Extending the Theory of Mind Framework to Embodied Artificial Agents

Author: Aleksandra Maria Jach
a.m.jach@student.tudelft.nl
Supervisors: Chirag Raman, Ojas Shirekar
Affiliations: EEMCS Faculty Delft University of Technology

TU Delft

1.Background

- **Theory of Mind (ToM)** is the ability to attribute mental states, such as beliefs, to oneself and others [1].
- ToM has been studied in the fields of **psychology**, **neurosciences**, and recently in **human-agent interactions (HAI)** [2], [3].
- ToM is a critical component of **social interactions** [1].
- **Embodiment** refers to artificial agents that exist in tangible spaces or digital environments, such as robots or chatbots [4].

Knowledge gap: integration of ToM in environments where multiple *embodied agents* interact *with one another.*

2.Research Question

How has the framework of **Theory of Mind** been incorporated to virtually and physically **embodied agents** with the ability to take **perspectives** of each other's points of view?

Sub-questions

RQ1: In what ways does Theory Of Mind differ when applied to computational agents and when applied to humans?

RQ2: What are the applications of Theory of Mind in multi-agent systems in which agents can take on each others' perspectives?

RQ3: What are the ways of implementing Theory Of Mind for multi-agent environments?

3.Methodology

- This **Systematic Literature Review** was performed by complying with PRISMA guidelines.
- I reviewed 611 records. After abstract and full text screening, **38** papers were **included** in this study (see *Figure 1*).

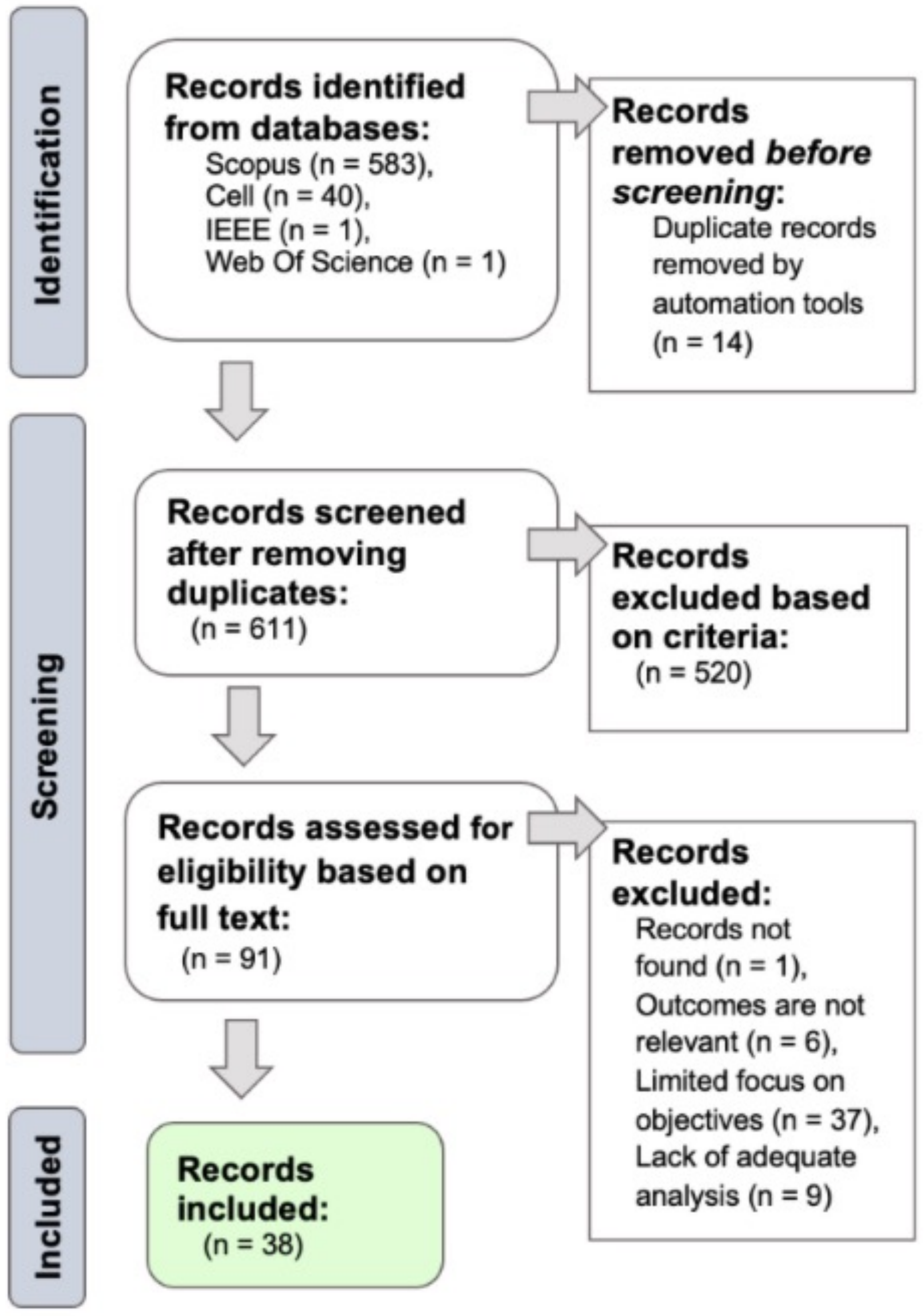
Identification
Records identified from databases:
Scopus (n = 583),
Cell (n = 40),
IEEE (n = 1),
Web Of Science (n = 1)
Records removed *before screening*:
Duplicate records removed by automation tools
(n = 14)
Screening
Records screened after removing duplicates:
(n = 611)
Records excluded based on criteria:
(n = 520)
Records assessed for eligibility based on full text:
(n = 91)
Records excluded:
Records not found (n = 1),
Outcomes are not relevant (n = 6),
Limited focus on objectives (n = 37),
Lack of adequate analysis (n = 9)
Included
Records included:
(n = 38)

Figure 1: PRISMA flow diagram used to demonstrate the process of finding relevant papers for this systematic review.

4.Results

RQ1: Themes regarding outlooks on ToM:

- General vs Task-specific ToM
- Theory Theory vs. Simulation Theory
- Different cognitive processes **overlap** with ToM (see *Figure 2*).
- ToM is a **complex** process, especially when using **higher-order** ToM.
- Humans often rely on simple reactive **strategies** instead of ToM.
- Humans use various methods to reduce their cognitive load when using ToM.

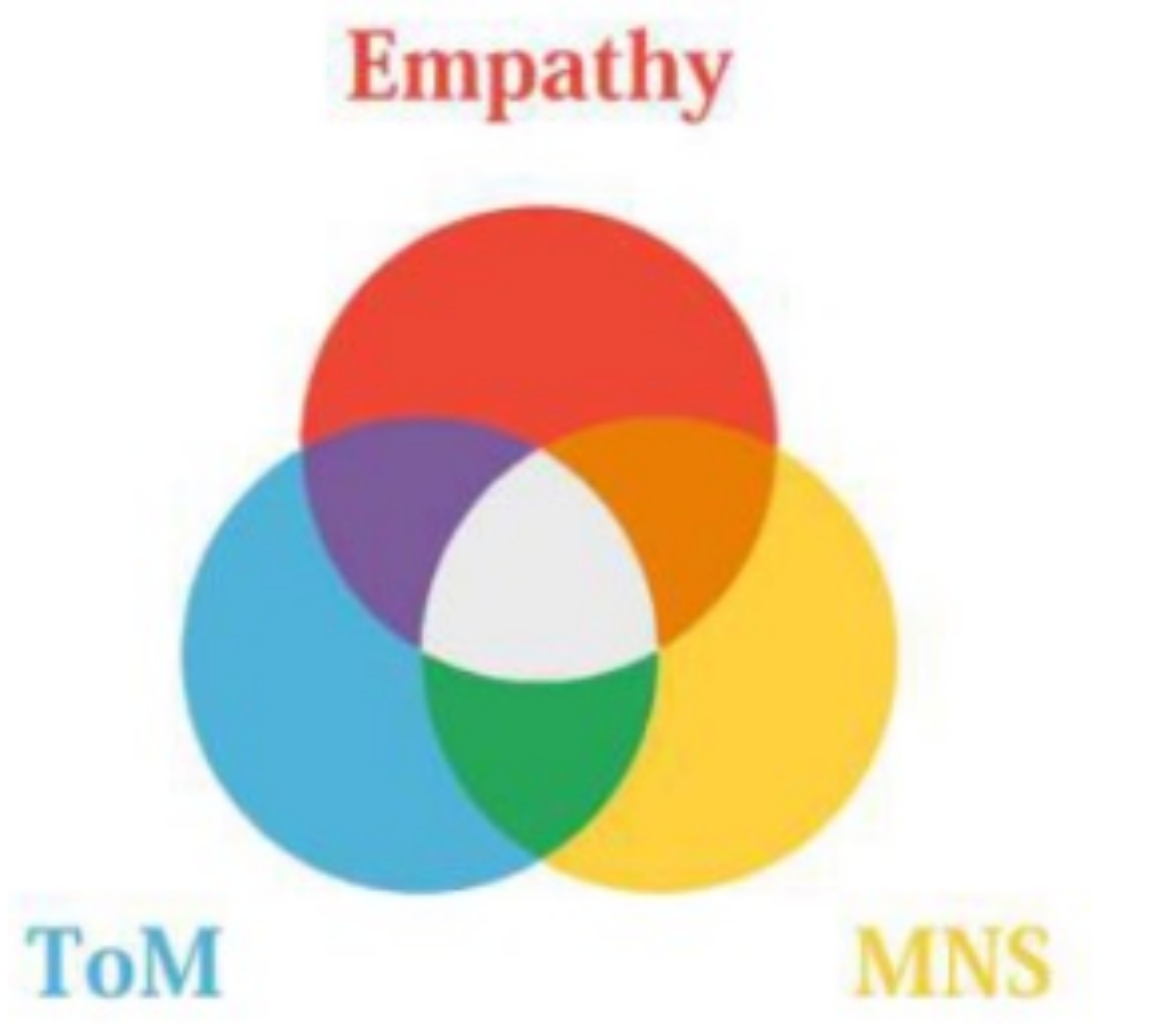

Figure 2: The overlap between Theory of Mind (ToM), empathy, and Mirror Neuron Systems (MNS) is challenging to quantify.
Source: Adapted from [3]

RQ2: Use cases:

- Collaborative, competitive, mixed settings & simulations
- Reducing uncertainty

RQ3: Implementation themes:

- Different methods for implementing ToM (see *Figure 3*)
- **Layered** architecture
- Constrains on beliefs
- Parameter update and adjustment

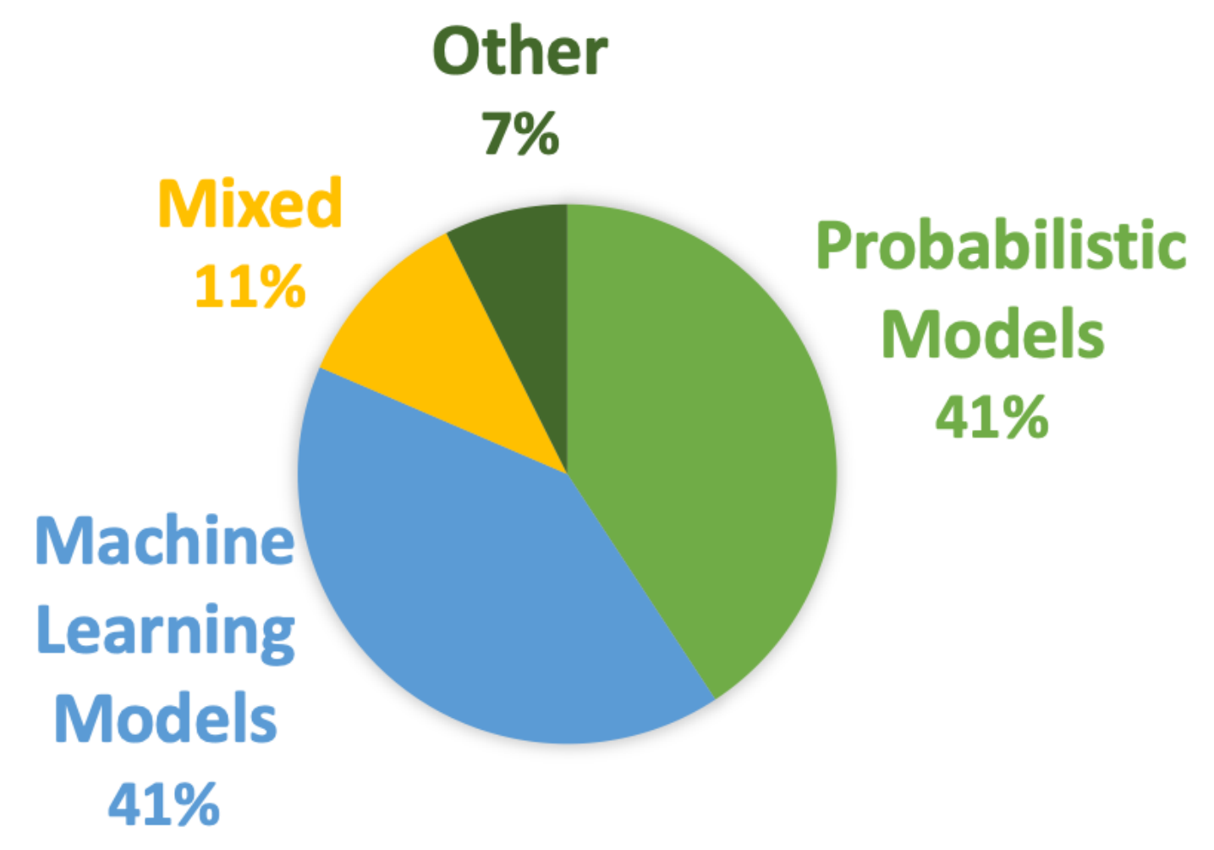

Figure 3: Distribution of ToM implementations in the selected studies.

5.Discussion and Conclusion

- Due to **ambiguity** of terminology in the literature, it is difficult to conceptualize ToM.
- **Integrating** ToM with other cognitive frameworks can improve efficiency and performance of ToM agents

Limitations

- Further research is needed on **probabilistic and machine learning** methods used to implement ToM
- The lack of of a linguistic analysis of word perspective-taking limits the depth of the current research.

References

[1] C. Langley, B. Cirstea, F. Cuzzolin, and B. Sahakian, "Theory of mind and preference learning at the interface of cognitive science, neuroscience, and AI: A review," *Frontiers in Artificial Intelligence*, vol. 5, 2022. DOI: 10.3389/frai.2022.778852
[2] J. Williams, S. Fiore, and F. Jentsch, "Supporting artificial social intelligence with theory of mind," *Frontiers in Artificial Intelligence*, vol. 5, 2022. DOI: 10.3389/frai.2022.750763
[3] D. Alcalá-López, K. Vogeley, F. Binkofski, and D. Bzdok, "Building blocks of social cognition: Mirror, mentalize, share?" *Cortex*, vol. 118, pp. 4-18, 2019. DOI: 10.1016/j.cortex.2018.05.006
[4] B. Lugrin, "Introduction to socially interactive agents," in *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics*, Volume 1: Methods, Behavior, Cognition, 1st ed. New York, NY, USA: Association for Computing Machinery, 2021, pp. 1-20, ISBN: 9781450387200