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

University of Stuttgart, Germany
imprs-is · SimTech · erc European Research Council Established by the European Commission · PUI Perceptual User Interfaces

Matteo Bortoletto, Constantin Ruhdorfer, Adnen Abdessaied, Lei Shi, Andreas Bulling

matteo.bortoletto@vis.uni-stuttgart.de

Motivation

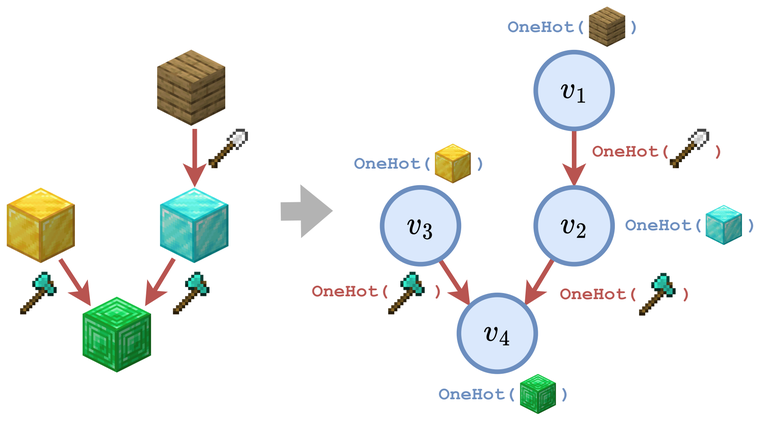
- **Theory of Mind** (ToM) refers to the ability to infer one's own and others' mental states → **crucial for collaboration**.
- It is imperative for AI agents to possess similar capabilities.
- Recent work has introduced **collaborative plan acquisition** (CPA) as a promising task for evaluating collaborative abilities in agents and their relation with Theory of Mind [1,2].
- As including ToM features in the models did non consistently improve performance, the nature of this connection remains unclear.

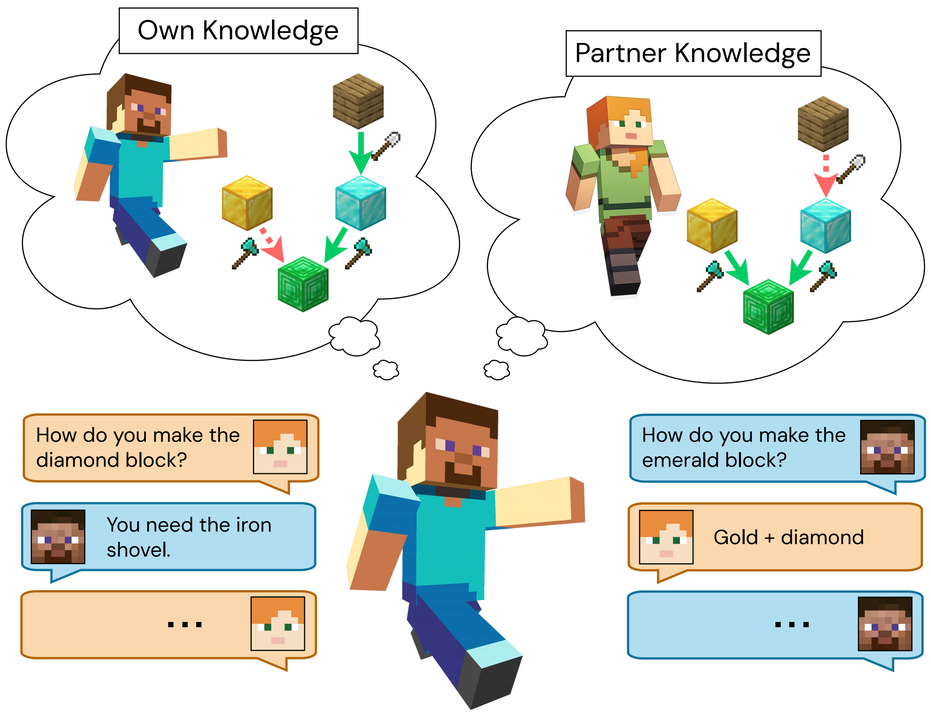
Our Contributions

1. **Graph-based representation** of plans **+ graph learning methods** significantly improves performance.
2. **Principled analyses** that suggest that learnt ToM features reflect latent patterns in the data with no perceivable link to ToM.

MindCraft [1,2]

Two players collaborate to craft a target material. 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.

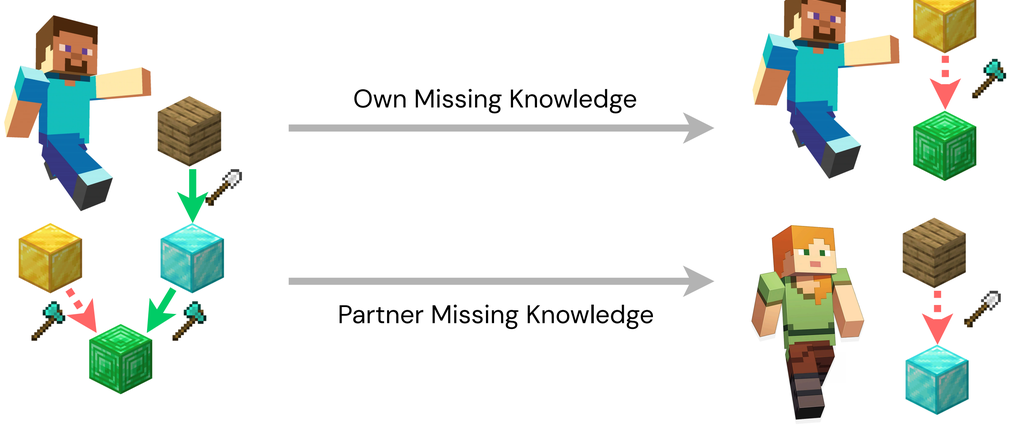




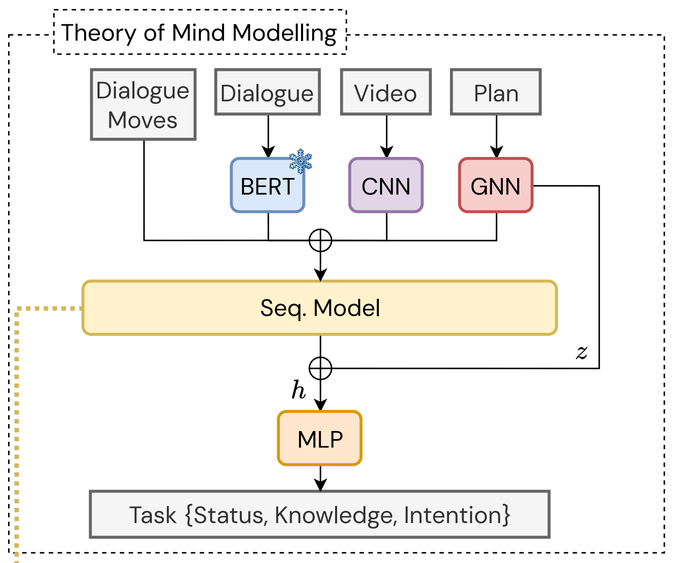
Theory of Mind Tasks

Task Status	Has your partner created GOLD_BLOCK so far?	Have you crafted GOLD_BLOCK yet?
Task Knowledge	Do you think the other player knows how to make BLUE_WOOL?	Do you know how to make BLUE_WOOL?
Task Intention	What do you think the other player is making right now?	What are you making right now?

Collaborative Plan Acquisition (CPA)



Improving ToM and CPA



Status

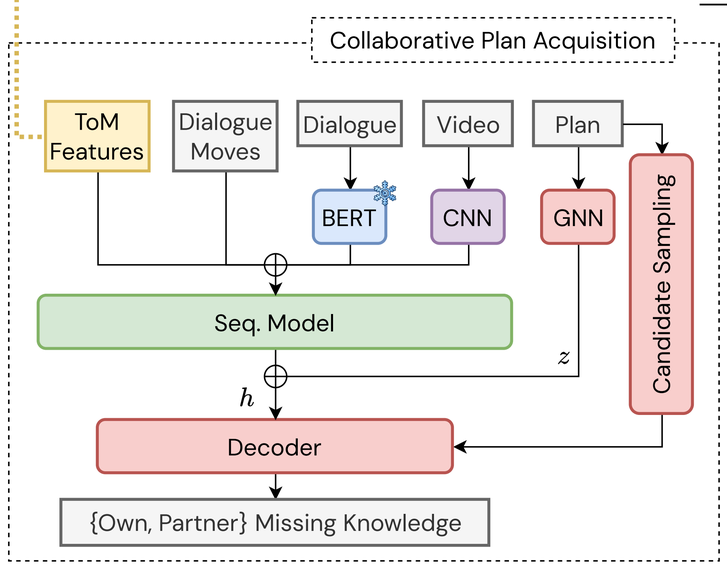
Modalities	Bara et al. (2023)	Ours	Human
M	47.7 ± 0.6	59.9 ± 0.7	67.0
D+M	45.5 ± 2.3	59.1 ± 0.6	67.0
D+V+M	45.2 ± 1.8	58.9 ± 0.8	67.0
V+M	47.3 ± 0.7	59.6 ± 0.4	67.0

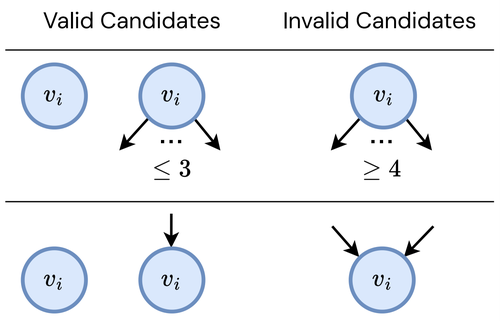
Knowledge

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

Intention

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





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	$\mathbf{56.9} \pm 0.6$	16.7 ± 0.1	57.6 ± 0.8	71.4 ± 1.0	56.2 ± 0.3
✓			45.9 ± 1.5	$\mathbf{57.3} \pm 0.6$	20.4 ± 1.4	58.0 ± 0.8	71.3 ± 1.6	56.5 ± 0.3
	✓		47.2 ± 1.1	$\mathbf{57.0} \pm 1.4$	20.1 ± 1.4	58.4 ± 0.5	74.3 ± 0.7	55.5 ± 1.9
		✓	47.4 ± 1.4	$\mathbf{57.2} \pm 0.5$	19.8 ± 1.7	57.9 ± 0.7	75.0 ± 1.0	56.5 ± 0.3
✓	✓		47.0 ± 1.4	$\mathbf{56.6} \pm 1.4$	20.9 ± 1.2	57.7 ± 0.5	73.1 ± 1.5	55.5 ± 1.9
✓		✓	45.9 ± 1.2	$\mathbf{57.5} \pm 0.6$	19.8 ± 0.8	58.4 ± 0.8	71.9 ± 1.5	56.5 ± 0.3
	✓	✓	46.9 ± 1.5	$\mathbf{57.5} \pm 0.6$	20.3 ± 1.8	58.5 ± 0.8	73.4 ± 1.2	56.4 ± 0.1
✓	✓	✓	45.5 ± 0.3	$\mathbf{56.7} \pm 0.7$	17.4 ± 0.1	57.1 ± 1.9	73.5 ± 0.5	56.6 ± 0.2

Probing for Theory of Mind

1. **No statistical significance** between models with ToM and without ToM features.
2. **Probing experiments**: Features trained on CPA yield similar performance.

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

3. Improvements in CPA tasks do **not correlate** with performance of on ToM tasks.
4. Trained models with **ground-truth mental state information** underperform those trained with learnt ToM features on OMK and PMK.

ToM Labels			OMK		PMK	
Status	*Knowledge*	*Intention*	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

Limits and Future Directions for Neural Theory of Mind

- Directly optimising a system for ToM may not represent an effective approach for progress.
- **Open-ended environments + self/unsupervised learning** is a more promising direction for future research.
- **Generation** instead of classification, leveraging large pre-trained models as prior.
- Work on **interpretability** methods.

ACL 2024 Bangkok, Thailand

[1] Bara, Cristian-Paul, et al. "MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks." *EMNLP* (2021).
[2] Bara, Cristian-Paul, et al. "Towards collaborative plan acquisition through theory of mind modeling in situated dialogue." *IJCAI* (2023).