strategies to maximize an agent's reward require it to develop a theory of mind. We show that reinforcement learning agents that model the mental states of others achieve significant performance improvements over agents with no such theory of mind model. Importantly, our best agents still fail to achieve performance comparable to agents with access to the gold-standard mental state of other agents, demonstrating that the modeling of theory of mind in multi-agent scenarios is very much an open challenge. Code can be found at https://github.com/msclar/symmtom.

1. Introduction

Human communication is shaped by the desire to efficiently cooperate and achieve communicative goals (Tomasello, 2009). Children quickly learn that other people have independent mental states, and that communicating is necessary to obtain information from or shape the intentions of those they interact with. Remembering and reasoning over others' mental states ensures efficient communication by avoiding having to repeat information, and contributes to achieving common goals with minimal effort.

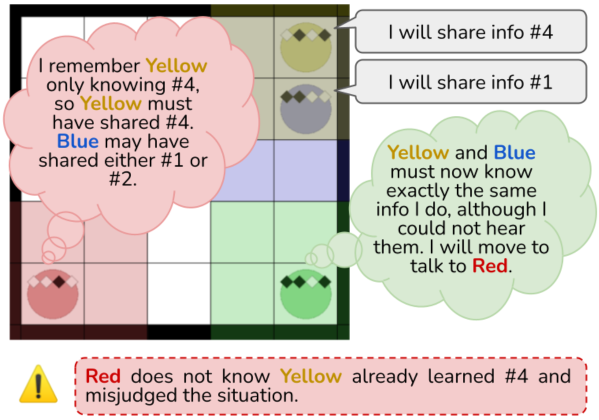

I will share info #4
I will share info #1
I remember Yellow only knowing #4, so Yellow must have shared #4. Blue may have shared either #1 or #2.
Yellow and Blue must now know exactly the same info I do, although I could not hear them. I will move to talk to Red.
Red does not know Yellow already learned #4 and misjudged the situation.

Figure 1. In SymmToM, agents aim to gain all available information (depicted as diamonds, black for known, white for unknown). Since hearing is limited to its neighbor cells, they must guess what happened beyond this range. Agents can see the whole grid, but mistakes in inferences may happen (as with the red agent).

Because of this, there is growing interest in developing agents that can exhibit this kind of behavior, referred to as Theory of Mind (ToM) by developmental psychologists (Premack & Woodruff, 1978). Previous work on agents imbued with such capabilities has focused mainly on two types of tasks. The former are tasks where the agent is a passive observer of a scene that has to predict the future by reasoning over others' mental states. These tasks may involve natural language (Nematzadeh et al., 2018) or be purely spatial (Gandhi et al., 2021; Rabinowitz et al., 2018; Baker et al., 2011). The latter are tasks where the theory of mind agent has a specific role, such as "the speaker" in speaker-listener scenarios (Zhu et al., 2021).

In contrast, human cooperation and communication is often multi-party, and rarely assumes that people have singular pre-specified roles. Moreover, human interlocutors are seldom passive observers of a scene but instead active participants. These dynamics mean human communication has additional complexities, such as the coordination between theory of mind, planning, and action, that are not easily tested in previous work. Therefore, we develop a

[*] Work during an internship at CMU. [1] Paul G. Allen School of Computer Science & Engineering, University of Washington [2] Language Technologies Institute, Carnegie Mellon University. Correspondence to: <msclar@cs.washington.edu, {gneubig,ybisk}@cs.cmu.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

more flexible environment, SymmToM, where we can study what happens when all participants must act as both speaker and listener. SymmToM is a fully symmetric multi-agent environment where all agents can see, hear, speak, and move, and are active players of a simple information-gathering game. To solve SymmToM, agents need to exhibit different levels of theory of mind, as well as efficiently communicate through a simple channel with a fixed set of symbols.

SymmToM is partially observable for all agents: even if agents have full vision, hearing may be limited. This also differentiates SymmToM from prior work, as modeling may require *probabilistic theory of mind*. In other words, agents need to not only remember and infer other agents' knowledge based on what they saw, but also estimate the probability that certain events happened. This estimation may be performed by assuming other agents' optimal behavior and processing the partial information available.

Despite its simple action space, SymmToM both fulfills the properties required for symmetric theory of mind to arise (which will be discussed in the following section), and empirically cannot be completely solved either by using well-known multi-agent deep reinforcement learning (RL) models, or even by tailoring those models to our task. In addition, all dimensions of complexity can be easily scaled to be more or less challenging, and we demonstrate how to test for different levels of theory of mind with corresponding metrics. Given this simplicity, flexibility, and difficulty, we contend that the SymmToM environment is an attractive first step towards testing the ability of agents to develop symmetric machine theory of mind.

2. Theory-of-Mind (ToM) Agents

A Theory-of-Mind agent can be defined as a modification of the standard multi-agent RL paradigm, where the agents' policies are conditioned on their beliefs about others. Formally, we define a reinforcement learning problem $\mathcal{M}$ as a tuple of a state space $\mathcal{S}$, action space $\mathcal{A}$, state transition probability function $T \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0{,}1]$, and reward $R \in \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, i.e. $\mathcal{M} := \langle \mathcal{S}, \mathcal{A}, T, R \rangle$. In this setting, an agent learns a (possibly probabilistic) policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ mapping states to actions to maximize their reward.

In a multi-agent RL setting each agent can potentially have its own state space, action space, transition probabilities, and reward function, so we can define an instance of $\mathcal{M}_i = \langle \mathcal{S}_i, \mathcal{A}_i, T_i, R_i \rangle$ for each agent i. For convenience, we can also define a joint state space $\mathcal{S} = \bigcup_i \mathcal{S}_i$ that describes the entire world in which all agents are interacting. Importantly, in this setting each agent will have its own view of the entirety of the world, described by a conditional observation function $\omega_i : \mathcal{S} \rightarrow \Omega_i$ that maps from the state of the entire environment to only the information observable by agent i. Since theory of mind is the ability to know (and act upon) the knowledge that an agent has, agents with *no theory of mind* will follow a policy that depends only on their current (potentially partial or noisy) observation of their environment: $\pi_i(a_{i,t} \mid \omega_i(s_t))$. Agents with *zeroth order theory of mind* (Flobbe et al., 2008; Hedden & Zhang, 2002) can reason over their own knowledge. These agents will be stateful, $\pi_i(\cdot \,|\, \omega_i(s_t), h_t^{(i)})$, where $h_t^{(i)}$ is i's hidden state. Hidden states are always accessible to their owner, i.e. i has access to $h_t^{(i)}$.

Agents with capabilities of reasoning over other agents' mental states will need to estimate $h_t^{(j)}$ for $j \neq i$. We denote i's estimation of j's mental state in time t as $\hat{h}_t^{(i,j)}$:

$$\pi_i(\cdot \mid \omega_i(s_t), h_t^{(i)}, \hat{h}_t^{(i,1)} \dots \hat{h}_t^{(i,i-1)}, \hat{h}_t^{(i,i+1)} \dots \hat{h}_t^{(i,n)})$$

How do we estimate $\hat{h}_t^{(i,j)}$? As a function of i's (the predicting agent) previous hidden state $t-1$, i's observation in $t-1$, and i's prediction of the hidden states of every agent in the previous turn:

$$\hat{h}_t^{(i+1)} = f(h_{t-1}^{(i)}, \omega_i(s_{t-1}), \hat{h}_{t-1}^{(i,1)} \dots \hat{h}_{t-1}^{(i,i-1)}, \hat{h}_{t-1}^{(i,i+1)} \dots \hat{h}_{t-1}^{(i,n)})$$

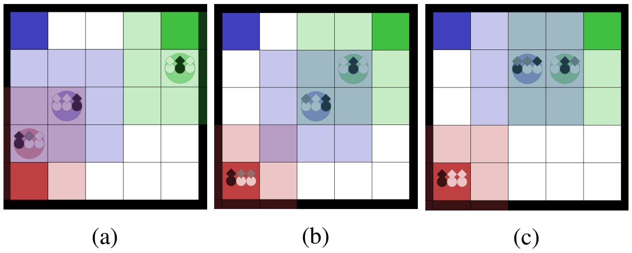
i's prediction of other agents' observation in $t-1$ is also crucial, but not explicitly mentioned since it can be computed using $\omega_i(s_{t-1})$. For the initial turn, $\hat{h}_0^{(i,j)}$ may be initialized depending on the problem: if initial knowledge is public, $\hat{h}_0^{(i,j)}$ is trivial; if not, $\hat{h}_0^{(i,j)}$ may be estimated.

3. Symmetric Theory-of-Mind

We define symmetric theory of mind environments as settings where theory of mind is required to perform a task successfully, and all agents have the same abilities. Having the same abilities means that all agents would have the same set of legal actions if placed in the same state (in terms of both location and knowledge), which is independent of the policy each agent executes. There are at least four defining characteristics for symmetric theory of mind to arise:

Symmetric action space. In symmetric theory of mind all agents are required to have the same action space (in contrast to, for example, theory of mind tasks in speaker-listener settings). Concretely, $\mathcal{A}_i = \mathcal{A}_j \neq \emptyset \; \forall i, j$.

Imperfect information. In perfect information scenarios all knowledge is public, making it impossible to have agents with different mental states. In theory of mind tasks in general, there could be a subset of agents with perfect information (e.g. a passive observer predicting future behavior). In symmetric theory of mind, since all agents have the same abilities and roles, all agents must have imperfect information. More precisely, ω_i –the subset



(a) (b) (c)

Figure 2. Example of three consecutive turns in an episode. There are three agents in a 5×5 grid, each with a hearing range of 1 (shaded in the same color of the agent). Fully-colored cells depict recharge bases. Information is represented by diamonds: black, gray, and white diamonds represent an information piece known first-hand, second-hand, and not known, respectively. Black circles show the information piece currently being said by each agent.

of the full state that agent i can observe if placed in each state– must not be the identity for any agent i.

Observation of others. Agents must have at least partial information of another agent to estimate its mental state. In contrast to passive-observer settings, in symmetric theory of mind every agent must be able to partially observe all others. More precisely, ω_i must observe at least partial information about $s_t^{(j)}$ (the subset of s_t that refers to agent j), although we do not require $s_t^{(j)} \neq \emptyset$ in every single turn. Moreover, if communication is allowed, it is desirable to partially observe or infer interactions between two or more agents to develop second order theory of mind (i.e. predicting what an agent thinks about what another agent is thinking) or higher.

Information-seeking behavior. It should be relevant for successfully performing the task to gather as much information as possible, and this information-gathering should involve some level of reasoning over other agent's knowledge. This is true for first-order theory of mind tasks in general, and can be formalized as $\pi^* \neq \pi$ for any zeroth-order theory of mind policy $\pi_i(\cdot \mid \omega_i(s_t), h_t^{(i)})$. In general, tasks can incorporate **perpetual** information seeking behaviors, to incentive efficient play even in long episodes. However, to achieve this with finish capacity, requires forgetting. Forgetting can be implemented as an explicit loss of knowledge under specific conditions, or degradation of memories. This introduces the concept of *information staleness*. Since information is not cumulative and the environment is only partially observable, agents will need to estimate whether what they knew to be true still holds in the present.

4. The SymmToM Environment

SymmToM is an environment where n agents are placed in a $w \times w$ grid world, and attempt to maximize their reward by gathering all the information available in the environment. Its construction mirrors the requirements specified above. There are c available information pieces, that each agent may or may not know initially. Information pieces known at the start of an episode are referred to as *first-hand information*. Each turn, agents may move through the grid to one of its four neighboring cells, and may speak exactly one of their currently known information pieces. More precisely, the action space of agent j is defined as follows:

$$\mathcal{A}_j = \{\textit{left}, \textit{right}, \textit{up}, \textit{down}, \textit{no move}\} \times \{1, \ldots, c\} \quad (1)$$

When an agent utters an information piece, it is heard by every agent in its hearing range (a $2h + 1 \times 2h + 1$ grid centered in each agent, with $2h + 1 < w$). The agents who heard the utterance can share this newly information with others in following turns. We refer to this as *second-hand information*, since it is learned –as opposed to *first-hand information*, given at the start of each episode. The state space is comprised of the position of the agents and their current knowledge:

$$\begin{aligned}\mathcal{S} = &\{\{(p_i, k_i), \text{ for } i \in \{1, \ldots, n\}\} \text{ where} \\ &p_i \in \{1, \ldots, w\} \times \{1, \ldots, w\}, \text{ and } k_i \in \{0,1\}^c\}\end{aligned}$$

Each agent aims to maximize their individual reward R_i via information seeking and sharing. Rewards are earned by hearing a new piece of information, giving someone else a new piece of information, or correctly using *recharge bases*. Recharge bases are special cells that reset an agent's knowledge in exchange for a large reward (e.g. $(n-1)c$ times the reward for listening to or sharing new information). Each agent has its own stationary recharge base during an episode. To trigger a base, an agent steps into its base having acquired all the available pieces of information, causing the agent to lose all the second-hand information it learned. Recharge bases guarantee that there is always reward to seek information. Concretely, let $s = \{(p_i, k_i), \text{ for } i \in \{1, \ldots, n\}\}$ be a state and $a_i = (a_i^{\text{dir}}, a_i^{\text{comm}}) \in \mathcal{A}_i$ an action, where a_i^{dir} represents the physical and a_i^{comm} the communicative action. We define agent i's reward R_i as the addition of three components. First, the reward for hearing new information, measured as the number of new information pieces heard by i. Second, the reward for hearing new information, computed as the number of agents that heard what i said and it was new to them. And lastly, the reward for using the recharge base correctly. Formally,

$$\begin{aligned}R_i(s,a_i) = &\sum_{i \neq j} \mathbb{1}\{||p_i - p_j||_\infty \leq h \text{ and } k_{i,a_j^{\text{comm}}} = 0\} \\ &+ \sum_{i \neq j} \mathbb{1}\{||p_i - p_j||_\infty \leq h \text{ and } k_{j,a_i^{\text{comm}}} = 0\} \\ &+ (a-1) \cdot c \cdot \mathbb{1}\{p_i = \text{base}_i \text{ and } k_j = \{1, \ldots, 1\}\}\end{aligned}$$

where $k_{i,a_j^{\text{comm}}} = 0$ represents that the a_j^{comm}-th element of k_i is unknown (i.e. zero).

A non-theory of mind agent can only achieve limited success. Without reasoning about its own knowledge (i.e. without *zeroth order theory of mind*), it does not know when to use a recharge base. Moreover, without knowledge about other agent's knowledge (i.e. without *first order theory of mind*) it is not possible to know which agents possess the information it is lacking. Even if it accidentally hears information, a non-first-order theory of mind agent cannot efficiently decide what to utter in response to maximize its reward. Higher order theory of mind is also often needed in SymmToM, as we will discuss further in §8.

Even though we only discussed a collaborative task for SymmToM, it can easily be extended for competitive tasks[1]. Moreover, all our models are also designed to work under competitive settings. SymmToM satisfies the desiderata we laid out in the previous section, as we will detail below:

Symmetric action space. As defined in Eq. 1, $\mathcal{A}_i = \mathcal{A}_j$ for all i, j. Only a subset may be available at a time since agents cannot step outside the grid, speak a piece they have not heard, or move if they would collide with another agent in the same cell, but they all share the same action space.

Imperfect information. Messages sent by agents outside of the hearing range will not be heard. For example, in Fig. 2a green sends a message but it is not heard by anyone, since it is outside of red's and blue's range. Hearing ranges are guaranteed not to cover the whole grid, since $2h+1 < w$.

Observation of others. Agents have perfect vision of the grid, even if they cannot hear what was said outside of their hearing range. Hence, an agent may see that two agents were in range of each other, and thus probably interacted, but not hear what was communicated. An example of this can be seen in Fig. 2a, where green observes blue and red interacting without hearing what was uttered. The uncertainty in the observation also differentiates SymmToM from prior work: to solve the task perfectly, an agent needs to assess the probability that other agents outside its hearing range shared a specific piece of information to avoid repetition. This estimation may be performed using the knowledge of what each agent knows (first order theory of mind), the perceived knowledge of each of the agents in the interaction (second order theory of mind), as well as higher order theory of mind.

Information-seeking behavior Rewards are explicitly given for hearing and sharing novel information, guaranteeing information-seeking is crucial in SymmToM. Recharge bases (Fig. 2b) ensure that the optimal solution is not for all agents to accumulate in the same spot and quickly share all the information available; and that the information tracking required is more complex than accumulating past events. Conceptually, with recharge bases we introduce an explicit and observable forgetting mechanism. As discussed in Section 3, this allows for perpetual information seeking and requires information staleness estimation.

5. Baseline Learning Algorithms and Bounds

To learn a strong baseline policy for SymmToM, we use MADDPG (Lowe et al., 2017), a well-known multi-agent actor-critic framework with centralized training and decentralized execution, to counter the non-stationarity nature of multi-agent settings. In MADDPG, each actor policy receives its observation space as input, and outputs the probability of taking each action. Notably, actors in MADDPG have no way of remembering past turns. This is a critical issue in SymmToM, as agents cannot remember which pieces they know, which ones they shared and to whom, and other witnessed interactions. To mitigate this, it is necessary to add a mechanism to carry over information from past turns, for example via incorporating a recurrent network as RMADDPG (Wang et al., 2020) does.

Perfect Information, Heuristic and Lower Bound Models Performance is difficult to interpret without simpler baselines. As a lower bound model we use the original MADDPG, that since it does not have recurrence embedded, should perform worse or equal to any of the modifications described above. We also include an oracle model (*MADDPG-Oracle*), that does not require theory of mind since it receives the current knowledge K for all agents in its observation space. The performance of MADDPG-Oracle may not always be achieved, as there could be unobserved communication with multiple situations happening with equal probability. Moreover, as the number of agents and size of the grid increases, current reinforcement learning models may not be able to find an optimal spatial exploration policy; they may also not be capable of inferring the optimal piece of information to communicate in larger settings. In these cases, MADDPG-Oracle may not perform optimally, so we also include a baseline with heuristic agents to compare performance.

Heuristic agents will always move to the center of the board and communicate round-robin all the information pieces they know until they have all the available knowledge. Then, they will move efficiently to their recharge base and come back to the center of the grid, where the process restarts. We must mention that this heuristic is not necessarily the perfect policy, but it will serve as a baseline to note settings where current multi-agent reinforcement learning models fail even with perfect information. Qualitatively, smaller

[1] There are many possible competitive extensions. For example, if we ended the trial when an agent steps successfully on their base for the b-th time ($b > 1$, to preserve the forgetting mechanism), giving that agent a positive reward and a negative one to all others, we would encourage competition.

settings have shown to approximately follow a policy like the heuristic just described.

6. Explicit Modeling of Symmetric Theory of Mind

In contrast to RMADDPG (Wang et al., 2020), we specifically design algorithms for our environment. This will ensure that we test the current limits of performance with known multi-agent deep reinforcement learning models. If even these models fail to solve the task, it will be a clear signal that there is more modeling research needed, and that SymmToM will be a useful benchmark to develop and test on. Intuitively, our model computes a matrix, $K \in \{0,1\}^{c \times n}$, that reflects the information pieces known by each agent from the perspective of the agent being modeled: K_{ij} reflects if the agent being modeled believes that agent j knows i. K is updated every turn and used as input of the following turn of the agent, obtaining the desired recurrent behavior. K is also concatenated to the usual observation space, to be processed by a two-layer ReLU MLP and obtain the probability distributions for speech and movement, as in the original MADDPG. There are several ways to approximate K. It is important to note that each agent can only partially observe communication, and thus it is impossible to perfectly compute K deterministically.

The current knowledge is comprised of first-hand information (the initial knowledge of every agent, F, publicly available) and second-hand information. Second-hand information may have been heard this turn (S, whose computation will be discussed below) or in previous turns (captured in the K received from the previous turn, noted $K^{(t-1)}$). Additionally, knowledge may be forgotten when an agent steps on a base having all the information pieces. To express this, we precompute a vector $B \in \{0,1\}^n$ that reflects whether each agent is currently on its base; and a vector $E \in \{0,1\}^n$ that determines if an agent is entitled to use their recharge base:

$$E_j = \mathbb{1}_{\sum_i K_{ij} = c} \text{ for all } j \in \{1, \ldots, n\}$$

We are then able to compute K as follows:

$$K_{ij}^{(t)} = (F_{ij} \text{ or } S_{ij} \text{ or } K_{ij}^{(t-1)}) \text{ and not } (B_j \text{ and } E_j) \quad (2)$$

F, $K^{(t-1)}$, and B are given as input, but we have not yet discussed the computation of the second-hand information S. S often cannot be deterministically computed, since our setting is partially observable. We will identify three behaviors and then compute S as the sum of the three:

$$S = S^{[0]} + S^{[1]} + S^{[2]}$$

For simplicity, we will assume that we are modeling agent k. $S^{[0]}$ will symbolize the implications of the information spoken by agent k: if agent k speaks a piece of information, they thus know that every agent in its hearing range must have heard it (first order theory of mind). $S^{[1]}$ will symbolize the implications of information heard by k: this includes updating k's known information (zeroth order theory of mind) and the information of every agent that is also in hearing range of the speaker heard by k. $S^{[2]}$ will symbolize the estimation of information pieces communicated between agents that are out of k's hearing range. Since we assume perfect vision, k will be able to see if two agents are in range of each other, but not hear what they communicate (if they do at all).

$S^{[0]}$ and $S^{[1]}$ can be deterministically computed. To do so, it is key to note that every actor knows the set of communicative actions $A \in \{0,1\}^{c \times n}$ performed by each agent last turn, given that those actions were performed in their hearing range. Moreover, each agent knows which agents are in its range, as they all have perfect vision. We precompute $H \in \{0,1\}^{n \times n}$ to denote if two given agents are in range.

Then, $S_{ij}^{[0]} = 1$ if and only if information piece i was said by k, and agents k and j are in hearing range of each other:

$$S_{ij}^{[0]} = A_{ik} \cdot H_{kj}$$

$S_{ij}^{[1]} = 1$ if and only if agent k (the actor we are modeling) heard some agent ℓ speaking information piece i, and agent j is also in range of agent ℓ. Note that agent k does not need to be in hearing range of agent j. More precisely,

$$S_{ij}^{[1]} = A_{i\ell} \cdot H_{k\ell} \cdot H_{\ell j}, \text{ for any agent } \ell$$

$S^{[2]}$ –the interactions between agents not in hearing range of the agent we are modeling– can be estimated in different ways. A conservative approach would be to not estimate interactions we do not witness ($S^{[2]} = 0$, which we will call *MADDPG-ConservativeEncounter (MADDPG-CE)*); and another would be to assume that every interaction we do not witness results in sharing a piece of information that will maximize the rewards in that immediate turn. We will call this last approach *MADDPG-GreedyEncounter (MADDPG-GE)*. MADDPG-GE assumes agents play optimally, but does not necessarily know all the known information and that could lead to a wrong prediction. This is particularly true during training, as agents may not behave optimally. The computation of $S^{[2]}$ for MADDPG-GE is as follows.

First, we predict the information piece U_ℓ that agent ℓ uttered. MADDPG-GE predicts U_ℓ will be the piece that the least number of agents in range know, as it will maximize immediate reward:

$$U_\ell = \arg\min_i \sum_j (K_{ij} \text{ and } H_{j\ell}) \in \{1, \ldots c\}$$

With this prediction, agent j will know information i if at least one agent in its range said it:

$$S_{ij}^{[2]} = 1 \text{ if } \exists\, \ell \neq k \text{ s.t. } U_\ell = j \text{ and } H_{j\ell} \text{ and } j \neq k \text{ else } 0$$

6.1. MADDPG-EstimatedEncounter

MADDPG-CE and MADDPG-GE are two paths to information sharing estimation, but neither estimate the probability of an agent knowing a specific piece of information. In *MADDPG-EstimatedEncounter (MADDPG-EE)*, known information of other agents is not binary, i.e. $K_{ij} \in [0, 1]$. This added flexibility can avoid making predictions of shared information based upon unreliable information.

MADDPG-EE estimates the probability that an agent j uttered each piece of information ($U_j \in \mathbb{R}^c$) by providing the current information of all agents in its range to an MLP:

$$U_j = \text{softmax}(f(K_{1j}, \dots, K_{cj}, \{K_{1\ell}, \dots, K_{c\ell} \text{ for all } \ell \text{ where } H_{j\ell}\})), \text{ with } f \text{ an MLP}$$

Then, the probability of having heard a specific piece of information will be the complement of not having heard it, which in turn means that none of the agents in range said it:

$$S_{ij}^{[2]} = 1 - \prod_{\ell, H_{j\ell}=1} 1 - U_\ell$$

Since MADDPG-EE requires functions to be differentiable, we use a differential approximation of Eq. 2. A pseudocode of MADDPG-EE's implementation can be found in Section A.4. MADDPG-EE solely focuses on first order theory of mind, and we leave to future work modeling with second order theory of mind. The structure of the model would be similar but with an order of magnitude more parameters.

7. Experiments

Next we compare the aforementioned algorithms. The observation space will be constituted of a processed version of the last turn in the episode, to keep the input size controlled. More precisely, the observation space is composed of: the position of all agents, all recharge bases, the current direction each agent is moving towards, what they communicated in the last turn, the presence of a wall in each of the immediate surroundings, and every agents' first-hand information. First-hand information is publicly available in our experiments to moderate the difficulty of the setup[2], but this constraint could also be removed. To lift this constraint, one approach would be to assume that $F_{ij} = 0$ for every unknown first-hand information, and learn K only based on heard interactions (modeled in $S[1]$).

We use reward as our main evaluation metric. This metric indirectly evaluates theory of mind capabilities, since information-seeking is at the core of SymmToM. We train through 60000 episodes, and with 9 random seeds to account for high variances. Our policies are parametrized by a two-layer ReLU MLP with 64 units per layer, as in the original MADDPG (Lowe et al., 2017). MADDPG-EE's function f is also a two-layer ReLU MLP with 64 units per layer.

We test two board sizes ($w \in \{6, 12\}$), two numbers of agents ($n \in \{3, 4\}$), and three quantities of information pieces ($c \in \{n, 2n, 3n\}$). Agents are placed randomly, and initial information is distributed randomly but equitably: each information piece is initially known by the same number of agents. Information exchange is simultaneous among agents. $h = 1$ for all our experiments: only agents' immediate neighbors will hear what they communicate.

Running experiments with the same number of turns for every setting would imply that agents can move less in combinations with larger values of w. Therefore, we set the length of each episode to $5w$, to make the length of each episode proportional to the grid size. Since the duration of the experiment is directly proportional to the length of the episodes, we settled on a small multiplier. $5w$ allows agents to move to each edge of the grid and back to the center. More design and experimental details can be found in §A.5.

7.1. Main Results

As we can observe in Table 1, there is a significant difference in performance between MADDPG-Oracle and MADDPG (MADDPG-Oracle is 127% better on average): this confirms that developing theory of mind and recurrence is vital to perform successfully in SymmToM. MADDPG-Oracle is often not an upper bound: when $c > n$, the heuristic performs better (92% on average, see details in §A.5). This shows that even with perfect information, it can be difficult to learn the optimal policy using MADDPG.

Moreover, models with recurrence perform significantly better than MADDPG (∼60% better), showing that remembering past information gives a notable advantage. As expected, recurrent models tailored to our problem resulted in better performance than a vanilla LSTM (RMADDPG). The performance of the best of the tailored models (MADDPG-{CE,GE,EE}) was 42% better on average than plain RMADDPG. LSTM was able to surpass the best of the tailored models only for $n=3, w=12, c=3n$.

Increasing c generally decreases global rewards for learned agents (on average, $c = 2n$ rewards are 74% of those for $c = n$, and $c = 3n$ rewards are 76.5% of $c = n$). This suggests that probabilistic decisions are harder to

[2] This simple setting is still partially observable, since the agents cannot hear interactions outside of their hearing range.

Table 1. Average reward per agent evaluated during 1000 episodes. 9 runs are averaged for each learned agent, using the best checkpoint to compensate for collapses in performance seen in Fig. 5 and 6. Values shown are individual rewards to normalize by the number of agents. Bold represents the best result of a learned imperfect-information model for each setting. Standard deviations appear in brackets.

agents (n)	3						4					
grid width (w)	6			12			6			12		
info pieces (c)	n	$2n$	$3n$	n	$2n$	$3n$	n	$2n$	$3n$	n	$2n$	$3n$
Heuristic (1000 trials)	$39_{[11]}$	$53_{[13]}$	$58_{[13]}$	$37_{[12]}$	$58_{[15]}$	$71_{[15]}$	$60_{[15]}$	$74_{[15]}$	$74_{[16]}$	$59_{[18]}$	$86_{[18]}$	$99_{[18]}$
MADDPG-Oracle	$42_{[7]}$	$48_{[6]}$	$41_{[6]}$	$56_{[18]}$	$40_{[12]}$	$41_{[17]}$	$70_{[4]}$	$47_{[6]}$	$32_{[5]}$	$55_{[10]}$	$33_{[6]}$	$33_{[11]}$
MADDPG	$40_{[2]}$	$18_{[2]}$	$15_{[1]}$	$39_{[16]}$	$33_{[16]}$	$11_{[1]}$	$34_{[6]}$	$22_{[3]}$	$13_{[0]}$	$17_{[4]}$	$15_{[2]}$	$14_{[1]}$
+RNN (RMADDPG)	$39_{[5]}$	$19_{[3]}$	$18_{[3]}$	$42_{[9]}$	$24_{[15]}$	$\mathbf{20}_{[7]}$	$30_{[4]}$	$20_{[2]}$	$16_{[1]}$	$32_{[7]}$	$17_{[3]}$	$16_{[1]}$
+Conservative (CE)	$36_{[4]}$	$26_{[3]}$	$33_{[4]}$	$49_{[12]}$	$55_{[34]}$	$14_{[1]}$	$\mathbf{40}_{[7]}$	$\mathbf{31}_{[3]}$	$25_{[3]}$	$30_{[12]}$	$25_{[8]}$	$16_{[1]}$
+Greedy (GE)	$37_{[7]}$	$\mathbf{26}_{[4]}$	$\mathbf{34}_{[4]}$	$\mathbf{52}_{[19]}$	$\mathbf{61}_{[26]}$	$14_{[2]}$	$34_{[11]}$	$30_{[4]}$	$\mathbf{26}_{[3]}$	$\mathbf{34}_{[8]}$	$\mathbf{28}_{[12]}$	$\mathbf{16}_{[2]}$
+Estimated (EE)	$\mathbf{41}_{[4]}$	$20_{[2]}$	$15_{[2]}$	$39_{[10]}$	$24_{[10]}$	$11_{[1]}$	$36_{[8]}$	$22_{[3]}$	$14_{[1]}$	$23_{[13]}$	$18_{[3]}$	$15_{[1]}$

learn, or impossible to successfully navigate when several events are equally likely. MADDPG-EE did not show improvements over the other agents, and in some cases performance decreased dramatically (e.g. $w=6$, $c=3n$). MADDPG-EE uses an MLP in its definition of $S^{[2]}$, which provides flexibility but complicated learning. We leave exploration of other probabilistic agents to future work, but the significant performance gap between learned models and the MADDPG-Oracle / heuristic shows there is ample space for improvement in this task, and hence proves SymmToM to be a simple yet unsolved benchmark.

Increasing n results in a 11% reduction of performance on average for learned models. Nonetheless, the heuristic improved its rewards by an average of 46%, given the larger opportunities for rewards when including an additional listener. Overall, this implies that increasing n also makes the setup significantly more difficult. Finally, increasing w did not have a conclusive result: for $n = 4$ it consistently decreased performance in 17%, but for $n = 3$ we saw an improvement of 18% and 61% for $c = n$ and $c = 2n$ respectively, and a decrease of 27% for $c = 3n$.

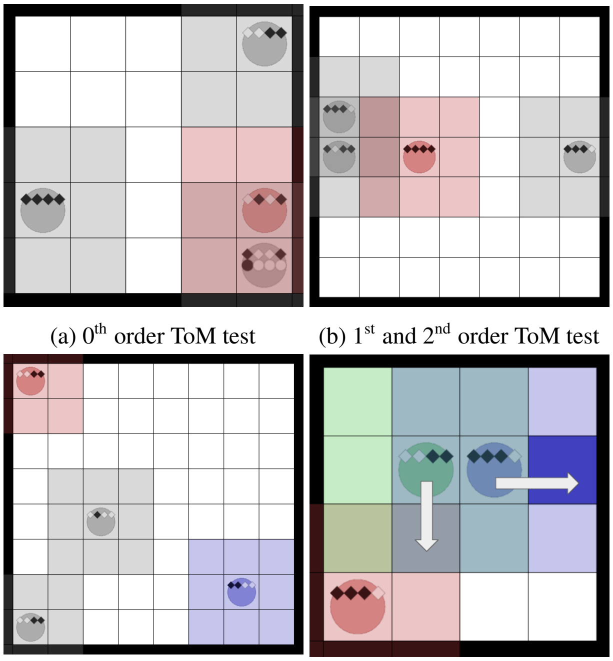
In sum, modifying c and n provides an easy way of making a setting more difficult without introducing additional rules.

8. Discussion

A classic example of a scenario specifically designed to test theory of mind is the Sally-Anne task (Wimmer & Perner, 1983). This false belief task, originally designed for children, aims to test if a passive observer can answer questions about the beliefs of another person, in situations where that belief may not match reality. If we were to use it for machine theory of mind, we could repeat the experiment and ask an agent to predict the position of an object varying the underlying conditions. This test is feasible because there is only one agent with freedom of action, which ensures that desired conditions are met every time. We can set up a similar setting in SymmToM if we allow for manual control of all agents but one, as shown in Fig. 3. Other tests besides the ones shown may be designed. In particular, in Fig. 3d we show an example of probabilistic theory of mind where two communicative events are equally likely, but one could modify this scenario to have different probabilities and test the expected value of the turns until red successfully shares an information piece. One could also design *retroactive deduction* tests: for example, in Fig. 3d if red communicates and receives no reward, it can deduce that green had received that information from blue. If there had been another agent (e.g. a yellow agent) in range of blue when it spoke to green, the red agent could also update its knowledge about yellow. Results and full discussion for the proposed tests are detailed in App. A.1. Models generally failed tests depicted in Fig. 3a and 3b, with significant variance between runs. As expected, $w = 12$ proved harder than the same test in a smaller grid. In $w = 6$ models often converged to a suboptimal but reasonable policy, whereas in $w=12$ efficient movement to a suboptimal goal was nontrivial. Notably, the second-order theory of mind test (Fig 3c) averaged $\sim 75\%$ success rate, which we hypothesize is due to having a mobile agent that the tested agent perceives as feedback.

Post-hoc analysis also has its challenges in multi-agent settings, even in the most direct cases. Thanks to our reward shaping, using recharge bases is always the optimal move when an agent has all the information available: an agent will have a reward of $(n-1)c$ for using the base, whereas it can only gain up to $n-1+c-1$ per turn if it decides not to use it. Even in this case, small delays in using the base may occur, for example if the agent can gather additional rewards on its path to the base. More generally, having multiple agents makes a specific behaviors attributable to any of the several events happening at once, or a combination of them.

Even though it may be difficult to establish causality when observing single episodes, we developed metrics


(a) 0th order ToM test (b) 1st and 2nd order ToM test

(c) 2nd order ToM test (d) Probabilistic ToM test

Figure 3. Example tests for 0th, 1st, 2nd order, and probabilistic Theory of Mind. We test red agents, immobilize gray agents, and control blue and green agents' movements. In Fig. 3a, red will go to the top right if it remembers to have heard the first piece, and to the left otherwise. In Fig. 3b, red will move to the right if and only if it assumes that the two agents on the left played optimally (red cannot hear them). In Fig. 3c, blue is controlled to ensure it will search the agent on the bottom left (its optimal play, in five moves). Red's optimal move is to meet blue, and hence must only move to the bottom left, even if the agent currently there will not provide any reward. In Fig. 3d, red will interact with green not knowing what blue previously shared with it. Red should be able to share the missing piece to green with an expected value of 1.5 turns.

that comparatively show which models are using specific features of the environment better than others. Reward can also be understood as a metric with a more indirect interpretation.

Post-hoc analyses of single episodes can also be blurred by emergent communication. Since agents were trained together, they may develop special meaning assignment to specific physical movements or messages. Even though qualitatively this does not seem to be the case for the models presented, tests should also account for future developments. This also implies that one should not over-interpret small differences in metrics.

We briefly describe the developed metrics below, full tables of results are available in Appendix A.2. All metrics are normalized by number of agents (i.e., they show the score for a single agent). This allows for better comparison between $n = 3$ and $n = 4$ settings.

Unsuccessful recharge base rate: Average times per episode an agent steps on its recharge base without having all the information available (i.e. wrong usage of the recharge base). Note that an agent may step on its base just because it is on the shortest path to another cell. Therefore, a perfect theory of mind agent will likely not have zero on this score; but generally, lower is better. See A.2 Table 4.

Wrong communication piece selection: Average times per episode an agent attempted to say information they currently do not possess. In these cases, no communication happens. Lower is better. See A.2 Table 5.

Useless communication piece selection: Average times per episode an agent communicated an information piece that everyone in its hearing range already knew, when having a piece of information that at least one agent in its range did not know. Lower is better. See A.2 Table 6.

Useless movement: Average times per episode an agent moves away from every agent that does not have the exact same information it has, given that the agent does not currently possess all the information available. This means that the agent is moving away from any possible valuable interaction. Lower is better. See A.2 Table 7.

A.2 contains full results tables. Briefly, we saw that MADDPG-CE and MADDPG-GE used recharge bases unsuccessfully at similar rates as Oracle, whereas RMADDPG performed 41% worse. Regarding information sharing, results suggest all models may be making wrong communicational decisions, but RMADDPG is more biased towards sharing redundant information when in-doubt, whereas MADDPG-CE and MADDPG-EE tend towards not communicating at all (the true effect of trying to share information one does not know).

9. Related Work

Theory-of-Mind has been studied for decades in cognitive science (Premack & Woodruff, 1978; Wellman, 1992; Astington & Baird, 2005). More recently, there has been work on developing agents that show that they can reason over the beliefs and goals of others (Rabinowitz et al., 2018; Rescorla, 2015). In many cases, models have been evaluated by being passive omniscient observers of a scene, either in a 2D (Rabinowitz et al., 2018) or natural language (Nematzadeh et al., 2018) world. Trained models are asked to predict the future given omniscient knowledge, but communication between observed agents is either non-existent or handcrafted. In cases where the modeled agent is active in the scene, movement or speech may be restricted for some agents but not others, leading to an asymmetric dynamic. For example, MADDPG (Lowe et al., 2017) has two tasks where oral communication is allowed, but there is only one speaker and the listener(s) have to react. Moreover, the speaker is immobile, in contrast to the listener(s). Other theory of mind speaker-listener tasks were evaluated only with two conversational agents, such as Zhu et al. (2021).

Work in reinforcement learning also often implicitly has some theory of mind modeling, especially in collaborative tasks. Even if the models can scale to multi-agent scenarios, models are often evaluated with only two agents for simplicity (Wang et al., 2020; Jain et al., 2019). Evaluating on only two agents often limits the opportunities for efficiently using higher-order theory of mind to solve a task: for example, agents never have to reason about i's assessment of j's modeling of k's mental state. One fully-symmetric multi-agent task often used in reinforcement learning is *cooperative navigation* (Lowe et al., 2017), a collaborative task that requires agents to cover landmarks without collision. Agents need to estimate where other agents will move, thus modeling their mental states. Traditionally, this task does not allow explicit communication between agents, resulting in impoverished theory of mind capabilities (Astington & Baird, 2005). Concurrently with our work, ToM2C (Wang et al., 2021) extended cooperative navigation and another related task (target coverage) to allow communication. Since all agents are symmetric, ToM2C may be understood as an example of multi-agent Symmetric Machine Theory of Mind. Nonetheless, some key differences arise: ToM2C only allows for targeted communication between pairs of sender and receiver, impeding deductions from bystanders of a specific sent message. Moreover, ToM2C only allows the sender to communicate the current estimation of the receiver goals, whereas –as detailed in Section 2– SymmToM allows agents to communicate pieces of information that they estimate are not known to people in their vicinity, but they never reveal this knowledge estimation directly. Information gathering plays a much more crucial role in SymmToM, and agents also need to be able to predict that other agents may forget information.

In the present work, we only focus on creating a task for analyzing complex reasoning over other agents' knowledge or lack thereof. Although theory of mind typically refers to reasoning over mental states, other aspects of theory of mind include understanding preferences, goals, intentions, and desires of others. Passive-observer benchmarks (Gandhi et al., 2021; Shu et al., 2021) have been proposed for evaluating the understanding of agent's goals and preferences, as well as understanding agent intentions (Ullman et al., 2009; Netanyahu* et al., 2021). Modeling is often analyzed by comparing to a human baseline, which is mainly possible due to the static nature of these datasets. Recently, Tejwani et al. (2021; 2022), developed a reinforcement learning framework called Social MDP, that incorporates social interactions into MDPs by reasoning recursively about the goals of other agents. As mentioned, reasoning about others goals' is another aspect of theory of mind, and it is complementary to our work. Social MDP's agents have full observation and only need to estimate other agent's goals based on their (fully observable) behavior. In SymmToM, in contrast to Social MDP, all agents have the same publicly-known information-sharing goal. What is unknown to SymmToM agents is the full state, particularly what other agents know at a given time: agents' reasoning aims to deduce interactions they did not witness. Although our reward fosters collaboration, agents in SymmToM do not directly gain from any increase or decrease in others' rewards as in Social MDP. Moreover, Social MDP's task does not have verbal communication, limiting communication to physical signaling.

10. Conclusions and Future Work

We defined a framework to analyze machine theory of mind (ToM) in a multi-agent symmetric setting, a richer and more realistic setup than theory of mind tasks currently used. Based on the four properties needed for symmetric theory of mind to arise, we provided a simplified setup on which to test the problem, and showed we can easily increase difficulty by growing the number of agents or communication pieces. Our goal in this work was not to solve symmetric theory of mind, but rather to give a starting point to explore more complex models in this area. Even with this minimal set of rules, SymmToM proves algorithmically difficult for current multi-agent deep reinforcement learning models, even when tailored to our specific task. We leave to future work to develop models that handle second-order theory of mind and beyond, and models that reevaluate past turns to make new deductions with information gained *a posteriori* (i.e., models that pass retroactive deduction tests). Another interesting direction is to replace the information pieces with constrained natural language: our communication sharing is binary, whereas in language there is flexibility to communicate different subsets of a knowledge base using a single sentence. It would also be interesting to test humans on this task. We hypothesize they may converge to a suboptimal policy –like the heuristic– due to our memory constraints and difficulty to methodically update and estimate knowledge. These should not be limiting factors for agents and thus we expect better performance in agent-agent interactions. We also think it would be valuable to test the differences in human performance if we alleviate memory limitations by allowing to take notes, and re-watching past turns.

Acknowledgements

This work was funded in part by NSF Award 2141751.

References

Astington, J. W. and Baird, J. A. *Why language matters for theory of mind*. Oxford University Press, 2005.

Baker, C., Saxe, R., and Tenenbaum, J. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.

Flobbe, L., Verbrugge, R., Hendriks, P., and Krämer, I. Children's application of theory of mind in reasoning and language. *Journal of Logic, Language and Information*, 17(4):417–442, 2008.

Gandhi, K., Stojnic, G., Lake, B. M., and Dillon, M. R. Baby intuitions benchmark (bib): Discerning the goals, preferences, and actions of others. *CoRR*, abs/2102.11938, 2021.

Hedden, T. and Zhang, J. What do you think i think you think?: Strategic reasoning in matrix games. *Cognition*, 85(1):1–36, 2002.

Jain, U., Weihs, L., Kolve, E., Rastegari, M., Lazebnik, S., Farhadi, A., Schwing, A. G., and Kembhavi, A. Two body problem: Collaborative visual task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6689–6699, 2019.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., and Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. pp. 6382–6393, 2017.

Nematzadeh, A., Burns, K., Grant, E., Gopnik, A., and Griffiths, T. Evaluating theory of mind in question answering. pp. 2392–2400, October-November 2018. doi: 10.18653/v1/D18-1261. URL https://aclanthology.org/D18-1261.

Netanyahu*, A., Shu*, T., Katz, B., Barbu, A., and Tenenbaum, J. B. Phase: Physically-grounded abstract social events for machine social perception. *35th AAAI Conference on Artificial Intelligence (AAAI)*, 2021. URL https://www.tshu.io/PHASE/.

Premack, D. and Woodruff, G. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4): 515–526, 1978.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., and Botvinick, M. Machine theory of mind. In *International conference on machine learning*, pp. 4218–4227. PMLR, 2018.

Rescorla, M. The computational theory of mind. 2015.

Shu, T., Bhandwaldar, A., Gan, C., Smith, K. A., Liu, S., Gutfreund, D., Spelke, E. S., Tenenbaum, J. B., and Ullman, T. Agent: A benchmark for core psychological reasoning. In *ICML*, pp. 9614–9625, 2021.

Tejwani, R., Kuo, Y.-L., Shu, T., Stankovits, B., Gutfreund, D., Tenenbaum, J. B., Katz, B., and Barbu, A. Incorporating rich social interactions into mdps. *arXiv preprint arXiv:2110.10298*, 2021.

Tejwani, R., Kuo, Y.-L., Shu, T., Katz, B., and Barbu, A. Social interactions as recursive mdps. In *Conference on Robot Learning*, pp. 949–958. PMLR, 2022.

Tomasello, M. *Constructing a language*. Harvard university press, 2009.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., and Tenenbaum, J. Help or hinder: Bayesian models of social goal inference. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

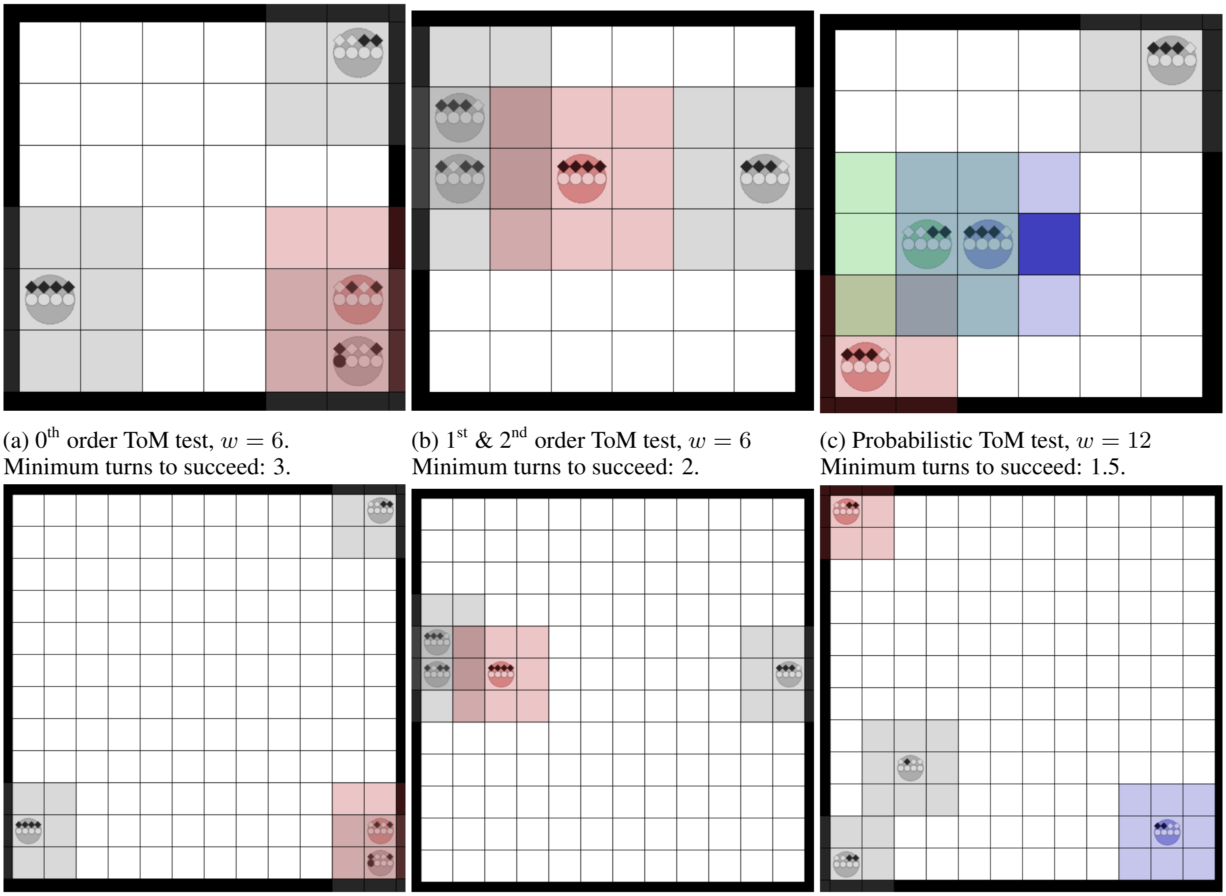
Wang, R. E., Everett, M., and How, J. P. R-maddpg for partially observable environments and limited communication. *arXiv preprint arXiv:2002.06684*, 2020.

Wang, Y., Zhong, F., Xu, J., and Wang, Y. Tom2c: Target-oriented multi-agent communication and cooperation with theory of mind. *arXiv preprint arXiv:2111.09189*, 2021.

Wellman, H. M. *The child's theory of mind*. The MIT Press, 1992.

Wimmer, H. and Perner, J. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983.

Zhu, H., Neubig, G., and Bisk, Y. Few-shot language coordination by modeling theory of mind. In *International Conference on Machine Learning*, pp. 12901–12911. PMLR, 2021.



(a) 0th order ToM test, $w = 6$.
Minimum turns to succeed: 3.

(b) 1st & 2nd order ToM test, $w = 6$
Minimum turns to succeed: 2.

(c) Probabilistic ToM test, $w = 12$
Minimum turns to succeed: 1.5.

(d) 0th order ToM test, $w = 12$
Minimum turns to succeed: 9.

(e) 1st & 2nd order ToM test, $w = 12$
Minimum turns to succeed: 8.

(f) 2nd order ToM test,
$w = 12$.

Figure 4. Depictions of rescaled tests from Figure 3, designed to match some of the parameter combinations already experimented on.

A. Appendix

A.1. Ad-Hoc Theory of Mind (ToM) tests

We test on the four examples shown in Figure 3, adapting the examples to fit one of the grid sizes we already experimented on. For the tests described in Figure 3a and Figure 3b, we test two different grid sizes: $w = 6$ and $w = 12$. For the tests described in Figure 3c and Figure 3d we only test $w = 12$ and $w = 6$ respectively. Image depictions of the exact test configurations can be seen in Figure 4.

We measure three metrics: *average success rate (SR)*, *average failure rate (FR)*, and *ratio of average turns to succeed vs. optimum (RATSO)*. Note that average success rate and average failure rate do not necessarily sum 1 since these two metrics only include trials where the agent reached any of the two proposed outcomes. If, for example, the agent never moved from the starting point, the trial would not be counted positively towards Avg. Success Rate nor Avg. Failure Rate. In addition, *ratio of average turns to succeed vs. optimum (RATSO)* is the ratio between the average turns it took to succeed in successful trials, and the optimum number of turns to succeed in a specific trial (lower is better, minimum possible value is 1.0).

For the tests in Figure 4a, 4b, 4d, and 4e, the trial ends when the red agent reaches the hearing range of one of the two possible target agents. The test depicted in Figure 4f is a pass/fail test: if red moves suboptimally at any point before meeting blue, the trial is declared as failed. This makes it a particularly difficult test to pass at random. Because of the nature of this second order theory of mind test, we only report the average success rate. Finally, for the probabilistic theory of mind test (Figure 4c) we want to measure how fast can red communicate all the information it has to green. The optimal number of turns is 1.5 (as described in Figure 3). Since this test can end either if all information has been shared to green, or if the

Theory of Mind test	Fig. 3a (0th order)								Fig. 3c (2nd order)
grid width (w)	6				12				12
	SR	FR	1-SR-FR	RATSO	SR	FR	1-SR-FR	RATSO	SR
MADDPG-Oracle	0%	100%	0%	–	0%	43%	57%	–	76%
MADDPG	0%	100%	**0%**	**1.11**	0%	22%	78%	–	74%
RMADDPG	0%	76%	24%	–	0%	32%	**68%**	–	74%
MADDPG-CE	0%	86%	14%	–	0%	**0%**	100%	–	69%
MADDPG-GE	0%	**57%**	43%	6.11	0%	18%	82%	–	**75%**
MADDPG-EE	**23%**	63%	14%	3.28	0%	11%	89%	3.00	69%

Theory of Mind test	Fig. 3b (1st and 2nd order)								Fig. 3d (probabilistic)	
grid width (w)	6				12				6	
	SR	FR	1-SR-FR	RATSO	SR	FR	1-SR-FR	RATSO	SR	RATSO
MADDPG-Oracle	0%	79%	21%	–	0%	37%	63%	–	71%	2.10
MADDPG	0%	83%	17%	**1.07**	0%	**24%**	76%	–	43%	3.17
RMADDPG	6%	83%	**11%**	1.60	0%	46%	54%	–	**70%**	2.71
MADDPG-CE	0%	83%	17%	8.33	0%	51%	**49%**	–	47%	2.25
MADDDPG-GE	**25%**	**58%**	17%	3.00	0%	51%	**49%**	–	29%	2.18
MADDPG-EE	8%	71%	21%	5.25	0%	47%	53%	–	45%	**1.89**

Table 2. Results for tests depicted in Fig. 4, evaluated during 1000 episodes for each of 9 different random seeds. SR means average success rate, FR means average failure rate, and RATSO is the ratio of average turns to succeed vs. the optimum turns to succeed. 1-SR-FR depicts the ratio of episodes where an agent did not reach any grid cell to terminated the test (either successfully or unsuccessfully) before the trial reached the maximum number of turns allowed ($5w$). A horizontal line means a metric could not be computed. Percentages are rounded to the nearest integer.

maximum number of turns has been reached, we will only report SR and RATSO. In other words, by design $FR = 0$ will always hold in this test.

Results are shown in Table 2. All tests show there is significant work to be done in improving agents' reasoning. Even in the Oracle setting, agents often fail the tests. For example, MADDPG-Oracle always fails the zeroth-order theory of mind test with $w = 6$ (depicted in Figure 4a). This shows that the trained model has learned a suboptimal but reasonable policy, since it moves towards an agent that will earn it a reward. In contrast, a high $1 - SR - FR$ in the test in Figure 4a shows that the agent never moved to the hearing range of any of the two possible "goal" agents – hence earning zero reward. Even though this would suggest MADDPG-GE performs the worst for this test, it is important to note that immobilizing agents introduces a new confounding variable (as all ad-hoc tests do, in line with what we argue in the main text). For example, if an agent sees that another one is not moving towards them, they might infer this agent is judging the interaction as useless and avoid interaction as well. In a test with a movement-controlled agent instead of all immobilized ones (second-order theory of mind test, Figure 4f) MADDPG-GE showed to perform the best among all learned agents, moving optimally in 75% of the trials. Success rate for the best of untrained models was only 33%, showing agents' learning significantly improves performance on this test.

As expected, tests in smaller grids showed to be easier than the same test performed on agents trained in a larger grid (See results for 0th and 1st+2nd order theory of mind in Table 2; no model was successful for $w = 12$). Since reward signals tend to be more sparse in larger grids, all models show larger values of $1 - SR - FR$. This may suggest that even efficiently moving towards a suboptimal goal may be a challenge, or that agents converged to a policy that plays a larger weight on making deductions based on other agents' movements. For $w = 6$, the best success rates were shown by MADDPG-*E models, although they still show ample room for improvement. As it can be seen in Table 3, even when average success rate is low, there sometimes exist seeds with exceptional performance. Concretely, there was one MADDPG-EE that was able to solve the 0th test for $w = 6$ to perfection.

Finally, in the probabilistic theory of mind test we see that no agent was able to consistently have perfect success in the task

	0^{th} order $(w = 6)$	0^{th} order $(w = 12)$	1^{st} order $(w = 6)$	1^{st} order $(w = 12)$	2^{nd} order test	probabilistic test
MADDPG-Oracle	0%	0%	0%	0%	82%	100%
MADDPG	0%	0%	1%	0%	81%	90%
RMADDPG	0%	0%	53%	0%	82%	99%
MADDPG-CE	0%	0%	0%	0%	80%	71%
MADDPG-GE	0%	0%	79%	0%	80%	60%
MADDPG-EE	100%	0%	37%	0%	83%	79%

Table 3. Average success rates (SR) of the best performing seed for each test. Showcases that although in most cases models are failing the tests, there are seeds performing better than most (even reaching perfect success rate, such as MADDPG-EE for the $w = 6$ test.)

(71% success rate was the highest achieved, by MADDPG-Oracle). This means that the red agent was not able to consistently communicate all the information to the green agent before the maximum number of turns was reached. Nonetheless, if we constrain ourselves to the successful trials, we see that MADDPG-EE was able to finish the test in less than twice the time of the theoretical optimum (1.89x, a similar rate as MADDPG-Oracle, whose RATSO was 2.1). This suggests that when agents succeed, they do so fairly quickly. For comparison, untrained agents have a RATSO between 5 to 7, showing the training procedure improved this metric significantly.

As we emphasized in the main text, many more tests can be proposed. Our released code base also allows for easily adding new tests to the suite.

A.2. Post-hoc analyses

RMADDPG had the worst scores for unsuccessful recharge base use rate and useless communication piece selection count (see Tables 4 and 6). RMADDPG scored 41% more than Oracle for unsuccessful base usage on average, and 64% more than Oracle on average for usage of a useless communication piece (in all our metrics, lower is better). The best tailored models (MADDPG-CE and MADDPG-GE) performed similarly to Oracle on average for these two metrics. In contrast, MADDPG-CE and MADDPG-GE performed significantly worse than Oracle for the wrong communication piece selection count (49% and 53% more than Oracle on average, see Table 5). This suggests that all models may be making wrong decisions, but RMADDPG is biased towards communicating redundant information whereas MADDPG-CE and MADDPG-EE tend towards not communicating at all (the true effect of trying to communicate something they are not allowed). Further analysis is needed to truly understand if these apparently wrong behaviors were done in turns where the agent had all the information available to make a better move, or if this is their default when they believe they have nothing of value to communicate. A priori RMADDPG bias seems more principled, but it still showed worse performance overall.

No learned model performed particularly better in the useless movement metric (average differences in performance were less than 15%, see Table 7), suggesting that they perform pointless movements in similar frequencies. It is important not to overinterpret small differences in these metrics. For example, a useless movement may be a signal of emergent communication. Furthermore, an agent may communicate something suboptimal for its immediate reward but this move may not affect its expected reward for the trial.

agents (n)	3						4					
grid width (w)	6			12			6			12		
info pieces (c)	n	$2n$	$3n$	n	$2n$	$3n$	n	$2n$	$3n$	n	$2n$	$3n$
MADDPG-Oracle	3.89	3.04	4.26	5.34	3.63	3.44	2.83	4.20	1.34	2.40	2.71	1.21
MADDPG	4.70	3.71	2.81	7.07	5.88	0.37	6.57	3.32	0.85	6.75	3.18	0.57
RMADDPG	6.09	5.22	3.04	9.46	6.39	4.92	5.57	4.61	0.87	7.17	2.42	**0.42**
MADDPG-CE	**3.94**	5.62	3.45	**4.60**	**3.70**	0.34	**3.23**	3.40	1.08	4.38	**1.45**	0.43
MADDPG-GE	4.01	5.75	3.86	5.34	4.55	**0.33**	4.19	3.11	1.03	**3.82**	1.94	0.60
MADDPG-EE	4.84	**3.65**	**2.54**	6.36	6.38	0.37	6.53	**2.91**	**0.72**	6.98	3.17	**0.42**

Table 4. Results for **unsuccessful recharge base usage rate**, normalized by agent. Bold lettering represents the best result of a learned imperfect-information model for each setting (lower is better).

agents (n)	3						4					
grid width (w)	6			12			6			12		
info pieces (c)	n	$2n$	$3n$	n	$2n$	$3n$	n	$2n$	$3n$	n	$2n$	$3n$
MADDPG-Oracle	3.06	3.22	3.36	10.76	9.84	5.76	2.68	4.86	1.88	7.95	5.85	3.59
MADDPG	4.77	0.39	0.30	11.29	4.25	1.12	5.82	0.24	1.33	11.59	**1.00**	**1.85**
RMADDPG	**3.13**	1.74	0.93	**8.86**	6.08	3.55	**3.37**	1.98	1.01	10.83	5.13	5.10
MADDPG-CE	6.46	3.70	3.81	11.00	5.60	3.51	7.14	4.03	3.84	14.34	4.35	11.37
MADDPG-GE	5.88	3.97	3.33	10.91	6.10	4.15	7.76	4.14	3.37	13.87	6.99	12.16
MADDPG-EE	4.40	**0.37**	**0.24**	9.53	**3.89**	**0.33**	5.38	**0.10**	**0.79**	**9.03**	1.39	2.85

Table 5. Results for **wrong communication piece selection count**, normalized by agent. Bold lettering represents the best result of a learned imperfect-information model for each setting (lower is better).

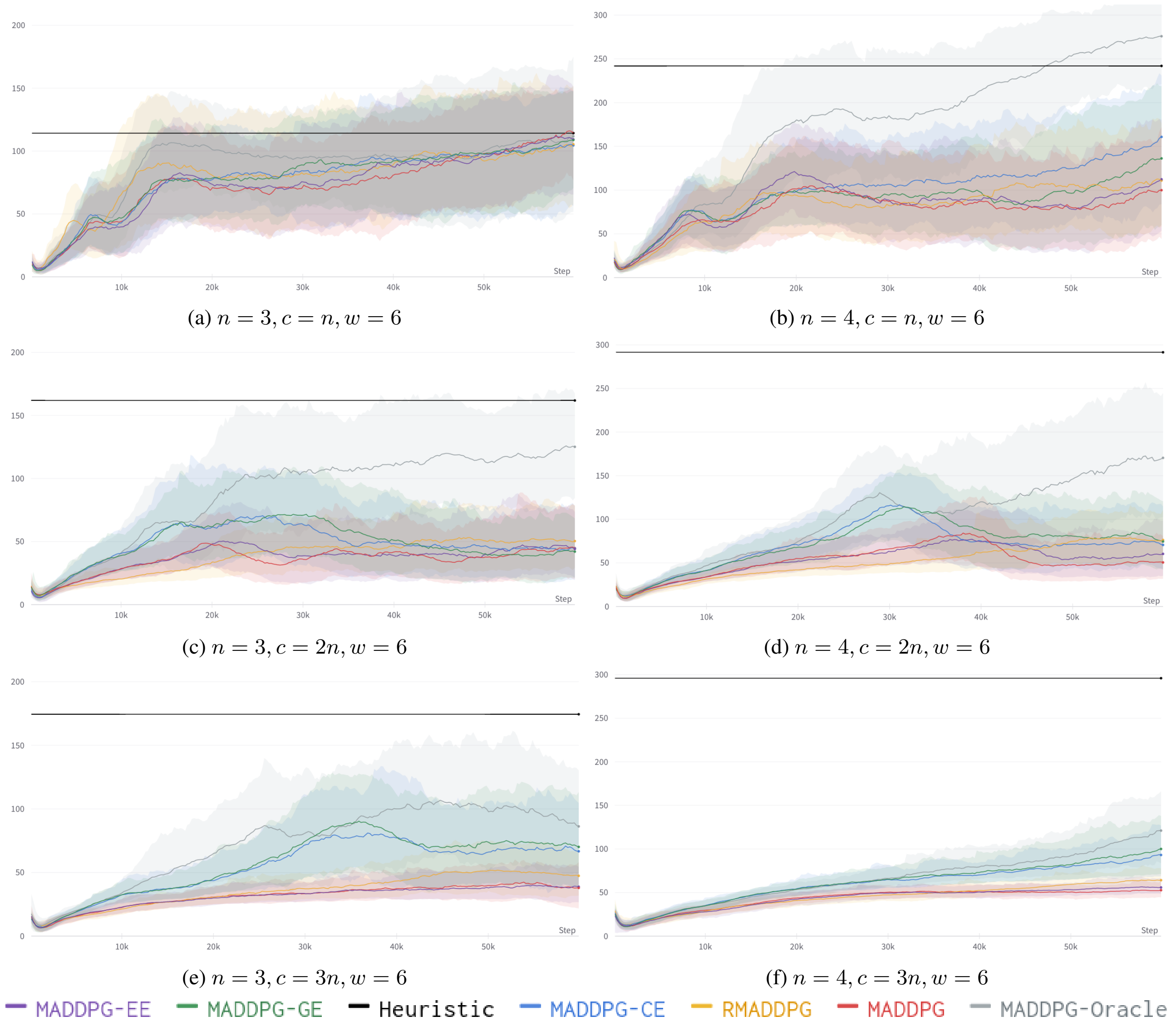
agents (n)	3						4					
grid width (w)	6			12			6			12		
info pieces (c)	n	$2n$	$3n$	n	$2n$	$3n$	n	$2n$	$3n$	n	$2n$	$3n$
MADDPG-Oracle	8.10	13.26	14.01	20.64	24.64	21.28	10.55	17.11	11.55	23.53	15.48	13.22
MADDPG	8.21	12.92	20.64	17.34	28.80	42.85	14.78	17.12	24.57	25.82	39.56	49.90
RMADDPG	8.17	16.83	18.42	18.23	34.85	38.28	**14.13**	19.66	24.51	28.75	37.00	50.26
MADDPG-CE	11.83	**9.28**	**9.69**	19.75	**12.91**	**8.19**	17.50	12.77	**15.05**	26.13	**11.32**	28.09
MADDPG-GE	11.55	9.52	10.19	19.72	13.45	10.11	17.28	**11.97**	15.75	24.58	16.61	**27.93**
MADDPG-EE	**7.79**	12.67	20.28	**15.92**	34.51	41.02	14.57	17.77	25.08	**24.09**	35.42	49.53

Table 6. Results for **useless communication piece selection count**, normalized by agent. Bold lettering represents the best result of a learned imperfect-information model for each setting (lower is better).

agents (n)	3						4					
grid width (w)	6			12			6			12		
info pieces (c)	n	$2n$	$3n$	n	$2n$	$3n$	n	$2n$	$3n$	n	$2n$	$3n$
MADDPG-Oracle	2.46	1.38	1.77	4.33	5.71	4.01	0.78	1.36	0.81	2.60	2.17	1.75
MADDPG	**2.03**	2.37	**1.42**	4.59	4.01	3.04	2.17	1.19	0.60	4.66	3.32	1.42
RMADDPG	2.93	2.59	1.50	**3.74**	4.27	3.74	**1.85**	1.44	**0.42**	**3.85**	3.04	**1.28**
MADDPG-CE	2.31	2.28	1.72	4.91	**2.17**	**2.93**	1.89	0.78	0.95	4.06	**0.92**	3.80
MADDPG-GE	2.66	2.23	1.73	4.26	2.21	3.50	1.96	**0.64**	0.87	4.31	1.70	3.43
MADDPG-EE	2.35	**2.19**	1.98	4.05	5.28	3.31	2.18	1.32	0.44	4.75	3.26	1.97

Table 7. Results for **useless movement count**, normalized by agent. Bold lettering represents the best result of a learned imperfect-information model for each setting (lower is better).

A.3. Training curves for all 9 random seeds combined



(a) $n = 3, c = n, w = 6$

(b) $n = 4, c = n, w = 6$

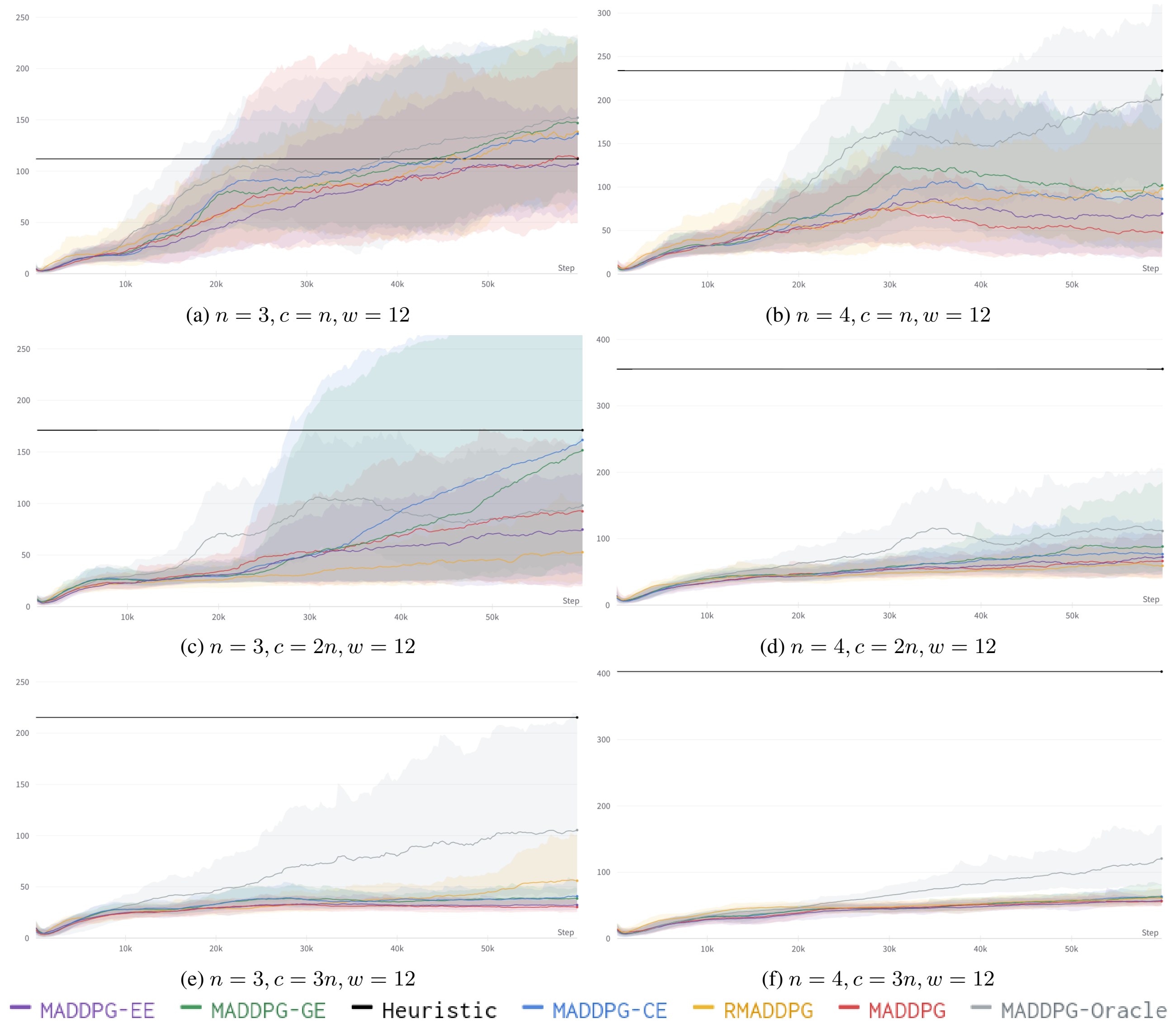
(c) $n = 3, c = 2n, w = 6$

(d) $n = 4, c = 2n, w = 6$

(e) $n = 3, c = 3n, w = 6$

(f) $n = 4, c = 3n, w = 6$

Figure 5. Average episode rewards throughout training for 60000 episodes for all combinations of $n \in \{3, 4\}$, $w = 6$, and $c \in \{n, 2n, 3n\}$.



(a) $n = 3, c = n, w = 12$

(b) $n = 4, c = n, w = 12$

(c) $n = 3, c = 2n, w = 12$

(d) $n = 4, c = 2n, w = 12$

(e) $n = 3, c = 3n, w = 12$

(f) $n = 4, c = 3n, w = 12$

Figure 6. Average episode rewards throughout training for 60000 episodes for all combinations of $n \in \{3, 4\}$, $w = 12$, and $c \in \{n, 2n, 3n\}$.

A.4. Pseudocode of MADDPG-EE

Algorithm 1 Actor implementation of MADDPG-EE, approximating K to make it differentiable.
Input: *observation*, $A \in \{0,1\}^{c\times n}$, $agent\_idx \in \{0, \dots n-1\}$, $F \in \{0,1\}^{c\times n}$, $K \in \{0,1\}^{c\times n}$, $B \in \{0,1\}^{n}$, $H \in \{0,1\}^{n\times n}$

```
Make agents not be in their own hearing range, to avoid talking to themselves from the previous turn. This would be problematic when using recharge bases.
    $H = H - \mathbb{1}_{n\times n}$

Compute $S^{[0]}$, all the heard information spoken by agent_idx:
    $S^{[0]} = \underset{n\ times}{copy}\, A[:, agent\_idx] \odot \underset{c\ times}{copy}\, H[agent\_idx, :]$
Compute $S^{[1]}$, all heard information by agent_idx, spoken by all agents:
    $S^{[1]} = (A \odot \underset{C\ times}{copy}\, H[agent\_idx, :]) \cdot H$
Compute $S^{[2]}$, an estimation of information pieces communicated between agents that were out of agent_idx's hearing range:
    $U_j = softmax(f_1(K_{1j}, \dots, K_{cj}, \{K_{1\ell}, \dots, K_{c\ell} \text{ for all } \ell \text{ where } H_{j\ell}\}))$, with $f_1$ an MLP
    $S^{[2]}_{ij} = 1 - \prod_{\ell, H_{j\ell}=1} 1 - U_\ell$ for all $i \in \{0, \dots, c-1\}$ and $j \in \{0, \dots, n-1\}$

$S = S^{[0]} + S^{[1]} + S^{[2]}$
$E_i = \mathbb{1}_{sum(K[:,i])=c} \in \{0,1\}^n$    for all $i \in \{0, \dots, c-1\}$
$K = step\big(F \cdot 100 + K + S - 2 \cdot \underset{c\ times}{copy}(B \odot E)\big)$
return $softmax(f_2([observation\ K])),\ K$    where $f_2$ is an MLP
```

A.5. Experiment Design Decisions and Considerations About Result Presentation

We tested with $n = 3$ and $n = 4$ and not larger numbers of agents, as the training time increases quadratically with n; also, the intrinsic difficulty of larger setup –even with perfect information– would possibly degrade performance to the point of making it impossible to compare models.

We select values of c as kn with $k \in \mathbb{N}$ so that every agent has k different information pieces at the start of each episode.

Experiments were run on a server with 256GB RAM, 2 18-core Intel E5-2699 processors @ 2.3GHz. All runs for $n = 3, w = 6$ took approximately 600 hours of compute; all runs for $n = 3, w = 12$ took approximately 1150 hours of compute; runs for $n = 4, w = 6$ took approximately 950 hours of compute; runs for $n = 4, w = 12$ took approximately 1800 hours of compute. We parallelized computation across the 18 cores. This made unfeasible to run experiments with $n > 4$, since all experiments for $n = 3$ and $n = 4$ took jointly 4500 hours of compute.

We decided to evaluate running additional episodes over the best checkpoints of each model because there was high variance for some runs, and drops in performance after achieving the highest rewards. Those results are the base of the discussion and can be seen in Table 1. Still, we share the training curves so that the reader can observe these behaviors in Fig. 5 and Fig. 6.

We used the same hyperparameters as the ones used in MADDPG, except with a reduced learning rate and tau ($lr = 0.001$ and $\tau = 0.005$). We used the same parameters for all our experiments.

We computed "A is 127% better than B on average" –and all other similar statements– as an average of relative increases. More precisely, if A_1,...,A_{12} and B_1,...,B_{12} are the average scores for each of the 12 settings for A and B respectively (see Table 1's columns), 127% better means $\text{mean}_i \frac{A_i}{B_i} = 2.27$.