Into the Minds of the Chefs

Using Theory of Mind for Robust Collaboration with Humans in Overcooked

anonymous [1]

Abstract

We propose ToM-FCP, a method leveraging Theory of Mind and Fictitious Co-Play to improve robustness in two agent collaboration in Overcooked. Theory of Mind acts as an inductive bias that allows our agents to reason about each other, thus helping them to be more robust when playing with novel partners. Our proposed method does not need human data and instead relies on adding training partner diversity by employing Fictitious Co-Play instead.

1. Introduction

Generally, when two agents collaborate they need to infer each others preferences, goals and future actions for effective coordination. Humans naturally infer these based on observation alone and thus can effectively coordinate with each other. This ability is referred to as Theory of Mind (ToM), a concept originally coined in the behavioral psychology literature (Heider & Simmel, 1944). This ability is so innate to the human experience that we engage in this behavior subconsciously and posses it even in early stages of infancy (Gergely et al., 1995; Gergely & Csibra, 1997; Repacholi & Gopnik, 1997; Carpenter et al., 2005). Still, we have a prior about the behavior of our cooperation partners and when their behavior differs too much from our expectations effective coordination protocols are no longer established and thus coordination becomes increasingly infeasible. With the rise of artificial intelligence (AI) new artificial cooperation partners will become increasingly common in our day-to-day life. Thus, enabling effective coordination becomes crucial.

We are interested in building artificial agents that are capable of cooperating with a diverse set of partners. Building such agents has proven to be a major problem for AI research (Klien et al., 2004; Mutlu et al., 2013; Bard et al., 2019; Carroll et al., 2019; Dafoe et al., 2020; Knott et al., 2021). While modern learning techniques such as self-play (Silver et al., 2017b; Hernandez et al., 2019) produce artificial agents that are capable of coordinating with *themselves*, their performance is often brittle when paired with other partners, both differently trained artificial as well as human ones.. Their coordination protocols never generalize and as a result they stubbornly execute their policy disregarding the behavior of the novel partner. This is referred a classical case of overfitting, due to the co-evolution of agents (Lanctot et al., 2017).

The co-evolution problem is particularly relevant when AI systems are applied in real life scenarios, where it highlights a second issue: Human safety in human-machine collaboration. The key insight here is that every human partner acts differently and we would want all interactions to be safe. When coordination protocols are not established or break down, cooperation fails and accidents are to be expected. In virtual and toy settings this might be of little concern but human-machine collaboration will increasingly occur in the physical world where physical harms are possible, e.g. a worker interacting with a robot arm, a pedestrian crossing the street in front of a self-driving car, an elderly person being assisted by a healthcare robot etc. In these scenarios coordination failure might induce significant bodily harm. This motivates making these systems *robust* and *safe*. With rule based or other forms of simple artificial agents, it might be possible to manually safeguard these against failures but for neural models such efforts might be infeasible. This is due to the fact that deep neural networks are little understood and interpretability work is still in its infancy.

Previous solutions to the co-evolution problem often focus on learning a human model (Sadigh et al., 2016; Carroll et al., 2019; Charakorn et al., 2020; Strouse et al., 2021; Knott et al., 2021; Yang et al., 2022) or constructing one to learn to coordinate with (Nikolaidis & Shah, 2013; Javdani et al., 2018; Choudhury et al., 2020) since getting data from actual humans at scale is highly time consuming and expensive. Thus, they focus on adding inductive biases (Hu & Foerster, 2019; Puig et al., 2020; Hu et al., 2020; Wang et al., 2020; Strouse et al., 2021; Yang et al., 2022) or

[1]Institute for Visualization and Interactive Systems, Department of Human Computer Interaction and Cognitive Systems, University of Stuttgart, Germany. Correspondence to: anonymous, matr. XXXXXX <anonymous@stud.uni-stuttgart.de >.

try to leverage existing human data, for instance by using behaviour cloning (Bain & Sammut, 1999; Torabi et al., 2018) or imitation learning (Abbeel & Ng, 2004; Ho & Ermon, 2016). This comes with the disadvantage of relying on a single human model for training. Ideally, we would like to improve robustness by diversifying the training partners and by improving their quality while needing no human data.

In this work, we propose to combine some of the recently suggested techniques such as Fictitious Co-Play (Strouse et al., 2021) and Theory of Mind (ToM) as an inductive bias without the need for human data. Our aim is to obtain agents able to generalize better to novel collaborative partners in ad-hoc, zero-shot collaboration settings.

2. Related Work

Human-AI Collaboration With the recent success of reinforcement learning in competitive environments (Silver et al., 2017b;a) attention has been also placed on settings in which agents need to *cooperate*. Especially, those in which agents need to do so with novel agents – known as the *ad-hoc coordination problem* (Stone et al., 2010; Stefano & Ramamoorthy, 2013) – since this has many practical applications. In this work we are interested in Human-AI coordination (Stefano & Ramamoorthy, 2013; Carroll et al., 2019), in which collaborative agents need to respond to human behaviour, which is often sub-optimal.

To adapt to playing with a sub-optimal partner Hu et al. (2020) proposed *Other Play* as a method for zero-shot collaboration. Their approach makes the assumption that an agent should be invariant to symmetries in the game which does not guarantee optimal behaviour for all cooperation partner (compare (Knott et al., 2021)). There have been many attempts to model human behaviour to (i) facilitate the training of a so called *ego-agent* against a more realistic human model (Nikolaidis & Shah, 2013; Javdani et al., 2018; Sadigh et al., 2016) or (ii) train a ego-agent with a diverse range of partners of different skills (Jaderberg et al., 2017; Charakorn et al., 2020; Zhao et al., 2021; Strouse et al., 2021).

To foster research in Human-AI collaboration Carroll et al. (2019) proposed Overcooked-AI (see Figure 1) as a challenging collaboration environment for multi-agent reinforcement learning which has become a recent focus in works on Human-AI collaboration (Knott et al., 2021; Charakorn et al., 2020; Nalepka et al., 2021; Fontaine et al., 2021; Zhao et al., 2021; Sarkar et al., 2022; Ribeiro et al., 2022). Overcooked tasks two agents to collaborate on cooking and delivering a soup together for which they get a joined reward. Notably, Overcooked consists of multiple layouts that are designed to facilitate certain types of challenges (see Figure 2 for an overview). The key insight of this work is that agents need to adjust to human play, which is often sub-optimal.

Improving Human-AI Collaboration in Overcooked With the recent emergence of Overcooked-AI (Carroll et al., 2019) as a benchmark several works used it for researching Human-AI collaboration (Charakorn et al., 2020; Nalepka et al., 2021; Fontaine et al., 2021; Zhao et al., 2021; Sarkar et al., 2022; Knott et al., 2021; Strouse et al., 2021; Ribeiro et al., 2022) Most related to use are the works of Knott et al. (2021) and Strouse et al. (2021). Strouse et al. (2021) introduced Fictitious Co-Play (FCP) as a training paradigm for training with greater partner diversity. The core idea in FCP is that if an agent is supposed to cooperate well with novel agents of different capabilities, it should be trained with a diverse set of partners of different skill. They achieved this by training an agent partner using a population of self-play agents and their past checkpoints taken during training. Knott et al. (2021) highlighted the importance of correct *evaluation* for making agents robust in human AI cooperation. Key to their paper is the question of how to evaluate agent performance when collaborating with humans, who are diverse and often unpredictable. As a result, they propose to apply the software engineering practice of unit-testing to Overcooked-AI, designing challenging test cases to evaluate the performance of the agent. In their unit-testing setup they tested three robustness techniques, including Theory of Mind. ToM is implemented using a rule-based approach that includes hand-picked values and heuristics, limiting its expressiveness and making it non-scalable to different problems, environments and partners. In contrast, we enrich agents with Theory of Mind by means of neural networks.

While adding diversity in the partner population is a known solution for the co-evolution problem, Charakorn et al. (2020) highlighted the difficulty in finding the correct strategy for adding diversity. They showed that partner sampling and population-based training are unreliable in adding diversity in all environments and propose the use of several pre-trained agents. Nalepka et al. (2021) further studied the human-aware agent trained by Carroll et al. (2019) in a modified Overcoocked environment to determine whether the increased performance was due to the increased flexibility of a human-aware versus self-play agent. In a human subject study they determined that humans indeed perform better with the human-aware agent but also that the subjects were more flexible in their interactions with the agent.

Machine Theory of Mind Theory of Mind (ToM) is often described as the ability of an agent to infer the goals, preferences and underlying mental states of other agents. This concept first emerged in the the cognitive psychology liter-

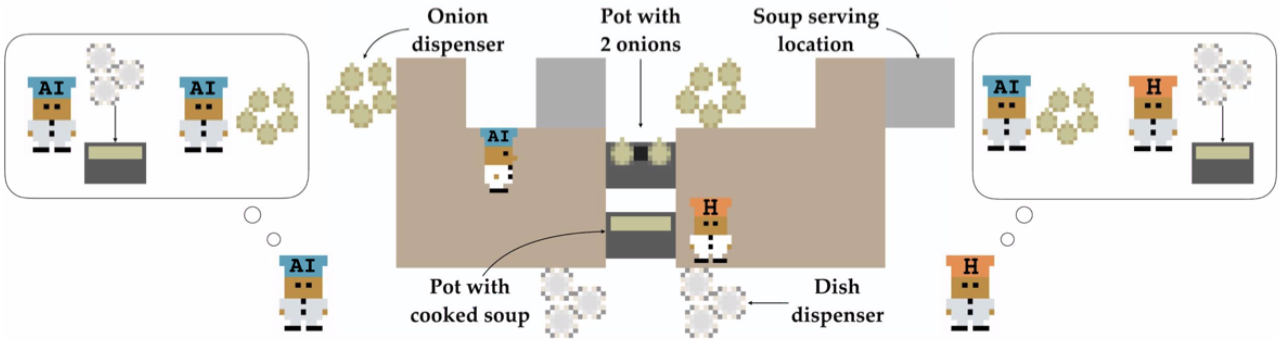

Figure 1. The Overcoocked environment (Carroll et al., 2019).

ature (Heider & Simmel, 1944). Rabinowitz et al. (2018) extended the concept to machines, coining the term Machine Theory of Mind. In their formulation, an observer is tasked with predicting the actions of agents in an environment where the assumption is that this requires the correct modeling of the agent's mental states. A number of recent works took inspiration from Rabinowitz et al. (2018) to include ToM in neural networks (Yuan et al., 2021; Sclar et al., 2022; Nguyen et al., 2022; Fuchs et al., 2021; Yu et al., 2022; Zhou et al., 2022; Nguyen et al., 2023; Bara et al., 2021; Liu et al., 2023). Especially in zero-shot cooperation setting where agents need to cooperate with agents never seen during training, ToM can act as an inductive bias towards building more realistic human models (Knott et al., 2021). Yuan et al. (2021) embedded ToM in an offline reinforcement learning algorithm by only updating a single agent at a time and having each agent hold a believe over the mental states of all other agents. In this work we take inspiration from Yuan et al. (2021) for embedding ToM in all agents for enhanced cooperation, especially in ad-hoc cooperation with humans.

AI Robustness & Safety With the ever increasing proliferation of AI systems in real-world applications, the social impacts of these systems become more and more relevant. This makes it necessary to give serious consideration towards challenges and risks, especially risks related to accidents (Bostrom, 2014; Amodei et al., 2016). Following Amodei et al. (2016) these accidents are defined as unwanted and unintended behaviour that emerges from an machine learning system due to a wrong or incomplete objective function, a mistaken learning process or other implementation errors. The impact of AI systems will likely be transformative and therefore making sure that these systems are robust with regard to accidents is a pressing issue. As discussed above these accident risks are especially relevant in real-world cooperation settings as here bodily harms are easily imaginable. Avoiding these is difficult since this not any ordinary multi-agent but a human-agent system that posses a unique challenge since humans are no idealized rational agents (Tversky & Kahneman, 1974). One avenue towards safe AI systems is to make sure that their objectives match our objectives, as discussed under the heading of value alignment (Fisac et al., 2020). Some alignment failures as for instance goal misgeneralization (Langosco et al., 2022) have already been observed in research settings.

We believe that ToM at is core is about reasoning about the mental states of others, thus taking into account their preferences, goals and objectives. An agent that has this ability therefore might prove more value aligned and thus us more robust as a collaboration partner for humans. Our work explores this hypothesis as we propose ToM as a robustness technique for human-agent interaction.

3. Key Novelty and Contributions

The key novelty in this work is to combine ToM modelling and FCP in an offline RL setting to train agents capable of collaborating with humans in the Overcooked environment. We view ToM and FCP as robustness techniques, which can produce agents that are flexible and adaptive to human partners with different skill levels. Importantly, our method does not require human data, which is often expensive to collect. In summary, our planned contribution is four-fold:

- A new approach combining ToM and FCP to produce agents capable of robustly collaborating with humans without human data.
- A series of ablation studies to study how ToM and FCP individually impact our agent's performance.
- A comparison of different ways to represent the mental state of an agent.
- A comprehensive set of metrics surveyed from the literature to evaluate an agent in collaborative settings,

in which average reward is not informative enough.

4. Approach

4.1. Overcooked

Carroll et al. (2019) specifically designed Overcoocked to be a challenging environment for two-agent collaboration which makes it ideal for testing our proposed approach. Within this environment the agents are tasked with cooking a soup (see Figure 1) by making use of onion dispensers, cooking stations, plate dispensers and soup serving locations. Agents can pick actions from the action space $\{up, down, left, right, noop, interact\}$ where $iteract$ has different effects depending on the object the agent is facing. The environment presents multiple layouts (see Figure 2) which have been designed to facilitate either low-level coordination or high-level strategy failures. A soup can be cooked by placing three onions into the cooking stations and waiting for 20 time steps, then arriving with a plate to pickup the soup and deliver it to the serving location. Formally, Overcooked is defined as an multi-agent Markov decision process (Boutilier, 1996) represented by the tuple $(S, \mathcal{N}, \{A_{i\in\mathcal{N}}\}, T, R)$. S is a set of environment states, $\mathcal{N}$ is the set of N agents, A_i is the action set available to the agent i, $T : S \times A_1 \times \cdots \times A_N \times S \to [0, 1]$ is the state transition function and $R : S \to \mathbb{R}$ the reward function.

4.2. Modelling Theory of Mind in Overcooked

We believe that the ToM model proposed by Knott et al. (2021) does not fully capture the potential that Theory of Mind has for robustness in collaborative reinforcement learning, for the following reasons:

1. It only allows the human model to have ToM capabilities. Ideally, both partners should be able to reason about the mental state of the other for better cooperation.

2. Their ToM capabilities are not expressive enough since they are based on simple rules that can not transfer to new tasks and environments. We would want these capabilities to generalize and thus aim for them to be learned.

3. Their ToM formulation is not scalable. Many problem settings and environments are much more complex and similar rules can not be easily found for them.

In contrast we propose adding ToM to the agents as an auxiliary training task as described by Yuan et al. (2021). In their formulation ToM emerges by having the agent reason about all other agents mental states by observing their actions alone. Additionally, we propose to combine this idea with FCP to add training partner diversity and optionally fine-tune our method on human data.

Following Yuan et al. (2021), we consider multi-agent reinforcement learning environments with $N = 2$ agents. At every time step t every agent i picks an action $a_i^t \in A_i$ given their own policy $\pi_i(a|\tau_i^t)$ where τ_i^t is the history of actions and observations of i, i.e. $\{o_i^0, a_i^0, \ldots, o_i^t, a_i^t\}$. Here, an observation o_i^t is a tuple of consisting of the physical state s given the action a_i^t, $O(s, a_i^t)$, and the mental state of the agent ω_i resulting in $o_i^t = (O(s, a_i^t), \omega_i)$ where $O(\cdot)$ is a function that encodes the state of the world. The complete game state $\tilde{S} = S \times \Omega_0 \times \cdots \times \Omega_N$ is given by the state space S and the agent mental states Ω_i. An agent's mental state Ω_i can be interpreted as representing its goal or intentions. Given the set of all actions $\mathbf{a}^t = \{a_0^t, \ldots, a_N^t\}$ at time step t the global state transition function is defined using the physical state transition function T and the agents' mental states ω_i.

Since there is no direct communication between agents they only act based on their observation and mutual inference. Thus, agent i acts according to its value function $Q_i : \tilde{S} \times A_i \to R$. An agent has to maintain a belief over the private states of all other agents, i.e. $b_i^t(\omega_{-i})$ where all other agents are denoted by the subscript $-i$. This lets us formalize the policy of an agent as

$$\pi_i(a|s^t, \omega_i^t, \tau_i^t) = \frac{\exp(\beta \sum_{\omega_{-i}\in\Omega_{-i}} b_i^t(\omega_{-i}|\tau_i^t) Q_i(a, s^t, \omega_i, \omega_{-i}))}{\sum_{a'\in A_i} \exp(\beta \sum_{\omega_{-i}\in\Omega_{-i}} b_i^t(\omega_{-i}|\tau_i^t) Q_i(a', s^t, \omega_i, \omega_{-i}))},$$

where the belief b_i^t is updated by traversing all possible mental states $\omega_{_i}$ and estimates how likely they are given its own observations, making use of an estimation of the the others policy $\hat{\pi}_{-i}$ by using a belief update function $f_{-i} : \Delta(\Omega_i) \times A_i \times S \to \Delta(\Omega_i)$ with $\Delta(\Omega_i)$ being a distribution over the private states Ω_i.

4.3. Method

In this work we aim to improve Human-AI collaboration by improving the quality and the diversity of training partners through ToM and FCP. We show an overview in Figure 3 and formalize our approach in Algorithms 3, 2 and 1.

Our first goal is to allow agents to explicitly reason about each other's mental state adapting the algorithm by Yuan et al. (2021) to the Overcooked environment. In Overcooked the goals of agents change during a game run (i.e. after picking up an onion you might then want to put it into the pot etc.). Thus, the assumption of Yuan et al. (2021) that agent's mental states are fixed for each game run is unrealistic. We will adapt and expand their algorithm to allow for non-fixed mental states (see Algorithm 2 and 1). This raises the following key questions: (i) What are the mental states of artificial

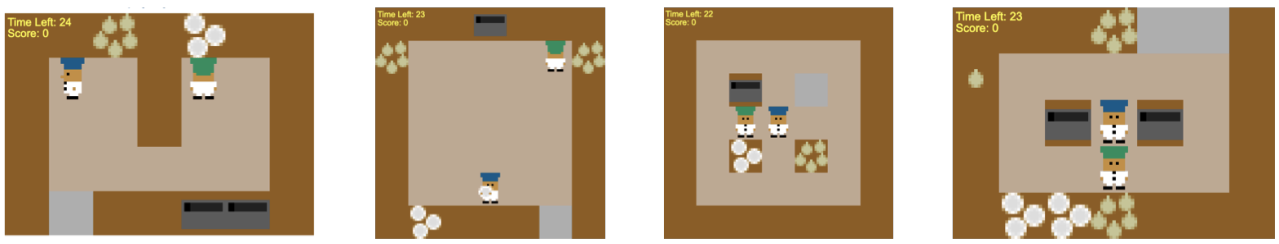

Figure 2. Overcooked-AI (Carroll et al., 2019) environment layouts; Taken from (Knott et al., 2021).

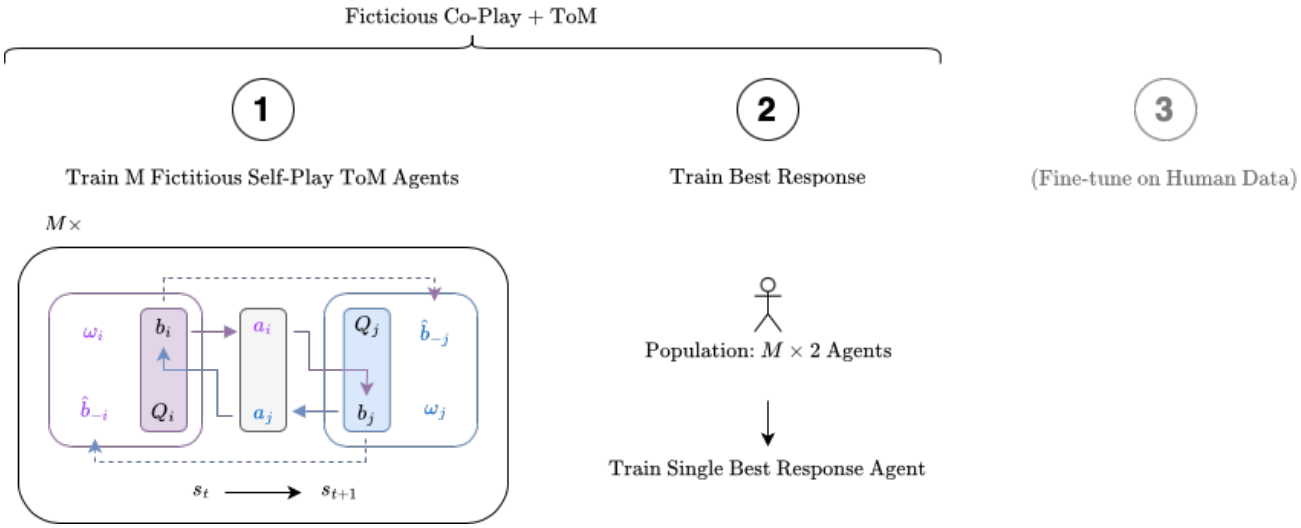

Figure 3. Overview of our approach. We first train M Self-Play agents as done by FCP where our agents have ToM capabilities (Yuan et al., 2021). We thereby obtain a population of $M \times 2$ agents that we then use to train a best response to. If time remains we propose to fine tune the resulting model on Human-Human data. Note that step 1 and 2 do not rely on any human gameplay data.

agents in Overcooked at any given time? (ii) How can we represent them such that other agents can predict them? To this end, first note that in offline RL it is easy to expand the dataset $\mathcal{D} = \{\tau^1, \dots, \tau^M\} = \{(s_t^k, a_t^k, r_t^k)_{t=0}^T\}_{k=1}^M$ of M trajectories adding belief annotations before using it for learning: $\mathcal{D} = \{(\omega_{1,t}^k, \dots, \omega_{N,t}^k, s_t^k, a_t^k, r_t^k)_{t=0}^T\}_{k=1}^M$. Since we want to be able to augment $\mathcal{D}$ on the fly and since we want this augmentation step to generalize to all environments it would be convenient to compute $\{\omega_{1:N}\}_{0:T}$ from $\mathcal{D}$ *itself*. In our work we explore three different ways of annotating our data with mental states.

1. Set $\{\omega_i\}_t$ to the next action an agent is about to perform, represented as a one-hot encoded vector. Here the mental state of an agent corresponds to the low-level action it is about to take.

2. Interpret and set $\{\omega_i\}_t$ to correspond to next strategic goal an agent has. So, for every timestep t in a trajectory τ find the next interaction the agent performs at step $t + n$ and set $\{\omega_i\}_t$ to the one-hot encoded interaction label. As an example, consider the sequence of actions {*right, right, noop, up, interact(pick_onions)*}. In this example then $\omega = \text{OneHot}(interact(pick\_onions))$.

3. Pick $\{\omega_i\}_t$ as a neural representation of agent i's policy network. Specifically, take a d-dimensional neural representation from the Q-network given the state s for all agents i first and use this as their mental state.

The second key advantage of our proposed method is that it needs no human data, which is in general expensive to obtain and time consuming. We achieve this by training our model with a population of diverse agents via FCP (Strouse et al., 2021) (see Algorithm 3). Specifically, we train our model in co-play with a partner sampled from a list of N partners, which are trained using self-play and initialized

Algorithm 1 Adaptive ToM Collaboration Emergence: Fixed Agent Update (short: ToMCollaborationFixedAgent). Adapted from Yuan et al. (2021).

Require: $\theta_i, \hat{\theta}_i, \eta_Q, \eta_\pi, \eta_f, \eta_l, i \in \{1, \dots, N\}, \mathcal{D}$
  repeat
    Agents sample actions according to their policy
    Agents update their beliefs
    Agents update their estimation of partners' beliefs
    Agents predict their partners skill levels $\hat{l}$
  until game ends
  Compute $\{\omega_{1:N}\}_{0:T}$ for new trajectory
  Update $\mathcal{D}$ with new trajectory
  Sample M trajectories $\{$
    $(\{\omega_{1:N}\}_{0:T}, l_{1:N}, \hat{l}_{1:N}, s_{0:T}, \mathbf{a}_{0:T}, r_{0:T})\}_{k=1}^{M}$
  $y_i^{t,(k)} = r_i^{t,(k)} + \gamma \max_{a \in A_i} Q_{\theta_i}($
    $a, s^{t+1,(k)}, w^{(k)}, \hat{b}_{-i}^{t,(k)})$
  $L^Q = \sum_{t,k} ||Q_{\theta_i}(a_i^{t,(k)}, s^{t,(k)}, w^{(k)}, \hat{b}_{-i}^{t,(k)}) - y_i^{t,(k)}||^2$
  $L^\pi = \sum_{t,k} H(\hat{\pi}_{-i}(\mathbf{a}_{-i}^t | s^{t,(k)}, w_{-i}^{(k)}), \mathbf{a}_{-i}^{t,(k)})$
  $L^f = KL(\bar{b}_{-i}^{t,(k)} || f_{-i}(\hat{b}_{-i}^{t-1,(k)}, a_i^{t,(k)}, s^{t,(k)}))$
  $L^l = \text{crossEntropy}(l_{1:N}, \hat{l}_{1:N})$
  $\theta_i \leftarrow \theta_i - \nabla_{\theta_i}(\eta_Q L^Q + \eta_\pi L\pi + \eta_f L^f + \eta_l L^l)$
  Periodically update $\hat{\theta}_i \leftarrow \theta_i$ for Q-learning

Algorithm 2 Adaptive ToM Collaboration Emergence (short: ToMCollaboration). Adapted from Yuan et al. (2021).

Require: learning rate η, batch size M
  Randomly initialize $\theta_i, \hat{\theta}_i, \eta_Q, \eta_\pi, \eta_f, \eta_l, i \in \{1, \dots, N\}$
  for each round **do**
    for $i \in \{1, \dots, N\}$ **do**
      Initialize replay buffer $\mathcal{D} \leftarrow \emptyset$
      while train agent i **do**
        Run ToMCollaborationFixedAgent(
          $\theta_i, \hat{\theta}_i, \eta_Q, \eta_\pi, \eta_f, i, \mathcal{D}$)
      end while
    end for
  end for

from checkpoints saved at different stages of training (begin, half way and end). The idea is that different checkpoints represent different skill levels. Since our agents have one of three different skill levels (begin, half way and end) we also task each agent with predicting the skill level of its partner as this also can be easily added as a label to the offline RL dataset as $l_i \in \{\text{beginner: } 0, \text{advanced: } 1, \text{expert: } 2\}$. This is inspired by humans who typically reason about the skill level of a cooperation partner and thereby adjusting their behaviour (i.e. working together with a toddler versus with an adult). Besides overcoming the need for human data, we view this approach as an additional source of robustness, helping agents to better deal with novel partners having different skill levels.

Our intuition is that adding ToM to collaborative agents will increase their robustness and thus have a different goal than Yuan et al. (2021) who where mainly interested in better overall performance. To recap, our method iterates on the method of Yuan et al. (2021) in three important ways: (i) we generalize their method to non-fixed mental states, (ii) we also predict the capabilitites of the partner and (iii) we integrate the approach with FCP for increased partner diversity.

Lastly, while our proposed method works without any human data, we believe that fine-tuning on human data could in principle improve our approach. Recent work supports this idea. Yang et al. (2022) proposed the optimal behaviour prior algorithm (OBP) for fine tuning with human-human data and Hong et al. (2023) introduced a method for learning from human-human game-play in an offline RL setting by augmenting the dataset $\mathcal{D}$. Given that in this work we want to focus on building agents that effectively collaborate with humans without the need for human data, we keep this different but promising line of work as optional.

Implementation Details As a starting point, we will use the agent architecture used in (Strouse et al., 2021), consisting of a ResNet for visual observations processing and an LSTM for temporal modelling. Given the sequential nature of our data, we will also consider testing deep learning architectures based on the transformer neural netowk (Vaswani et al., 2017). Additionally, we plan to apply conservative Q-Learning (CQL) (Kumar et al., 2020) – a regularization technique for handling distributional shifts in offline RL.

For our experiments we plan to use PantheonRL, a Python package for multi-agent reinforcement learning provided by Sarkar et al. (2022) that also implements the Overcooked-AI environment. Additionally, it also features many other environments and training paradigms which makes additional evaluations possible.

4.4. Evaluation

Following Knott et al. (2021), we evaluate our agent we will use two metrics: (i) average success rate and (ii) average validation reward, both averaged across the four original layouts. However, as Knott et al. (2021) pointed out, average reward cannot be a sufficient metric to evaluate robustness. They proposed to use a collection of manually designed unit tests that capture edge cases in collaboration for each task. While they admit that these can never be complete, we are convinced that they are nonetheless more informative compared to average reward. Therefore, we also evaluate our collaborative agent on their unit tests.

Algorithm 3 FictitiousCoPlay$_{\text{ToM}}$

Require: $\mathcal{E}$, a distribution of E environments
  Randomly initialize N collaboration agents via their parameters $\theta_C^i,\ i \in N$
  for Randomly pick $e \in \mathcal{E}$ **do** ▷ Train Diverse Pool of Partners; see (Yang et al., 2022) step 1
    for Randomly pick $i \in N$ **do**
      Run ToMCollaboration Algorithm 2
    end for
  end for
  Train best response to population θ_C^N using ToMCollaborationFixedAgent Algorithm 1 ▷ see (Yang et al., 2022) step 2

We note that different papers using the Overcooked-AI environment do not report the same or comparable metrics. This is mainly due to the different research question that each try to answer. While some report average reward (Zhao et al., 2021) or number of deliveries (Strouse et al., 2021; Nalepka et al., 2021), others follow individual metrics specific to their research questions (Ribeiro et al., 2022; Fontaine et al., 2021), for instance unit-testing (Knott et al., 2021). It thus seems to be necessary to compare the different metrics that these works employ and find a good measure for evaluating zero-shot collaboration. To this end, we plan to track many evaluation metrics and reason about their information content regarding novel partner cooperation. Which metrics to report on will be evaluated as part of the thesis. Our goal is that this clarifies the state of current research and acts as a guideline.

5. Intended Outcomes

Core to this work is to improve Human-AI collaboration via quality and diversity of training partners and models. Our work has the following intended outcomes:

- A method for training agents capable of cooperating efficiently in zero-shot collaboration settings, i. e. with novel agents.
- Evaluate whether an agent performs well in zero-shot collaboration and finding an informative measure or set of measures to quantify that.
- A series of ablation studies to study how ToM and FCP individually impact our agent's performance and unit tests using the suite provided by Knott et al. (2021).

6. Mandatory and Optional Goals

6.1. Mandatory Goals

This thesis can be split in three milestones that correspond to mandatory goals:

1. The adaption of Adaptive ToM Collaboration for non-fixed private state and its integration into Overcoocked as an improvement for the quality of the model.
2. Implement FCP on top of that for a diversification of training partners.
3. Evaluate the method by:
   (a) Surveying the literature for current metrics for evaluating performance in the Overcooked environment.
   (b) Evaluate the ideas of unit testing for checking robustness of Knott et al. (2021) for the obtained model.
   (c) Report performance in terms of an metric assemble if appropriate but at least via (i) success rate, (ii) average validation reward and (iii) unit test success.

Additionally, I also commit to the following goals:

4. Evaluate and survey SOTA Q-Learning architectures.
5. Investigate CQL as additional training regularization.
6. Use the Q function of the model to investigate the models behavior and look for robustness failures qualitatively such as misgeneralization (Langosco et al., 2022).

6.2. Optional Goals

Human-Human Data As shortly outlined in the methods section, human-human gampeplay data can be used to fine-tune a model. Specifically, to adjust to failures that are somehow innate to humans. If time allows, we would like to integrate the optimal behaviour prior algorithm (Yang et al., 2022) and look into possible inspirations from the work of Hong et al. (2023).

Human Study Optionally, we want to evaluate the performance of the trained agents when paired with human players. We will ask participants to play together with our agents and other baselines, and analyze both objective performance as well as subjective preferences.

Other Environments Next, while the most important literature to this project is from the Overcooked-AI environment, other cooperative games environments like Pursuit Evastion, Waterworld, Multi-Agent Walker and Multi-Ant from Gupta et al. (2017) as provided by the PettingZoo[1] library (Terry et al., 2021) exist and could potentitally be interesting.

7. Schedule

The thesis is scheduled to start on April, 15th 2023 and end on October, 15th 2023. We outline our schedule in Figure 4.

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1, 2004.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P. F., Schulman, J., and Mané, D. Concrete problems in AI safety. *CoRR*, abs/1606.06565, 2016. URL http://arxiv.org/abs/1606.06565.

Bain, M. and Sammut, C. A framework for behavioural cloning. In *Machine Intelligence 15, Intelligent Agents [St. Catherine's College, Oxford, July 1995]*, pp. 103–129, GBR, 1999. Oxford University. ISBN 0198538677.

Bara, C.-P., CH-Wang, S., and Chai, J. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1112–1125, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.85. URL https://aclanthology.org/2021.emnlp-main.85.

Bard, N., Foerster, J. N., Chandar, S., Burch, N., Lanctot, M., Song, H. F., Parisotto, E., Dumoulin, V., Moitra, S., Hughes, E., Dunning, I., Mourad, S., Larochelle, H., Bellemare, M. G., and Bowling, M. The hanabi challenge: A new frontier for AI research. *CoRR*, abs/1902.00506, 2019. URL http://arxiv.org/abs/1902.00506.

Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, 2014.

Boutilier, C. Planning, learning and coordination in multiagent decision processes. In *Proceedings of the 6th Conference on Theoretical Aspects of Rationality and Knowledge*, TARK '96, pp. 195–210, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc. ISBN 1558604179.

Carpenter, M., Call, J., and Tomasello, M. Twelve-and 18-month-olds copy actions in terms of goals. *Developmental science*, 8(1):F13–F20, 2005.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., and Dragan, A. On the utility of learning about humans for human-ai coordination. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/f5b1b89d98b7286673128a5fb112cb9a-Paper.pdf.

Charakorn, R., Manoonpong, P., and Dilokthanakul, N. Investigating partner diversification methods in cooperative multi-agent deep reinforcement learning. In *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part V 27*, pp. 395–402. Springer, 2020.

Choudhury, R., Swamy, G., Hadfield-Menell, D., and Dragan, A. D. On the utility of model learning in hri. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '19, pp. 317–325. IEEE Press, 2020. ISBN 9781538685556.

Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., Larson, K., and Graepel, T. Open problems in cooperative AI. *CoRR*, abs/2012.08630, 2020. URL https://arxiv.org/abs/2012.08630.

Fisac, J. F., Gates, M. A., Hamrick, J. B., Liu, C., Hadfield-Menell, D., Palaniappan, M., Malik, D., Sastry, S. S., Griffiths, T. L., and Dragan, A. D. Pragmatic-pedagogic value alignment. In *Robotics Research: The 18th International Symposium ISRR*, pp. 49–57. Springer, 2020.

Fontaine, M. C., Hsu, Y.-C., Zhang, Y., Tjanaka, B., and Nikolaidis, S. On the importance of environments in human-robot coordination. *arXiv preprint arXiv:2106.10853*, 2021.

Fuchs, A., Walton, M., Chadwick, T., and Lange, D. Theory of mind for deep reinforcement learning in hanabi. *CoRR*, abs/2101.09328, 2021. URL https://arxiv.org/abs/2101.09328.

[1] PettingZoo is part of a Python package named PantheonRL (Sarkar et al., 2022) which also additionally provides Overcooked-AI and other environments in a single library which makes it a great basis for the implementation.

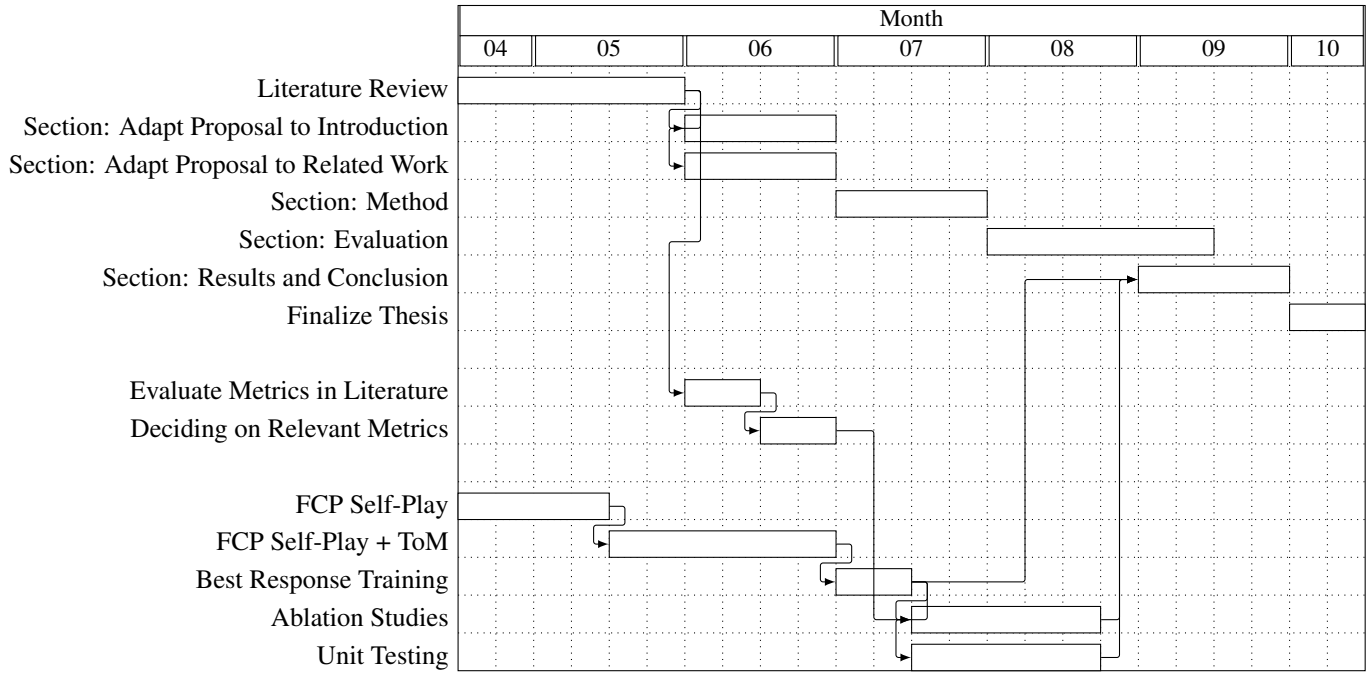

Figure 4. Schedule

Gergely, G. and Csibra, G. Teleological reasoning in infancy: The infant's naive theory of rational action. *Cognition*, 63(2):227–233, 1997. doi: 10.1016/s0010-0277(97)00004-8.

Gergely, G., Nádasdy, Z., Csibra, G., and Bíró, S. Taking the intentional stance at 12 months of age. *Cognition*, 56(2):165–193, 1995. ISSN 0010-0277. doi: https://doi.org/10.1016/0010-0277(95)00661-H. URL https://www.sciencedirect.com/science/article/pii/001002779500661H.

Gupta, J. K., Egorov, M., and Kochenderfer, M. Cooperative multi-agent control using deep reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems*, pp. 66–83. Springer, 2017.

Heider, F. and Simmel, M. An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2):243, April 1944. doi: 10.2307/1416950. URL https://doi.org/10.2307/1416950.

Hernandez, D., Denamganaï, K., Gao, Y., York, P., Devlin, S., Samothrakis, S., and Walker, J. A. A generalized framework for self-play training. In *2019 IEEE Conference on Games (CoG)*, pp. 1–8, 2019. doi: 10.1109/CIG.2019.8848006.

Ho, J. and Ermon, S. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Hong, J., Dragan, A., and Levine, S. Learning to influence human behavior with offline reinforcement learning, 03 2023.

Hu, H. and Foerster, J. N. Simplified action decoder for deep multi-agent reinforcement learning. *CoRR*, abs/1912.02288, 2019. URL http://arxiv.org/abs/1912.02288.

Hu, H., Lerer, A., Peysakhovich, A., and Foerster, J. N. "other-play" for zero-shot coordination. *CoRR*, abs/2003.02979, 2020. URL https://arxiv.org/abs/2003.02979.

Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W. M., Donahue, J., Razavi, A., Vinyals, O., Green, T., Dunning, I., Simonyan, K., et al. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017.

Javdani, S., Admoni, H., Pellegrinelli, S., Srinivasa, S. S., and Bagnell, J. A. Shared autonomy via hindsight optimization for teleoperation and teaming. *The International Journal of Robotics Research*, 37(7):717–742, 2018. doi: 10.1177/0278364918776060.

Klien, G., Woods, D., Bradshaw, J., Hoffman, R., and Feltovich, P. Ten challenges for making automation a "team player" in joint human-agent activity. *IEEE Intelligent Systems*, 19(6):91–95, 2004. doi: 10.1109/MIS.2004.74.

Knott, P., Carroll, M., Devlin, S., Ciosek, K., Hofmann, K., Dragan, A. D., and Shah, R. Evaluating the robustness of collaborative agents. *CoRR*, abs/2101.05507, 2021. URL https://arxiv.org/abs/2101.05507.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. *CoRR*, abs/2006.04779, 2020. URL https://arxiv.org/abs/2006.04779.

Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Perolat, J., Silver, D., and Graepel, T. A unified game-theoretic approach to multiagent reinforcement learning, 2017. URL https://arxiv.org/abs/1711.00832.

Langosco, L. L. D., Koch, J., Sharkey, L. D., Pfau, J., and Krueger, D. Goal misgeneralization in deep reinforcement learning. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12004–12019. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/langosco22a.html.

Liu, A., Zhu, H., Liu, E., Bisk, Y., and Neubig, G. Computational language acquisition with theory of mind. *arXiv preprint arXiv:2303.01502*, 2023.

Mutlu, B., Terrell, A., and Huang, C.-M. Coordination mechanisms in human-robot collaboration. In *Proceedings of the Workshop on Collaborative Manipulation, 8th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 1–6. Citeseer, 2013.

Nalepka, P., Gregory-Dunsmore, J., Simpson, J., Patil, G., and Richardson, M. Interaction flexibility in artificial agents teaming with humans. 07 2021.

Nguyen, D., Nguyen, P., Le, H., Do, K., Venkatesh, S., and Tran, T. Learning theory of mind via dynamic traits attribution. *arXiv preprint arXiv:2204.09047*, 2022.

Nguyen, D., Nguyen, P., Le, H., Do, K., Venkatesh, S., and Tran, T. Memory-augmented theory of mind network. *arXiv preprint arXiv:2301.06926*, 2023.

Nikolaidis, S. and Shah, J. Human-robot cross-training: Computational formulation, modeling and evaluation of a human team training strategy. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 33–40, 2013. doi: 10.1109/HRI.2013.6483499.

Puig, X., Shu, T., Li, S., Wang, Z., Tenenbaum, J. B., Fidler, S., and Torralba, A. Watch-and-help: A challenge for social perception and human-ai collaboration. *CoRR*, abs/2010.09890, 2020. URL https://arxiv.org/abs/2010.09890.

Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S. M. A., and Botvinick, M. M. Machine theory of mind. *CoRR*, abs/1802.07740, 2018. URL http://arxiv.org/abs/1802.07740.

Repacholi, B. M. and Gopnik, A. Early reasoning about desires: evidence from 14- and 18-month-olds. *Developmental psychology*, 33 1:12–21, 1997.

Ribeiro, J. G., Martinho, C., Sardinha, A., and Melo, F. S. Assisting unknown teammates in unknown tasks: Ad hoc teamwork under partial observability. *CoRR*, abs/2201.03538, 2022. URL https://arxiv.org/abs/2201.03538.

Sadigh, D., Sastry, S., Seshia, S. A., and Dragan, A. D. Planning for autonomous cars that leverage effects on human actions. In *Robotics: Science and Systems*, 2016.

Sarkar, B., Talati, A., Shih, A., and Dorsa, S. Pantheonrl: A marl library for dynamic training interactions. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (Demo Track)*, 2022.

Sclar, M., Neubig, G., and Bisk, Y. Symmetric machine theory of mind. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19450–19466. PMLR, 17–23 Jul 2022.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T. P., Simonyan, K., and Hassabis, D. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *CoRR*, abs/1712.01815, 2017a. URL http://arxiv.org/abs/1712.01815.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, October 2017b. doi: 10.1038/nature24270. URL https://doi.org/10.1038/nature24270.

Stefano, A. and Ramamoorthy, S. A game theoretic model and best-response learning method for ad hoc coordination in multi-agente systems. In *Proc. of the 12th International Conference on Autonomous Agents and Multiagent Systems*, 2013.

Stone, P., Kaminka, G., Kraus, S., and Rosenschein, J. Ad hoc autonomous agent teams: Collaboration without pre-coordination. volume 3, 01 2010.

Strouse, D., McKee, K., Botvinick, M., Hughes, E., and Everett, R. Collaborating with humans without human data. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 14502–14515. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/797134c3e42371bb4979a462eb2f042a-Paper.pdf.

Terry, J., Black, B., Grammel, N., Jayakumar, M., Hari, A., Sullivan, R., Santos, L. S., Dieffendahl, C., Horsch, C., Perez-Vicente, R., et al. Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:15032–15043, 2021.

Torabi, F., Warnell, G., and Stone, P. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pp. 4950–4957. AAAI Press, 2018. ISBN 9780999241127.

Tversky, A. and Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157): 1124–1131, 1974. doi: 10.1126/science.185.4157.1124. URL https://www.science.org/doi/abs/10.1126/science.185.4157.1124.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, R. E., Wu, S. A., Evans, J. A., Tenenbaum, J. B., Parkes, D. C., and Kleiman-Weiner, M. Too many cooks: Coordinating multi-agent collaboration through inverse planning. *CoRR*, abs/2003.11778, 2020. URL https://arxiv.org/abs/2003.11778.

Yang, M., Carroll, M., and Dragan, A. D. Optimal behavior prior: Data-efficient human models for improved human-ai collaboration. *ArXiv*, abs/2211.01602, 2022.

Yu, M., Sang, Y., Pu, K., Wei, Z., Wang, H., Li, J., Yu, Y., and Zhou, J. Few-shot character understanding in movies as an assessment to meta-learning of theory-of-mind. *ArXiv*, abs/2211.04684, 2022.

Yuan, L., Fu, Z., Zhou, L., Yang, K., and Zhu, S. Emergence of theory of mind collaboration in multiagent systems. *CoRR*, abs/2110.00121, 2021. URL https://arxiv.org/abs/2110.00121.

Zhao, R., Song, J., Haifeng, H., Gao, Y., Wu, Y., Sun, Z., and Wei, Y. Maximum entropy population based training for zero-shot human-ai coordination. *arXiv preprint arXiv:2112.11701*, 2021.

Zhou, P., Zhu, A., Hu, J., Pujara, J., Ren, X., Callison-Burch, C., Choi, Y., and Ammanabrolu, P. An ai dungeon master's guide: Learning to converse and guide with intents and theory-of-mind in dungeons and dragons. *arXiv preprint arXiv:2212.10060*, 2022.