

# Limits of Theory of Mind Modelling in Dialogue-Based Collaborative Plan Acquisition

Matteo Bortoletto   Constantin Ruhdorfer<sup>†</sup>   Adnen Abdessaied<sup>†</sup>  
Lei Shi   Andreas Bulling  
University of Stuttgart, Germany  
matteo.bortoletto@vis.uni-stuttgart.de

## Abstract

Recent work on dialogue-based collaborative plan acquisition (CPA) has suggested that Theory of Mind (ToM) modelling can improve missing knowledge prediction in settings with asymmetric skill sets and knowledge. Although ToM was claimed to be important for effective collaboration, its real impact on this novel task remains under-explored. By representing plans as graphs and exploiting task-specific constraints we show that, as performance on CPA nearly doubles when predicting one’s own missing knowledge, the improvements due to ToM modelling diminish. This phenomenon persists even when evaluating existing baseline methods. To better understand the relevance of ToM for CPA, we report a principled performance comparison of models with and without ToM features. Results across different models and ablations consistently suggest that features learnt for ToM tasks are more likely to reflect latent patterns in the data with no perceivable link to ToM. This finding calls for a deeper understanding of the role of ToM in CPA and beyond, as well as new methods for modelling and evaluating mental states in computational collaborative agents.

## 1 Introduction

Dialogue-based human-AI collaboration is an interaction in which humans and artificial intelligent (AI) agents converse to achieve a shared goal or task (Streeck et al., 2011). When humans collaborate with each other, they rely on two main abilities: Verbal communication and Theory of Mind (ToM), i.e. the ability to infer one’s own and others’ mental states (Premack and Woodruff, 1978). To succeed in collaborating with humans, it is therefore imperative for AI agents to possess similar capabilities (Williams et al., 2022).

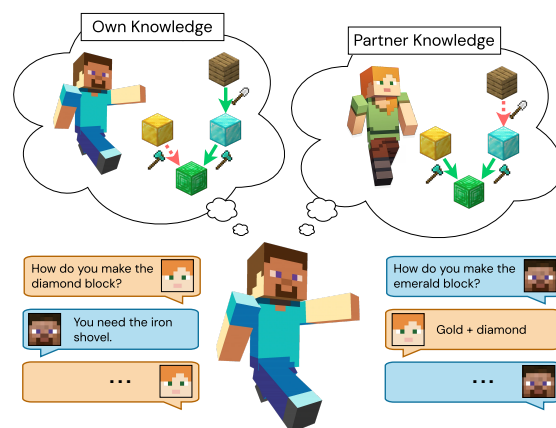



Figure 1: Collaborative plan acquisition in Minecraft involves inferring one’s own and the partner’s missing knowledge ( $\dashrightarrow$ ) through situated dialogue, starting from individual partial plans ( $\rightarrow$ ) to achieve a shared goal.

Recent work on this topic has introduced collaborative plan acquisition (CPA) as a promising task for evaluating collaborative abilities in agents (Bara et al., 2023). Starting from asymmetric knowledge and skill-sets of two collaborating agents, the goal of CPA is to infer one’s own missing knowledge (OMK) and the partner’s missing knowledge (PMK) to achieve a shared goal by engaging in a multi-round situated dialogue (see Figure 1). To study CPA, the authors used *MindCraft* – a multi-modal collaborative dialogue-based benchmark grounded in the sandbox game Minecraft (Bara et al., 2021). They also proposed a sequence-to-sequence CPA model that used visual observations, plan, and dialogue history as input. Their empirical results showed a large difference in performance between predicting OMK and PMK. Furthermore, they found that while including a subset of ToM features improved performance, using all ToM features resulted in nearly the same performance as using none.

In this work, we systematically analyse the limits

<sup>†</sup>Equal second-author contribution.

 The project web page is accessible [here](#).

of ToM modelling in dialogue-based collaborative plan acquisition. We first propose to represent plans as directed graphs, each represented by a node feature matrix, a connectivity matrix, and an edge feature matrix. This starkly contrasts previous works (Bara et al., 2021, 2023) that represented plans as lists consisting of materials and tools processed by GRU (Cho et al., 2014). Our novel structured representation allows us to elegantly frame CPA as a link prediction task (Liben-Nowell and Kleinberg, 2007) and apply negative sampling constrained on *MindCraft* plans’ structure for efficient training. Our proposed representation not only doubles the performance of predicting OMK compared to (Bara et al., 2023) but also bridges the gap to predicting PMK.

However, our evaluations show no significant performance difference when using ToM features – neither for our method nor for the baselines from Bara et al. (2023). We thus conduct extensive analyses using diagnostic probing, correlation analysis, and ground-truth ToM labels as input. Results across different models and ablations consistently suggest that learned ToM features are less associated with mental states and more aligned with revealing latent patterns within the data.

In summary, the contributions of our work are two-fold: (1) We propose a novel graph-based representation of plans for CPA and show that applying graph learning methods simultaneously doubles the performance of predicting OMK and closes the gap to predicting PMK; (2) We report principled analyses across different models and ablations that suggest that learnt ToM features reflect latent patterns in the data with no perceivable link to ToM.

## 2 Related Work

### 2.1 Dialogue-based Human-AI Collaboration

Collaborative dialogue systems are designed to work with humans towards achieving a shared goal (Rich et al., 2001; Bohus and Rudnicky, 2009; Allen et al., 2002; Streeck et al., 2011). Early works were based on scripts (Traum, 2017), employed planning (Papaioannou et al., 2018), or modelled the dialogue as a collection of information states (Larsson and Traum, 2000). More recent work focused on neural sequence-to-sequence models to learn from dialogue corpora (Wen et al., 2015; Dong et al., 2023). Neural approaches have also been explored for collaborative dialogues taking place when participants are working on

a shared artefact within a co-observed environment (Narayan-Chen et al., 2019; Kim et al., 2019; Jayannavar et al., 2020; Bara et al., 2021, 2023). Another line of work explored the role of Theory of Mind (Premack and Woodruff, 1978) in dialogue-based collaboration, focusing on simulated textual environments (Qiu et al., 2022; Zhou et al., 2023), or on human gameplay (Bara et al., 2023).

Our work focuses on the *MindCraft* environment (Bara et al., 2021, 2023) in which two agents with asymmetric skill sets and knowledge converse to achieve a shared goal in a Minecraft world.

### 2.2 Computational Theory of Mind

With recent advances in AI, an increasing number of works studied means to equip models with Theory of Mind capabilities based on deep learning approaches (Rabinowitz et al., 2018; Bara et al., 2021; Gandhi et al., 2021; Zhou et al., 2023; Liu et al., 2023; Bortoletto et al., 2024), partially observable Markov decision processes (Doshi et al., 2010; Han and Gmytrasiewicz, 2018) or via Bayesian approaches (Baker et al., 2009; Lee et al., 2019; Buehler and Weisswange, 2020; Fan et al., 2021). Within these, one line of work focused on inferring beliefs, actions, or instructions solely *as an observer* of agent behaviour (Rabinowitz et al., 2018; Grant et al., 2017; Duan et al., 2022). An emerging second line of work explored ToM *from the perspective* of interacting agents (Wang et al., 2021; Qiu et al., 2022; Bara et al., 2023), highlighting the significance of ToM in collaborative tasks. Bara et al. (2023) claimed that integrating ToM features improves collaborative plan acquisition (CPA). However, the limits and failure cases of ToM-enabled agents are poorly understood, particularly whether they really model mental states or exploit dataset biases. This work aims to address these concerns and assess ToM modelling on CPA.

## 3 Problem Formulation

*MindCraft* (Bara et al., 2021) was introduced as a multi-modal benchmark for studying ToM modelling within collaborative tasks. It involves two players collaborating through dialogue in a 3D block world to craft a target material by manipulating blocks using specific tools (see Figure 3, left). Players initially receive a partial plan as an incomplete directed AND-graph and a tool allowing each to interact with a set of specific blocks. Endowed with complementary knowledge and skill sets, play-

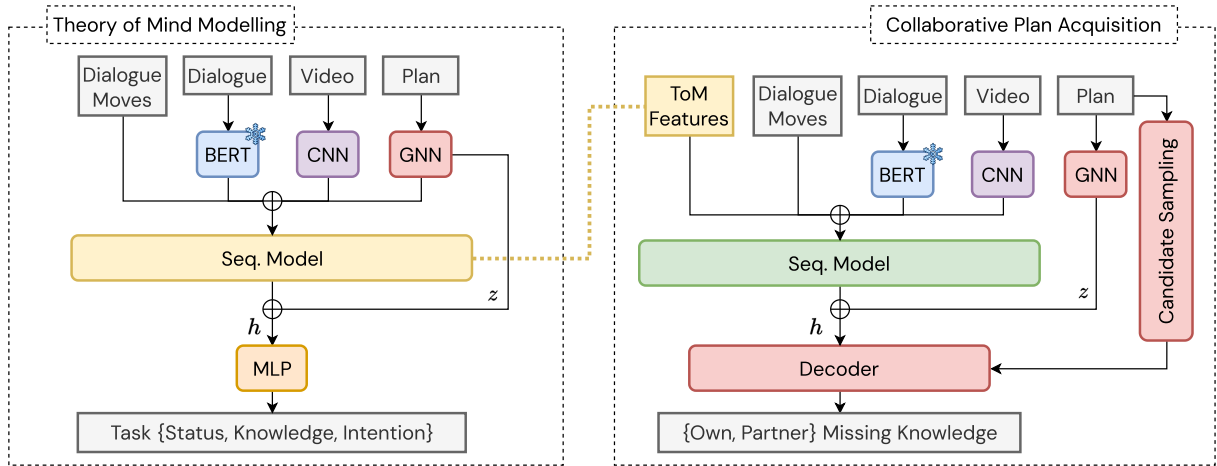


Figure 2: Models’ architecture for Theory of Mind (ToM) modelling and collaborative plan acquisition (CPA). Following (Bara et al., 2021, 2023), we train one model for each ToM task (*Status*, *Knowledge*, *Intention*) and CPA task (Own Missing Knowledge, Partner’s Missing Knowledge) and freeze the BERT weights during training (indicated by ❄️) for a fair comparison with the baseline.

ers must communicate via the in-game chat to craft the target material and reason about each other’s mental states. The ToM tasks introduced by Bara et al. (2021) are specifically designed to capture mental state information pertinent to collaboration. During gameplay, players are presented with pop-up questions every 75 seconds, each paired by type:

1. *Task Status*: Predict if one of the two players has created a specific material. For instance, Player 1 is asked: “Has your partner created GOLD\_BLOCK so far?” and Player 2 is asked: “Have you crafted GOLD\_BLOCK yet?” Possible answers are YES, NO, or MAYBE.
2. *Player Knowledge*: Predict whether players know how to craft a material or if they believe their partner knows. For example, Player 1 is asked: “Do you think the other player knows how to make BLUE\_WOOL?” and Player 2 is asked: “Do you know how to make BLUE\_WOOL?” Possible answers are YES, NO, or MAYBE.
3. *Player Intention*: Predict which material a player is making at the current time step. For example, Player 1 is asked: “What do you think the other player is making right now?” and Player 2 is asked: “What are you making right now?”. Possible answers are the different types of blocks in the game or NOT\_SURE.

In this work, we focus on a recent extension of *MindCraft* by Bara et al. (2023) in which they proposed collaborative plan acquisition (CPA) and

explored the role of ToM modelling in predicting players’ missing knowledge while executing the crafting tasks. CPA is formulated as follows:

**Definition 1** Consider a joint plan as a directed AND-graph  $\mathcal{P} = (V, E)$ , where the nodes  $V$  denote (sub-)goal materials, and edges  $E$  denote temporal constraints between the sub-goals. In a collaborative plan acquisition problem, two agents  $i$  and  $j$  start with partial plans  $\mathcal{P}_i = (V, E_i)$ ,  $E_i \subseteq E$ , and  $\mathcal{P}_j = (V, E_j)$ ,  $E_j \subseteq E$ . Given a sequence of visual observations  $O_i^t$  and a joint dialogue history  $D^t$  at time  $t$ , agent  $i$  has to infer their own missing knowledge  $\bar{E}_i = E \setminus E_i$  and the partner  $j$ ’s missing knowledge  $\bar{E}_j = E \setminus E_j$ .

ToM and CPA tasks are closely related but fundamentally different: ToM tasks focus on exploring players’ *beliefs* about the game state and their partner’s mental states, while CPA tasks involve predicting *missing information* from players’ partial plans. Additional details are in §A.1 and §A.2.

## 4 Method

### 4.1 Baseline

As a baseline we used the model proposed by Bara et al. (2023) that embeds dialogue utterances using a frozen pre-trained BERT model (Devlin et al., 2019) and represents each utterance by the features obtained from processing the corresponding [CLS] token using linear layers with tanh activation. A CNN and GRU were used to process the video frames and partial plans, respectively. During training, a model was first trained for each of the three

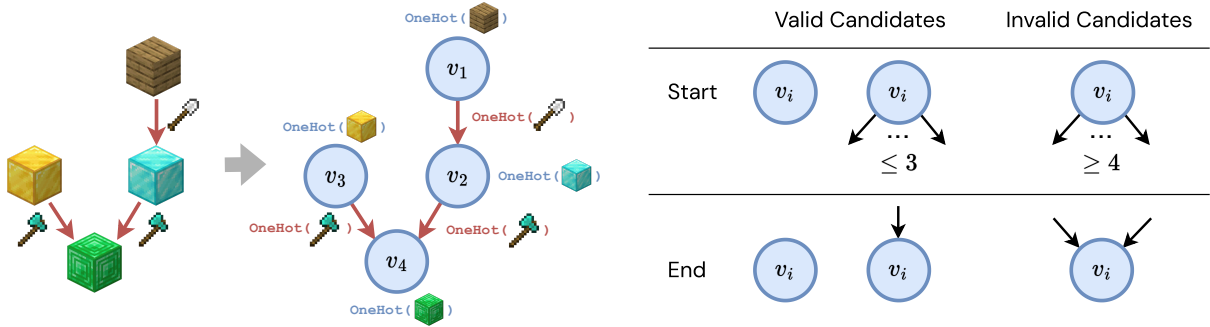


Figure 3: **Left:** Example of a plan graph. Nodes represent materials, edges connect each material to the requisite components for its synthesis, and edge features denote the tool needed to interact with a material. **Right:** Our candidate sampling strategy for predicting OMK: Valid candidate start-nodes must have an out-degree smaller than four whereas candidate end-nodes must have an in-degree less than or equal to one.

ToM tasks of §3 as classification tasks, requiring them to predict players’ answers. Afterwards, two separate models were trained on OMK and PMK by feeding the concatenation of the input modalities and the learnt ToM features into an LSTM (Hochreiter and Schmidhuber, 1997) followed by an MLP that outputs softmax scores for each possible missing link.

## 4.2 GNN-based Missing Knowledge Prediction

**Representing Plans as Graphs.** We propose a modification to the method by (Bara et al., 2023) that includes representing plans as graph objects and using a GNN-based encoder-decoder together with candidate sampling to predict missing edges (see Figure 2). More specifically, as shown in Figure 3, nodes represent materials, edges connect each material to the prerequisite components for its synthesis, and edge features denote the specific tool needed to interact with the material. This structured representation allows us to more naturally represent the dependencies between materials and information about the tools involved in the crafting process. Most importantly, it allows us to elegantly frame CPA as a link prediction task.

**Theory of Mind Modelling.** We modify the ToM modelling architecture of (Bara et al., 2021) with two key changes: First, we replaced the GRU plan encoder with (Brody et al., 2022), followed by average graph pooling. Second, we substituted the LSTM with a single-block Transformer (Vaswani et al., 2017). The rest of the architecture remains unchanged to ensure a fair comparison.

**Missing Knowledge Prediction.** By representing plans as graphs we can perform missing knowl-

edge prediction by applying negative sampling – a common technique used for graph completion (Yang et al., 2015; Schlichtkrull et al., 2018). A GNN encoder  $g$  first maps each node  $v_i \in V$  to a real-valued vector  $z_i = g(v_i) \in \mathbb{R}^d$ . We then add  $\Omega$  negative edges to the original graph. In conventional link prediction, negative edges are sampled randomly starting from the complete graph, and the goal is to classify edges as true or fake. In contrast, our approach begins with an incomplete graph and the task is to predict missing edges. Randomly selecting negative edges poses the risk of missing the edges we aim to predict. Our improved approach uses the plan structural constraints of *MindCraft* to narrow down the pool of all possible  $V^2 \setminus E_{\{i,j\}}$  edges to a set of *valid* candidates  $\Omega$  (see Figure 3, right). Specifically, valid candidate start-nodes must have an out-degree smaller than four whereas candidate end-nodes must have an in-degree less than or equal to one. We also exclude the starting set of game materials from the candidate end-nodes. We call this technique *candidate sampling*.

Afterwards, a decoder classifies edges as positive or negative by relying on the node embeddings. In particular, we learn a scoring function  $s : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  using a linear layer  $f$  that takes as input the concatenation of the two node embeddings corresponding to the candidate edge, and the output  $c$  of the sequence model, that serves as a context:  $s(v_i, v_j) = \hat{y}_{ij} = f(z_i \oplus z_j \oplus c)$ , where  $\oplus$  denotes the concatenation operator. In contrast to (Bara et al., 2023) who used an LSTM to process the sequential data, we used a single Transformer block. Finally, we optimise with the binary cross-entropy loss function  $\mathcal{L}$ , which maximises the likelihood of positive edges while minimising that of

<i>Status</i>			
Modalities	Bara et al. (2023)	Ours	Human
M	47.7 ± 0.6	59.9 ± 0.7	67.0
D+M	45.5 ± 2.3	59.1 ± 0.6	67.0
D+V+M	45.2 ± 1.8	58.9 ± 0.8	67.0
V+M	47.3 ± 0.7	59.6 ± 0.4	67.0
<i>Knowledge</i>			
Modalities	Bara et al. (2023)	Ours	Human
M	51.5 ± 1.1	57.9 ± 0.2	58.0
D+M	50.0 ± 1.5	57.2 ± 1.5	58.0
D+V+M	50.2 ± 1.1	57.5 ± 1.7	58.0
V+M	50.5 ± 1.6	57.6 ± 1.8	58.0
<i>Intention</i>			
Modalities	Bara et al. (2023)	Ours	Human
M	9.1 ± 0.2	11.7 ± 2.2	46.0
D+M	8.7 ± 2.1	11.1 ± 1.8	46.0
D+V+M	10.5 ± 2.3	12.1 ± 2.4	46.0
V+M	9.0 ± 0.3	13.4 ± 1.9	46.0

Table 1: Performance comparison on the three ToM tasks using different combinations of modalities: dialogue moves (M), dialogue (D), and video frames (V). We report the F1 scores obtained by the baseline (Bara et al., 2023), our model, and humans.

sampled negative edges:

$$\mathcal{L} = - \sum_{(i,j) \in E} \log(\sigma(\hat{y}_{ij})) - \sum_{(i,k) \in \Omega} \log(1 - \sigma(\hat{y}_{ik}))$$

where  $\sigma$  indicates the sigmoid function. Additional details about the model’s architecture and training are provided in §A.3.

## 5 Experiments

### 5.1 Theory of Mind Modelling

We first report the performance of our model on the three ToM tasks introduced in Section 3. As summarised in Table 1, our model outperforms the baseline<sup>1</sup> on all three tasks, underlining the efficiency of the proposed GNN-based approach. Notably, as highlighted in **green** in Table 1, our model manages to even match human performance in the *Knowledge* task. However, performance on the other two tasks is still far from a human level, especially on *Intention*. This might be attributed to the fact that, unlike *Knowledge* that does not require temporal modelling and could be solved by

<sup>1</sup>Despite training the baseline model (Bara et al., 2023) using the official code, its performance slightly deviated from the original paper, and discussions with the authors did not yield clarity. See §A.4 for further details and comparisons.

using plan information, both *Status* and *Intention* require accurate temporal modelling, which has to be kept coherent across the different input modalities. As can also be seen from the table, ablations of different input modalities have little impact on the final performance of our model and the baseline for all ToM tasks.

### 5.2 Collaborative Plan Acquisition (CPA)

Subsequently, we evaluate our model on the CPA task following Bara et al. (2023). We always use dialogue moves as input since they were shown to have a positive impact on performance. As can be seen from Table 2, our model consistently achieves overall F1 scores of over 56.6 thereby significantly outperforming the baseline of Bara et al. (2023) in all evaluation settings.

**Own Missing Knowledge (OMK).** We first analyse the task of predicting one’s own missing knowledge. As can be seen in Table 2, our model manages to double the performance of the baseline<sup>2</sup> by consistently achieving F1 scores of over 57%. In stark contrast to the baseline, which performs best when using only ToM features extracted from *Intention*, our model’s best performance is obtained by additionally incorporating features extracted from *Knowledge*. The benefit of these features on CPA is expected and intuitively makes sense since *Knowledge* was the ToM task for which our models achieved human-level performance (see Table 1).

**Partner’s Missing Knowledge (PMK).** Second, we evaluate the task of predicting the partner’s missing knowledge. In contrast to prior work (Bara et al., 2023), our evaluations reveal a significantly reduced performance gap between predicting the different types of missing knowledge (OMK vs PMK) as can be seen in the second part of Table 2. As highlighted in **blue**, this can be attributed to our proposed candidate sampling approach that, contrarily to naive sampling, effectively narrowed down the pool of valid candidate edges for one’s own missing knowledge to a similar order of magnitude as that of the partner’s. The difference in performance compared to the baseline is likely due to the choice of cost function used for training.

**Statistical Tests.** Although our model attained improved results on CPA, especially in predicting one’s own missing knowledge, it did so without

<sup>2</sup>In this case, the scores are higher than the ones reported in (Bara et al., 2023).

ToM Features			Overall		OMK			PMK	
<i>Status</i>	<i>Knowledge</i>	<i>Intention</i>	Bara et al. (2023)	Ours	Bara et al. (2023)	Ours (NS)	Ours	Bara et al. (2023)	Ours
			46.6 ± 1.6	<b>56.9 ± 0.6</b>	27.7 ± 2.3	23.7 ± 2.5	57.6 ± 0.8	65.4 ± 0.2	56.2 ± 0.3
✓			46.7 ± 2.0	<b>57.3 ± 0.6</b>	26.1 ± 2.5	26.6 ± 2.4	58.0 ± 0.8	67.2 ± 1.2	56.5 ± 0.3
	✓		47.4 ± 1.7	<b>57.0 ± 1.4</b>	28.0 ± 1.8	24.7 ± 2.6	58.4 ± 0.5	66.8 ± 1.5	55.5 ± 1.9
		✓	47.2 ± 1.9	<b>57.2 ± 0.5</b>	28.0 ± 2.6	26.0 ± 1.2	57.9 ± 0.7	66.3 ± 0.8	56.5 ± 0.3
✓	✓		47.6 ± 1.5	<b>56.6 ± 1.4</b>	28.4 ± 1.4	25.2 ± 0.3	57.7 ± 0.5	66.8 ± 1.5	55.5 ± 1.9
✓		✓	47.6 ± 1.7	<b>57.5 ± 0.6</b>	28.4 ± 1.8	27.2 ± 1.3	58.4 ± 0.8	66.8 ± 1.5	56.5 ± 0.3
	✓	✓	47.2 ± 1.7	<b>57.5 ± 0.6</b>	27.6 ± 1.9	27.7 ± 0.7	58.5 ± 0.8	66.8 ± 1.5	56.4 ± 0.1
✓	✓	✓	47.4 ± 1.8	<b>56.7 ± 0.7</b>	27.9 ± 2.0	26.6 ± 2.9	57.1 ± 1.9	66.8 ± 1.5	56.6 ± 0.2

Table 2: Performance comparison on CPA when training with learnt ToM features. We report the overall F1 scores as well those for own (OMK) and partner (PMK) missing knowledge prediction. NS = Naive Sampling.

ToM Task	ToM	OMK	PMK	Random
<i>Status</i>	60.6	51.6	49.5	46.7
<i>Knowledge</i>	50.9	49.8	50.8	45.1
<i>Intention</i>	10.2	14.1	13.0	9.3

Table 3: F1 scores on ToM tasks for logistic regression models trained using ToM features, CPA features and random noise. OWM and PMK indicate features coming from our model trained, without ToM features as input, on one’s own and partner’s missing knowledge prediction, respectively.

relying much on the learned ToM features. This can be seen from the results of the different ablated versions of Table 2. To study the effect of ToM features on CPA performance, we performed paired t-tests between our model trained without ToM features and versions of our model trained with different sets of ToM features. We can see that the ToM features did not result in any statistically significant performance difference on CPA since all tests resulted in  $p > 0.05$ . Notably, performing the significance testing on the baseline model of (Bara et al., 2023) yielded the same behaviour, i.e.  $p > 0.05$  across all model versions. Therefore, we challenge the utility of the ToM features in CPA by posing the question of whether these features represent actual information about mental states or reflect latent patterns in the data. We empirically answer this by performing various principled experiments ranging from diagnostic probing to correlation analysis over substituting ToM features with ground-truth labels.

### 5.3 Probing for Theory of Mind

Motivated by our results, we formulate the following research question: *Does ToM modelling as proposed by Bara et al. (2023) actually capture mental state information?* To answer this question, we conducted extensive analyses to study the impact of

ToM modelling on CPA from different angles.

#### 5.3.1 Diagnostic Probing

The ToM features used in CPA are obtained by learning the different tasks of Section 3. As a result, we expect such features to exclusively hold some information about the mental state that other models, when trained on different tasks than ToM, simply lack. To validate this intuitive hypothesis, we used diagnostic probing (Alain and Bengio, 2017; Adi et al., 2017; Conneau et al., 2018; Hupkes et al., 2018) and trained a simple logistic regression (LR) model to perform the three ToM classification tasks. We trained the LR model with different inputs in each experiment and tested its performance on the ToM tasks using the test split. More specifically, we considered four different input scenarios: the vanilla ToM features used in the previous experiments, the hidden representations of the transformer from models predicting OMK and PMK (output of the green model in Figure 2), and finally random noise. As seen in Table 3, a LR model trained on ToM features can perform reasonably well on the three tasks, especially *Status*. However, when trained with missing knowledge features, i.e. features completely optimised in the absence of ToM, the LR model achieves comparable performance in *Knowledge* and even better performance in *Intention*.

These findings open up two possible scenarios: (1) The learnt ToM features are more likely to represent latent patterns in the data with no perceivable link to ToM; (2) ToM capabilities spontaneously emerge from training models on CPA.

#### 5.3.2 Correlation Analysis

In this experiment, we explored whether improvements in CPA tasks correlate with the performance of models on ToM tasks, which are then used to extract ToM features. Intuitively, if the ToM fea-

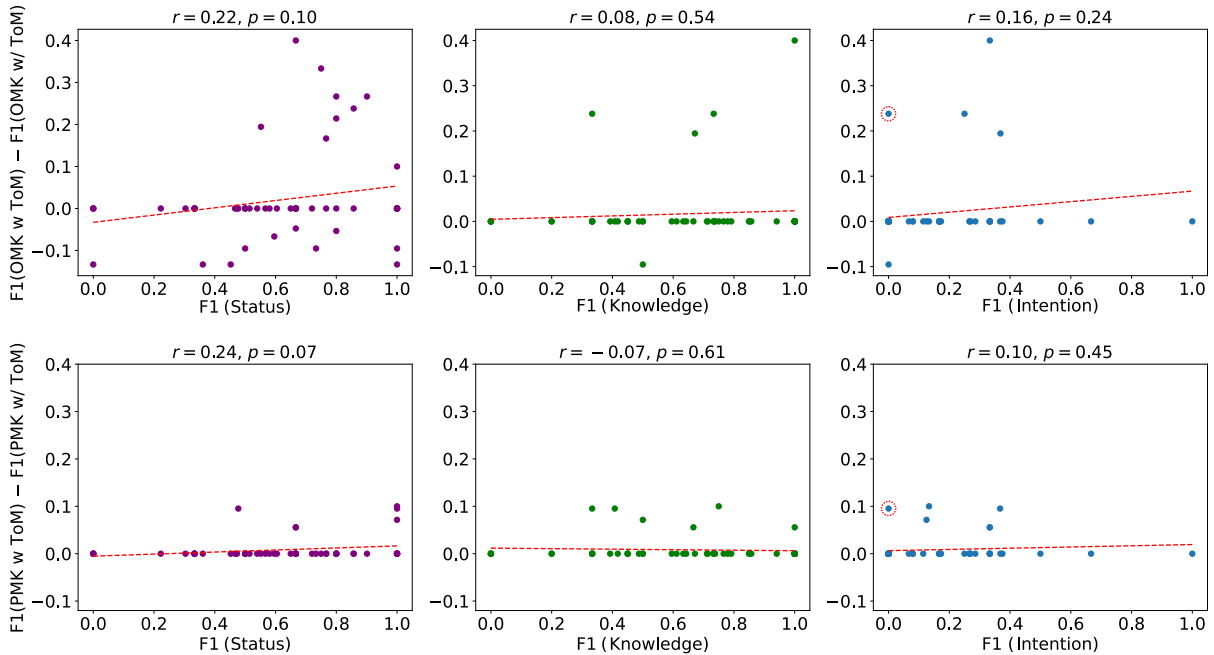


Figure 4: Correlation between F1 scores in the ToM tasks and the difference between F1 scores on OMK (**top**) and PMK (**bottom**) obtained by our model trained with and without ToM features as input. The red dotted line represents a linear fit. The red dotted circle indicates game sessions for which performance on CPA improved even if the F1 score on ToM task is zero.

tures truly benefit the model, we anticipate a strong positive correlation between the performance on the individual task ToM tasks and that of CPA. To validate this, we chose the best-performing model for each ToM feature in CPA and calculated the performance difference relative to the model without any ToM features (refer to Table 2). Next, we calculated the Pearson correlation coefficient to measure the relationship between this difference in performance and the F1 score on ToM tasks. Results for both OMK and PMK are reported in Figure 4. Given that  $r < 0.3$  and  $p \gg 0.05$  in all cases, we can conclude that *there is no correlation between the performance on ToM tasks and improvements on CPA*. It is worth noting that for some cases, CPA improved even if the F1 score on ToM was zero (highlighted with red dotted circles in Figure 4).

### 5.3.3 Incorporating ToM Ground-Truth

In our final experiment, we aim to assess the utility of mental state information in CPA by answering the question: *To what extent do models gain from including ground-truth information about mental states?* We replicate our experiments from Section 5.2 to explore this. However, instead of the learnt ToM features, we feed models with a one-hot encoding of the corresponding ToM question-answer pair. This encoding process follows the

same methodology employed by Bara et al. (2021). As can be seen from a comparison of Table 4 and Table 2, the baseline and our model trained with ground-truth mental state information consistently under-perform those trained with learnt ToM features on OMK and PMK, respectively. The remaining results are generally on par. Furthermore, upon comparing the initial row in Table 4 with the subsequent rows, we note that incorporating ToM ground-truth yields similar scores to those achieved without, except for the baseline in PMK. This underscores a significant limitation of the ToM task representation: *the collected ground-truth mental states are not beneficial for CPA*. This finding, in conjunction with our diagnostic probing analysis, suggests that models trained to infer mental states may be learning information more closely associated with other correlations in the data, rather than representing the mental states.

## 6 Limits and Future Directions for Neural Theory of Mind

A key insight of our work is that current approaches for CPA, rather than learning ToM, seem to merely exploit latent correlations in the data that have little to do with mental states. This is highlighted by the lack of impact of the proposed ToM features on

ToM Labels			OMK		PMK	
<i>Status</i>	<i>Knowledge</i>	<i>Intention</i>	Bara et al. (2023)	Ours	Bara et al. (2023)	Ours
			26.3 ± 1.9	58.2 ± 0.3	60.9 ± 3.2	51.5 ± 4.7
✓			26.8 ± 1.6	58.5 ± 0.6	66.0 ± 1.9	51.5 ± 4.7
	✓		26.8 ± 1.6	58.3 ± 0.2	66.0 ± 1.9	51.5 ± 4.7
		✓	26.8 ± 1.6	58.2 ± 0.3	66.0 ± 1.9	52.2 ± 3.4
✓	✓		26.6 ± 1.2	58.3 ± 0.2	66.0 ± 1.9	51.5 ± 4.7
✓		✓	27.0 ± 1.4	58.4 ± 0.2	66.0 ± 1.9	51.5 ± 4.7
	✓	✓	26.9 ± 1.6	58.6 ± 0.5	66.0 ± 1.9	51.0 ± 4.2
✓	✓	✓	26.6 ± 1.1	58.5 ± 1.3	66.0 ± 1.9	51.5 ± 4.7

Table 4: Performance comparison on CPA when training with ground-truth ToM labels. We report F1 scores for own (OMK) and partner (PMK) missing knowledge prediction.

CPA, as shown in Table 2 and Table 4. This finding is surprising and worrisome at the same time and calls for a fundamental re-assessment of how to equip computational agents with ToM capabilities and how to evaluate them. Despite research on this topic still being in its infancy, the problem of correctly learning neural ToM has recently been put more and more under scrutiny (Sap et al., 2022; Aru et al., 2023). Our results underline in a directly observable way that the acquisition of comprehensive ToM capabilities cannot be reduced to merely passing a specific, narrow set of tasks. The main rationale for this conclusion is that we still do not have a task for which possessing ToM capabilities is both a *necessary and sufficient* prerequisite for its resolution. Current ToM benchmarks rely on tasks that seem to *intuitively* require ToM to be solved. However, these tasks can often be solved by just exploiting shortcuts within the data (Le et al., 2019; Aru et al., 2023; Bortoletto et al., 2024). **As a result, we posit that directly optimising an agent or system for ToM may not represent an effective approach for progress.**

Instead, recent work proposed the use of open-ended environments to study ToM with the aim of observing whether these capabilities emerge through interactions with other agents (Aru et al., 2023). Minecraft represents a good candidate environment for multi-agent collaboration in an open world. However, the way Bara et al. (2021, 2023) frame *MindCraft* is still limited to specific tasks and requires extensive data collection efforts. One possible solution could be to transform *MindCraft* into a reinforcement learning environment with a focus on less constrained collaborative tasks. While Bara et al. (2021, 2023) suggest modelling ToM

as a supervised learning task, the way humans acquire ToM is more nuanced and largely unsupervised (Ruffman, 2023). **We believe that the development of open-ended environments combined with learning ToM capabilities in an unsupervised, human-like manner is a more promising direction for future research.**

ToM capabilities are deeply linked to language acquisition (Tomasello, 2005). In the context of dialogue-based collaboration, another interesting future direction could be to learn ToM from generation instead of classification (Liu et al., 2023). Current approaches could be further improved by building a more general and robust world model, e.g., by leveraging a pre-trained language or video-language model as a more general prior. Finally, in addition to developing suitable environments and learning algorithms, effective and interpretable methods to evaluate whether agents have truly learned ToM will be crucial. We see three exciting directions in this regard: probing (Niven and Kao, 2019), mechanistic interpretability (Wang et al., 2023), and concept learning (Oguntola et al., 2021; Chen et al., 2020). The work of Oguntola et al. serves as an inspiring example, where agents learn human-interpretable concepts that represent beliefs about other agents in a simple multi-agent reinforcement learning setting.

## 7 Conclusion

In this work, we demonstrated that applying task-specific constraints to plan graphs reduces significantly the performance gap between predicting OMK and PMK in *MindCraft*. At the same time, improvements from ToM modelling diminish, raising concerns about current approaches. Our ex-



periments and analyses consistently suggest that current ToM modelling approaches learn features that likely reflect latent patterns in the data, with no perceivable link to ToM. This finding calls for a deeper understanding of the role of ToM in CPA and beyond, as well as for new methods to model and evaluate mental states in collaborative agents.

## 8 Limitations

We identify two main limitations of our work. First, our strategy of selecting candidate edges for own missing knowledge prediction is specific to the structure of the task as presented by [Bara et al. \(2021\)](#). While we recognise the task-specific nature of our strategy, it is crucial to note that [Bara et al. \(2023\)](#) also leverages task-specific constraints by assuming that the partner’s missing knowledge is present in one’s own. However, our approach can still be used in other settings without the assumption that one’s own missing knowledge covers that of the partner. Crucially, our approach does not challenge or undermine the fundamental conclusions drawn about modelling and evaluating ToM. Neither does it serve as the cause for the diminishing improvements on CPA observed when including ToM features.

Second, our current analysis is limited to the *MindCraft* dataset. To the best of our knowledge, it is the only environment that studies the role of ToM in CPA making it the natural candidate for our analysis. However, our work lays the basis of a systematic study of ToM in CPA in general and can also inform future work targeting new environments or datasets.

## 9 Ethical Impact

Our work is foundational, far away from particular applications or any potential societal impact. However, it is important to keep in mind that claims about modelling and predicting mental states potentially have huge ethical impact. Caution is imperative when dealing with sensitive aspects of individuals’ inner experiences and emotions. Mishandling such information could lead to privacy breaches, potential stigmatisation, or the misuse of personal data. Additionally, there is a risk of reinforcing biases or misinterpreting complex psychological nuances, which may have unintended consequences on individuals’ well-being. Lastly, resonating with our findings, the use of models that predict mental states by merely exploiting heuris-

tics and spurious patterns in the data rather than genuinely modelling Theory of Mind introduces significant ethical challenges. Therefore, ethical considerations and responsible practices are crucial to ensure a respectful and appropriate use of technology in this domain.

## 10 Acknowledgements

M. Bortoletto and A. Bulling were funded by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme under grant agreement No 801708. L. Shi was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy – EXC 2075 – 390740016. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting C. Ruhdorfer. The authors would like to especially thank Hsiu-Yu Yang, Manuel Mager, Pavel Denisov, Ekta Sood, and Anna Penzkofer for their support and the numerous insightful discussions.

## References

- Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. [Fine-grained analysis of sentence embeddings using auxiliary prediction tasks](#). In *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France.
- Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations*.
- James F. Allen, Nate Blaylock, and George Ferguson. 2002. [A problem solving model for collaborative agents](#). In *Proceedings of the 1st International Joint Conference on Autonomous Agents & Multiagent Systems*, pages 774–781, Bologna, Italy. Association for Computing Machinery.
- Jaen Aru, Aqeel Labash, Oriol Corcoll, and Raul Vicente. 2023. [Mind the gap: challenges of deep learning approaches to theory of mind](#). *Artif. Intell. Rev.*, 56(9):9141–9156.
- Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. 2009. [Action understanding as inverse planning](#). *Cognition*, 113(3):329–349.
- Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. [MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

- Cristian-Paul Bara, Ziqiao Ma, Yingzhuo Yu, Julie Shah, and Joyce Chai. 2023. [Towards collaborative plan acquisition through theory of mind modeling in situated dialogue](#). In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, pages 2958–2966, Macao, SAR, China.
- Dan Bohus and Alexander I. Rudnicky. 2009. [The ravenclaw dialog management framework: Architecture and systems](#). *Comput. Speech Lang.*, 23(3):332–361.
- Matteo Bortoletto, Lei Shi, and Andreas Bulling. 2024. [Neural Reasoning About Agents’ Goals, Preferences, and Actions](#). In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*.
- Shaked Brody, Uri Alon, and Eran Yahav. 2022. [How attentive are graph attention networks?](#) In *Proceedings of the 10th International Conference on Learning Representations*, Virtual Event.
- Moritz C. Buehler and Thomas H. Weisswange. 2020. [Theory of mind based communication for human agent cooperation](#). In *Proceedings of the 1st IEEE International Conference on Human-Machine Systems*, pages 1–6, Rome, Italy. IEEE.
- Zhi Chen, Yijie Bei, and Cynthia Rudin. 2020. [Concept whitening for interpretable image recognition](#). *Nature Machine Intelligence*, 2(12):772–782.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. [Learning phrase representations using RNN encoder–decoder for statistical machine translation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. [What you can cram into a single \\$&!#\\* vector: Probing sentence embeddings for linguistic properties](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2023. [A survey of natural language generation](#). *ACM Comput. Surv.*, 55(8):173:1–173:38.
- Prashant Doshi, Xia Qu, Adam Goodie, and Diana L. Young. 2010. [Modeling recursive reasoning by humans using empirically informed interactive pomdps](#). In *9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, volume 1–3, pages 1223–1230, Toronto, Canada. IFAA-MAS.
- Jiafei Duan, Samson Yu, Nicholas Tan, Li Yi, and Cheston Tan. 2022. [Boss: A benchmark for human belief prediction in object-context scenarios](#). *arXiv preprint arXiv:2206.10665*.
- Lifeng Fan, Shuwen Qiu, Zilong Zheng, Tao Gao, Song-Chun Zhu, and Yixin Zhu. 2021. [Learning triadic belief dynamics in nonverbal communication from videos](#). In *Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition*, pages 7312–7321, Virtual. Computer Vision Foundation / IEEE.
- Kanishk Gandhi, Gala Stojnic, Brenden M. Lake, and Moira R. Dillon. 2021. [Baby intuitions benchmark \(BIB\): discerning the goals, preferences, and actions of others](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 9963–9976, virtual. Curran Associates, Inc.
- Erin Grant, Aida Nematzadeh, and Thomas L. Griffiths. 2017. [How can memory-augmented neural networks pass a false-belief task?](#) In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, London, UK.
- Yanlin Han and Piotr J. Gmytrasiewicz. 2018. [Learning others’ intentional models in multi-agent settings using interactive pomdps](#). In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, pages 5639–5647, Montréal, Canada. Curran Associates, Inc.
- Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. [Array programming with NumPy](#). *Nature*, 585(7825):357–362.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- J. D. Hunter. 2007. [Matplotlib: A 2d graphics environment](#). *Computing in Science & Engineering*, 9(3):90–95.
- Dieuwke Hupkes, Sara Veldhoen, and Willem H. Zuidema. 2018. [Visualisation and ‘diagnostic classifiers’ reveal how recurrent and recursive neural networks process hierarchical structure](#). *J. Artif. Intell. Res.*, 61:907–926.

- Prashant Jayannavar, Anjali Narayan-Chen, and Julia Hockenmaier. 2020. [Learning to execute instructions in a Minecraft dialogue](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2589–2602, Online. Association for Computational Linguistics.
- Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2019. [CoDraw: Collaborative drawing as a testbed for grounded goal-driven communication](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513, Florence, Italy. Association for Computational Linguistics.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA.
- Staffan Larsson and David R. Traum. 2000. [Information state and dialogue management in the trindi dialogue move engine toolkit](#). *Nat. Lang. Eng.*, 6(3–4):323–340.
- Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. [Revisiting the evaluation of theory of mind through question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.
- Jin Joo Lee, Fei Sha, and Cynthia Breazeal. 2019. [A bayesian theory of mind approach to nonverbal communication](#). In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, pages 487–496, Daegu, South Korea. IEEE.
- David Liben-Nowell and Jon M. Kleinberg. 2007. [The link-prediction problem for social networks](#). *J. Assoc. Inf. Sci. Technol.*, 58(7):1019–1031.
- Andy Liu, Hao Zhu, Emmy Liu, Yonatan Bisk, and Graham Neubig. 2023. [Computational language acquisition with theory of mind](#). In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda.
- Wes McKinney. 2010. [Data structures for statistical computing in python](#). In *Proceedings of the 9th Python in Science Conference 2010 (SciPy 2010)*, pages 56–61, Austin, Texas. scipy.org.
- Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. [Collaborative dialogue in Minecraft](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy. Association for Computational Linguistics.
- Timothy Niven and Hung-Yu Kao. 2019. [Probing neural network comprehension of natural language arguments](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.
- Ini Oguntola, Joseph Campbell, Simon Stepputtis, and Katia P. Sycara. [Theory of mind as intrinsic motivation for multi-agent reinforcement learning](#). *arXiv preprint arXiv:2307.01158*.
- Ini Oguntola, Dana Hughes, and Katia P. Sycara. 2021. [Deep interpretable models of theory of mind](#). In *Proceedings of the 30th IEEE International Conference on Robot & Human Interactive Communication*, pages 657–664, Vancouver, BC, Canada. IEEE.
- The pandas development team. 2020. [pandas-dev/pandas: Pandas](#).
- Ioannis Papaioannou, Christian Dondrup, and Oliver Lemon. 2018. [Human-robot interaction requires more than slot filling - multi-threaded dialogue for collaborative tasks and social conversation](#). In *FAIM/ISCA Workshop on Artificial Intelligence for Multimodal Human Robot Interaction*, pages 61–64, Stockholm, Sweden. ISCA.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [Pytorch: An imperative style, high-performance deep learning library](#). In *Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035, Vancouver, BC, Canada. Curran Associates, Inc.
- David Premack and Guy Woodruff. 1978. [Does the chimpanzee have a theory of mind?](#) *Behavioral and Brain Sciences*, 1(4):515–526.
- Liang Qiu, Yizhou Zhao, Yuan Liang, Pan Lu, Weiyang Shi, Zhou Yu, and Song-Chun Zhu. 2022. [Towards socially intelligent agents with mental state transition and human value](#). In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 146–158, Edinburgh, UK. Association for Computational Linguistics.
- Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew M. Botvinick. 2018. [Machine theory of mind](#). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 4215–4224, Stockholm, Sweden. PMLR.
- Charles Rich, Candace L. Sidner, and Neal Lesh. 2001. [COLLAGEN: applying collaborative discourse theory to human-computer interaction](#). *AI Mag.*, 22(4):15–26.
- Ted Ruffman. 2023. [Belief it or not: How children construct a theory of mind](#). *Child Development Perspectives*, 17(2):106–112.

- Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. [Neural theory-of-mind? on the limits of social intelligence in large LMs](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. [Modeling relational data with graph convolutional networks](#). In *The Semantic Web - 15th International Conference*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607, Heraklion, Crete, Greece. Springer.
- Jürgen Streeck, Charles Goodwin, and Curtis LeBaron. 2011. *Embodied interaction: Language and body in the material world*. Cambridge University Press.
- Michael Tomasello. 2005. *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.
- David Traum. 2017. *Computational Approaches to Dialogue*. Routledge.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, Long Beach, CA, USA. Curran Associates, Inc.
- Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. [SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python](#). *Nature Methods*, 17:261–272.
- Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. [Interpretability in the wild: a circuit for indirect object identification in GPT-2 small](#). In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda.
- Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. [Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant](#). In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.
- Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. [Semantically conditioned LSTM-based natural language generation for spoken dialogue systems](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.
- Jessica Williams, Stephen M. Fiore, and Florian Jentsch. 2022. [Supporting artificial social intelligence with theory of mind](#). *Frontiers in Artificial Intelligence*, 5.
- Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. [Embedding entities and relations for learning and inference in knowledge bases](#). In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA.
- Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. 2023. [I cast detect thoughts: Learning to converse and guide with intents and theory-of-mind in dungeons and dragons](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11136–11155, Toronto, Canada. Association for Computational Linguistics.

## A Appendix

### A.1 MindCraft

MindCraft (Bara et al., 2021) is an interactive environment based on Minecraft in which the objective of each game is to create a randomly generated goal material. Some starting materials are already present in the environment, and the others are created by the players by:

- *Mining*: hit a specific block with a tool to create a new block. This allows to generate infinite new blocks of a certain type and move them around.
- *Combining*: fuse two materials to obtain a new one.

Players are free to navigate in the environment, move blocks around and chat. Being the field of view first-person, players have partial observability of the environment.

Players possess *asymmetric knowledge and skill sets*, where each player is provided with a partial plan (such as a partial knowledge graph or recipe) along with a specific tool to interact with certain blocks. Their knowledge and skills are complementary: Player 1’s plan contains the information missing in Player 2’s plan, and vice versa. Likewise, Player 1 can interact with blocks that Player 2 cannot, and vice versa. This inherent asymmetry encourages communication between players and the need to reason about each other’s mental states.

Bara et al. (2021) introduced three ToM question answering tasks that are specifically designed to capture mental state information that is pertinent to collaboration. Players are presented with three questions, each paired by type: if one player is asked about their partner’s beliefs, the other player is presented with the same question regarding their own beliefs (with respect to the same question):

1. *Task Status*: Predict if a specific material has been created by one of the two players. For example, if Player 1 is asked: Has the other player made GOLD\_BLOCK until now?, then Player 2 is asked: Have you created GOLD\_BLOCK until now?. Possible answers are YES, NO, or MAYBE. This task probes players’ belief of the game state.
2. *Player Knowledge*: Predict whether a player knows how to create a specific material or if they believe their partner knows. For example,

if Player 1 is asked: Do you think the other player knows how to make BLUE\_WOOL?, then Player 2 is asked: Do you know how to make BLUE\_WOOL?. Possible answers are YES, NO, or MAYBE. This task probes players’ belief of their and their partner’s current knowledge.

3. *Player Intention*: Predict which material a player is making at the current time step. For example, if Player 1 is asked: What do you think the other player is making right now?, then Player 2 is asked: What are you making right now?. Possible answers are the different types of block in the game or NOT\_SURE.

Mental state annotations are collected using periodic pop-ups that interrupt the game every 75 seconds.

Computational models are trained and evaluated by predicting the answer to the ToM tasks from a player’s perspective, given a history of observations, the chat dialogue, the perceived actions in the shared environment, and the partial plan. Models are trained to minimise the cross entropy loss. They output  $c$  logits, with  $c$  being the number of possible classes: three for *Status* and *Knowledge*, and 22 for *Intention* (21 possible materials + NOT\_SURE). Models’ final prediction is obtained by taking the argmax of the logits.

#### A.1.1 Data Modalities

The *MindCraft* dataset comprises various modalities, including first-person video streams for each player, chat dialogue and corresponding dialogue moves, and players’ plans. Bara et al. (2021) established a timestep of  $\Delta t = 1$  second, ensuring that each timestep has an associated video frame. However, not all timesteps include dialogue utterances or questions, as players do not necessarily exchange messages every second and questions pop-up every 75 seconds. The players’ plan is static and known from the beginning of the game. The models generate predictions at timestep  $t$ , coinciding with when questions are posed to the players. These predictions rely on the player’s plan, video stream, and dialogue history (if available) up to time  $t$ .

#### A.1.2 Dataset Statistics

Bara et al. (2021) report that the original MindCraft dataset includes 100 games, with an average of 20.5 dialogue exchanges per game, for a total of 2091 exchanges. Games last between 1 minute and

22 seconds to 27 minutes and 26 seconds, with the average game lasting 7 minutes and 23 seconds. A total of 12 hours, 18 minutes, and 33 seconds of in-game interaction was recorded. Between 5 and 10 objects are used in a game, and between 7 and 11 steps are necessary in each game to achieve the goal. The dataset is randomly partitioned into 60% for training, 20% for validation, and 20% for testing, with the condition to keep similar distributions of game lengths.

## A.2 Collaborative Plan Acquisition

Bara et al. (2023) extended MindCraft by collecting 60 additional game sessions and defined an additional task: *collaborative plan acquisition* (CPA). In CPA a model has to predict, from a player’s perspective, its own missing knowledge (OMK) and the other player’s missing knowledge (PMK). CPA was introduced to explore the role of ToM modelling in predicting players’ missing knowledge while executing the crafting tasks. CPA is formulated as follows:

**Definition 1** Consider a joint plan as a directed AND-graph  $\mathcal{P} = (V, E)$ , where the nodes  $V$  denote (sub-)goal materials, and edges  $E$  denote temporal constraints between the sub-goals. In a collaborative plan acquisition problem, two agents  $i$  and  $j$  start with partial plans  $\mathcal{P}_i = (V, E_i)$ ,  $E_i \subseteq E$ , and  $\mathcal{P}_j = (V, E_j)$ ,  $E_j \subseteq E$ . Given a sequence of visual observations  $O_i^t$  and a joint dialogue history  $D^t$  at time  $t$ , agent  $i$  has to infer their own missing knowledge  $\bar{E}_i = E \setminus E_i$  and the partner  $j$ ’s missing knowledge  $\bar{E}_j = E \setminus E_j$ .

Predictions for OMK and PMK are made at  $t = T$ , where  $T$  is the final timestep of the game.

While intrinsically linked, ToM and CPA tasks exhibit a fundamental distinction: ToM tasks directly explore players’ beliefs regarding the game state and their partner’s mental states, whereas CPA tasks involve predicting the absent information from players’ partial plans. In ToM tasks, the ground truth is based on players’ beliefs, which may be true or false, whereas in CPA tasks, the ground truth is formally determined by  $\bar{E}_i = E \setminus E_i$  for OMK and  $\bar{E}_j = E \setminus E_j$  for PMK. Bara et al. (2023) discuss the partial overlap between *Task Knowledge* and PMK, highlighting their difference: *Task Knowledge* probes whether a *single* piece of knowledge is known by the partner, while PMK involves predicting whether the partner shares *each* piece of the player’s knowledge.

### A.2.1 Data Modalities

CPA is conducted in the MindCraft environment, therefore the data modalities utilised mirror those in the ToM tasks: first-person video streams for each player, chat dialogue along with corresponding dialogue moves, and the players’ plans. In addition to the aforementioned modalities, models for CPA receive ToM features extracted from the sequence-to-sequence model trained on the ToM tasks discussed in §A.1. These ToM features are tensor representations of dimension 1024 that are incorporated into the model’s input. A timestep of  $\Delta t = 1$  second is maintained. In contrast to the ToM tasks, CPA models generate predictions at timestep  $t = T$ , i.e., at the end of each game.

### A.2.2 Formalising ToM Tasks

Based on the formalism introduced for CPA, we can formalise the ToM tasks as follows:

1. *Task Status*: Predict if a specific material  $V_k \in V$  has been created by one of the two players.
2. *Player Knowledge*: Predict whether a player knows how to create a specific material  $V_k$ , i.e.,

$$\{e_{V_k}^n\}_{n>1} \stackrel{?}{\in} E_i$$

or if they believe their partner knows, i.e.

$$\{e_{V_k}^n\}_{n>1} \stackrel{?}{\in} E_j,$$

with  $e_{V_k} \in E$  being an edge with end-node  $V_k$  and  $n$  being the number of materials needed to craft  $V_k$ .

3. *Player Intention*: Predict which material  $V_k \in V$  a player is making at the current time step  $t$ .

## A.3 Technical Details

### A.3.1 GNN Plan Encoder

We propose a modification to the method by Bara et al. (2023) that includes representing plans as graph objects and using a GNN-based encoder-decoder together with candidate sampling to predict missing knowledge, i.e., missing edges (see Figure 2).

In the encoding phase, given a graph, we compute the node embeddings using GATv2 convolutions (Brody et al., 2022). The node features (one-hot encoding of materials) are first projected using a single linear layer with hidden dimension 128, followed by GELU activation and dropout. The same procedure is applied to edge features

<i>Status</i>			
Modalities	Bara et al. (2023)	Ours	Human
M	56.0 ± 0.8	59.9 ± 0.7	67.0
D+M	54.6 ± 1.1	59.1 ± 0.6	67.0
D+V+M	59.3 ± 1.0	58.9 ± 0.8	67.0
V+M	59.3 ± 1.7	59.6 ± 0.4	67.0
<i>Knowledge</i>			
Modalities	Bara et al. (2023)	Ours	Human
M	54.7 ± 2.5	57.9 ± 0.2	58.0
D+M	56.2 ± 1.9	57.2 ± 1.5	58.0
D+V+M	57.6 ± 1.0	57.5 ± 1.7	58.0
V+M	56.4 ± 2.5	57.6 ± 1.8	58.0
<i>Intention</i>			
Modalities	Bara et al. (2023)	Ours	Human
M	14.9 ± 0.2	11.7 ± 2.2	46.0
D+M	12.1 ± 1.0	11.1 ± 1.8	46.0
D+V+M	13.5 ± 0.6	12.1 ± 2.4	46.0
V+M	13.8 ± 1.7	13.4 ± 1.9	46.0

Table 5: Performance comparison on the three ToM tasks using different combinations of modalities: dialogue moves (M), dialogue (D), and video frames (V). F1 scores for the baseline are reported from the original paper (Bara et al., 2023) rather than from our execution of the official code.

(one-hot encoding of tools). Then the first GATv2 convolution is applied, which has hidden dimension 128, four heads, GELU activation and dropout. The final node embeddings are generated by a second GATv2 convolution with output dimension 128 and one head.

In the decoding phase, we evaluate potential missing edges by scoring them against the relevant node embeddings and the context vector. This involves combining the two node embeddings associated with the edge and the context vector, then passing the resulting concatenation through a linear layer with an output dimension of one. The output logits are subsequently fed into a sigmoid function with a threshold of 0.5 to determine whether the edge exists or not.

### A.3.2 Transformer

We use a single-block Transformer (Vaswani et al., 2017) with output dimension 1024 as sequence-to-sequence model. The input consists of concatenated features from various modalities, such as video, dialogue, plan graph, and dialogue moves (and ToM features in the case of CPA), as depicted in Figure 2. Our Transformer incorporates positional encoding (Vaswani et al., 2017) and utilises

a causal attention mask to ensure that each token attends only to previous tokens during self-attention computation. We utilise eight attention heads to compute attention scores over the concatenated input features, including ToM features for CPA tasks. This ensures that the model attends to ToM features. Following the baseline models (Bara et al., 2021, 2023), our models utilise zero-padding when an input modality is absent.

### A.3.3 Training

Models are trained using PyTorch (Paszke et al., 2019) with 1, 42 and 123 as random seeds. All models were trained on a single GPU card, taking approximately 60 minutes for the baselines (17,946,302 parameters) and our models (9,222,100 parameters) on ToM. For CPA, training takes approximately 60 minutes for the baselines (33,698,691 parameters) and 20 minutes for our models (13,364,641 parameters). For the baselines (Bara et al., 2023), we used default parameters reported in the code. For our models, we used the Adam optimiser (Kingma and Ba, 2015) with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.99$ ,  $\epsilon = 10^{-8}$ , and a learning rate of  $\eta = 1 \cdot 10^{-4}$ . We did not perform any exhaustive hyper-parameter tuning but just tried a set of reasonable values:  $\{1 \cdot 10^{-5}, 5 \cdot 10^{-4}, 1 \cdot 10^{-4}, 5 \cdot 10^{-4}\}$ .

### A.4 Comparison to Bara et al. (2023)

Although we trained the baseline of (Bara et al., 2023) using the official code<sup>3</sup> with default hyper-parameters, its performance slightly deviated from the original paper. We contacted the authors asking for clarifications and details that are not documented in the paper/code. Discussions with them did not yield a clear answer, and they were unable to provide their model files. Therefore, in Table 1 and Table 2 we provided results obtained from our runs, reproducible by using our code provided as supplementary material. In the interest of completeness, we include a comparison between our results and those reported by Bara et al. (2023) in Table 5 for ToM tasks and in Table 6 for CPA tasks. On the ToM tasks, the scores achieved by our model are generally on par with those reported by Bara et al. (2021). On CPA, our model still outperforms the baseline of Bara et al. (2023) that, based on the originally reported scores, on average performs slightly worse in some tasks (see Table 6, Overall).

<sup>3</sup><https://github.com/sled-group/collab-plan-acquisition>

ToM Features			Overall		OMK		PMK	
Status	Knowledge	Intention	Bara et al. (2023)	Ours	Bara et al. (2023)	Ours	Bara et al. (2023)	Ours
			44.1 ± 0.6	<b>56.9</b> ± 0.6	16.7 ± 0.1	57.6 ± 0.8	71.4 ± 1.0	56.2 ± 0.3
✓			45.9 ± 1.5	<b>57.3</b> ± 0.6	20.4 ± 1.4	58.0 ± 0.8	71.3 ± 1.6	56.5 ± 0.3
	✓		47.2 ± 1.1	<b>57.0</b> ± 1.4	20.1 ± 1.4	58.4 ± 0.5	74.3 ± 0.7	55.5 ± 1.9
		✓	47.4 ± 1.4	<b>57.2</b> ± 0.5	19.8 ± 1.7	57.9 ± 0.7	75.0 ± 1.0	56.5 ± 0.3
✓	✓		47.0 ± 1.4	<b>56.6</b> ± 1.4	20.9 ± 1.2	57.7 ± 0.5	73.1 ± 1.5	55.5 ± 1.9
✓		✓	45.9 ± 1.2	<b>57.5</b> ± 0.6	19.8 ± 0.8	58.4 ± 0.8	71.9 ± 1.5	56.5 ± 0.3
	✓	✓	46.9 ± 1.5	<b>57.5</b> ± 0.6	20.3 ± 1.8	58.5 ± 0.8	73.4 ± 1.2	56.4 ± 0.1
✓	✓	✓	45.5 ± 0.3	<b>56.7</b> ± 0.7	17.4 ± 0.1	57.1 ± 1.9	73.5 ± 0.5	56.6 ± 0.2

Table 6: Performance comparison on CPA when training with learnt ToM features. F1 scores for the baseline are reported from the original paper (Bara et al., 2023) rather than from our execution of the official code.



Figure 5: Chat from a game in which our ToM model cannot solve the *Player Intention* ToM task. In the same game, integrating the corresponding ToM features into the CPA model enhances the performance on PMK.

It is important to note that our main contributions remain unaffected by this mismatch: First, our improvement on OMK stands, both if we compare our baseline results and those reported in Bara et al. (2023). Notably, our results for the baselines ( $\sim 0.28$ ) are higher than the ones in the original paper ( $\sim 0.21$ ). Second, our analyses are not contingent on absolute results but rather on the relationship with the ToM tasks.

#### A.4.1 Qualitative Example

Figure 5 shows the chat from a game in which our ToM model cannot solve the *Player Intention* ToM task. However, *on the same game*, integrating the corresponding ToM features into the CPA model enhances the performance on the PMK task by approximately 10 points. This particular game

instance is highlighted by the dotted red circle in the rightmost plot of Figure 4. We speculate the ToM model’s struggle with the *Player Intention* task may arise from the initial game part where players’ beliefs are *misaligned*, which could result in a *false belief* (cf. Figure 5). Despite this, the CPA model still benefits from the inclusion of ToM features, suggesting that ToM models may actually be learning information that is more closely associated with other correlations in the data, rather than representing the mental states.

#### A.5 Tools

We performed our data analysis using NumPy (Harris et al., 2020), Pandas (pandas development team, 2020; McKinney, 2010), and SciPy (Virtanen et al., 2020). Figures were made using Matplotlib (Hunter, 2007).

#### A.6 Infrastructure

We ran our experiments on a server running Ubuntu 22.04, equipped with NVIDIA Tesla V100-SXM2 GPUs with 32GB of memory and Intel Xeon Platinum 8260 CPUs.