Dimensions of Theory of Mind Attribution

Exploring higher-order mental state attribution to humans and machines

Olle Ageskär, Ida Allander, Markus Degerstedt, Olof Hyland, Philip Nguyen, Hilda Nääs, Nils Wickman
Supervisors: Sam Thellman & Emily Hofstetter

Abstract

This study aims to investigate individuals' attribution and preference concerning Theory of Mind (ToM) abilities in five different agents, ("*you*", *five-year-old child, self-driving car, AI-chatbot and virtual assistant*). To accomplish this objective, data was collected through the administration of surveys that aimed to assess participants' attribution and preference regarding ToM in AI-artifacts.

The key findings of the study reveal a pattern among the participants, wherein they attributed lower ToM abilities to the technical artifacts compared to the human agents. Furthermore, the findings indicate a preference for higher ToM in the agent *self-driving car*, despite attributing lower capabilities. In contrast, for the *AI-chatbot* and *virtual assistant* agents, participants expressed a preference for lower ToM abilities. A correlation was observed between the attribution of fundamental cognitive capabilities, such as having beliefs and desires, and the possession of higher-order ToM abilities.

The study's conclusions highlight the challenge of constructing a survey that effectively captures participants' perceptions while minimizing excessive cognitive demands. In future research, a valuable direction would be to collect quantitative data that more accurately captures individuals' perceptions and the underlying causes. Exploring people's attributions and desires concerning AI-agents holds scientific value within the research field and is also of great interest to external stakeholders and actors involved in the development of interactive AI-artifacts. We posit that research in human-AI interaction stands to benefit from insights derived from human perspectives, encompassing personal attributions and preference.

1. Introduction

The advancement of technology and integration of Artificial Intelligence (AI) in applications and technical artifacts reveals questions concerning the presence of human characteristics in these technical artifacts. To develop and refine human–machine interactions in everyday contexts, it's necessary to understand how humans interpret and attribute abilities to these types of artifacts. Previous studies have investigated mental state attribution to technical artifacts. However, a significant aspect that remains unexplored pertains to the extent to which individuals attribute higher-order ToM abilities to AI-artifacts.

This study aims to investigate individuals' attributions and preferences for five agents ("*you*", *five-year-old child*, *virtual assistant*, *self-driving car* and *AI-chatbot*) to exhibit capabilities associated with different orders of Theory of Mind. To achieve the primary objective the following three research questions were investigated:

RQ1 - Is there a difference between how people attribute varying orders of Theory of Mind to human agents ("*you*" and *five-year-old child*) compared to technical artifacts with different degrees of implemented AI *(virtual assistant*, *self-driving car* and *AI-chatbot*)?

RQ2 - Is there a difference between how people attribute Theory of Mind to technical artifacts and their preference for these artifacts to possess Theory of Mind?

RQ3 - Is there a correlation between attributing higher and lower orders of Theory of Mind?

2. Background

Theory of Mind (ToM) refers to the cognitive ability that allows individuals to attribute mental states such as beliefs, desires, and intentions, to oneself and others, and to understand that others have beliefs, desires, intentions, and perspectives that may differ from one's own (Leslie, 2001).

The concept of ToM and its various orders, as defined by Lowry (2016), forms the bedrock of our investigation. The main distinction between first-

(ToM_1) and second-order (ToM_2) Theory of Mind is that ToM_1 is about being aware of other people's beliefs and desires whereas ToM_2 is about being aware of what others think about other people's beliefs and desires. This study will also refer to the term zero-order theory of mind (ToM_0), that represents the fundamental capacity to hold beliefs or desires. Generally speaking, normally-developing children master first-order ToM by the age of five (Wellman, et al., 2001). The agent *5-year-old-child* is included in the study to compare the technical agents' to a well established development stage.

Upon encountering a computer system with even rudimentary linguistic capabilities, it is not uncommon for adult users to assume the presence of a more sophisticated language comprehension (Suchman, 2006). These assumptions often stem from our limited experience with entities that utilize language, which until recently, has been primarily other humans. This study will focus on how these innate human tendencies to attribute mental states affect the perception of AI-artifacts.

Mou et al. (2020) present the correlation between attributing higher ToM, and trusting an agent. Additionally, Troshani et al. (2020) found that anthropomorphism and intelligence were integral to user trust and interaction with AI applications and created a "social presence". This perceived social presence within AI systems points to a possible inclination of users to attribute ToM to AI.

However, Ivarsson and Lindwall (2023) found that AI artifacts with human-like voices decreased the reliability of interaction compared to those with non-human voices, if the user was unsure if the agent was human or not. This perspective raises important considerations for this study, highlighting potential negative implications of overly anthropomorphized AI artifacts.

Finally, the phenomenon of correspondence bias, as presented by Gilbert & Malone (1995), is the tendency of people to attribute dispositional characteristics to other agents based on observed behavior, leading to broader conclusions about the actors' abilities, disregarding potential logical limitations. According to the theory, people make attributions because doing so enables them to derive intentions and thereby control the extent to which others' behavior can affect them. This concept plays a vital role in this study, helping understand why people may attribute more abilities to AI artifacts than they truly possess and the implications of such attributions.

3. Method

A survey-based design was employed in this study. Initially, the study was run on Amazon Mechanical Turk (MTurk) to gather responses from a larger number of participants, but due to concerns about the quality of the data obtained, a second study was conducted using convenience sampling. Thus, this paper contains two separate results sections, each corresponding to one of the studies.

The survey consisted of 123 questions, where 60 regarded *attribution* of ToM in agents, 60 regarded *preference* for ToM in agents, and three questions were for the purpose of quality control. The questions included two questions regarding attribution of ToM and two questions regarding preference of ToM per order from zero to two, giving a total of 12 instances of questions. An example of attribution of first-order Theory of Mind (ToM_1) is "Which of these agents is most likely to know what others know?" and for preference for ToM_1 "Which of these agents would you most prefer to know what others know?". Further examples can be found in supplementary materials (p. 6). Each question was structured around agent pairs, given ten permutations. For each agent-pair and question, the participants were asked to decide which agent they attributed or preferred ToM to using a Likert scale of 0-4. These scores were then normalized to fall within a range of zero to one by dividing each score by the maximum possible score. A score of one corresponds to fully attribute or desire ToM to that agent. This was done for both the attributed and desired ToM for each order from zero to two.

A G*Power analysis yielded a sample size of 124 for independent t-test and 55 for One-Way ANOVA. Due to an expected data loss the sample for data collection was set to 350 participants in Study 1, out of the initial pool, only 56 passed our data quality measures. The exclusion criteria were as follows:

1. *Control questions:* Only participants who answered all three control questions correctly were included in the final dataset. This was to ensure comprehension and attentiveness to the survey questions.
2. *Response pattern:* Participants who selected more than 50% of their answers on either the two leftmost or two rightmost options were excluded. This criterion was based on a noticeable trend of non-differential responding in the survey of Study 1.
3. *Completion time:* Participants who spent less than 15 minutes on the survey were excluded. This was to

ensure adequate time for thoughtful consideration of each question.

After the data had been screened, Study 1 consisted of 14 women, 42 men, 0 non-binary, and 0 others. The age of the participants lacked normal distribution and ranged between 22-42 years (M = 30.1, SD = 3.3).

To address the quality issues observed in the initial data collection, a second survey was conducted. This time, a convenience sample of 34 participants were recruited by researchers through in-person recruitment and through social media posts (16 women, 17 men, 1 non-binary, and 0 others). The age of the participants lacked normal distribution and ranged between 19-57 years (M = 28.4, SD = 10.8).

To improve the quality of the data collected in the second survey, an important modification was made to the questionnaire. Specifically, an additional response option was included: "I don't understand the question". This option facilitated an assessment of participants' understanding of the survey questions and helped identify any potential issues related to clarity or comprehension.

4. Results

4.1 Research Question 1

- Is there a difference between how people attribute varying orders of Theory of Mind to human agents (“*you*” and *five-year-old child*) compared to technical artifacts with different degrees of implemented AI *(virtual assistant*, *self-driving car* and *AI-chatbot*)?

STUDY 1

Separate One-Way ANOVA tests were conducted for each order of ToM, yielding the following results: ToM$_0$ (F = 10.49, p < .001, Cohen's f = 0.14), ToM$_1$ (F = 5.08, p < .001, Cohen's f = 0.10), ToM$_2$ (F = 3.28, p = .013, Cohen's f =0.08). The results show a similar attribution of combined ToM for the human agents, with a slightly higher attribution to the technical artifacts.

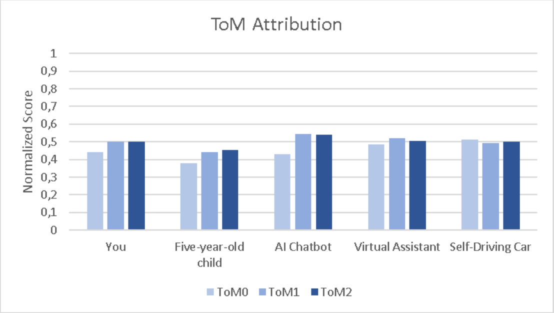

Figure 1. Mean of attributed ToM for each agent clustered by ToM order. The higher the assigned value, the more participants have chosen to attribute the abilities to the agents in the comparative questions.

STUDY 2

Separate One-Way ANOVA tests were conducted for each order of ToM, yielding the following results: ToM$_0$ (F = 180.20, p < .001, Cohen's f = 0.71), ToM$_1$ (F = 29.30, p < .001, Cohen's f = 0.35), ToM$_2$ (F = 23.10, p < .013, Cohen's f = 0.32). The results show a higher attribution of combined ToM for the human agents, compared to the technical artifacts.

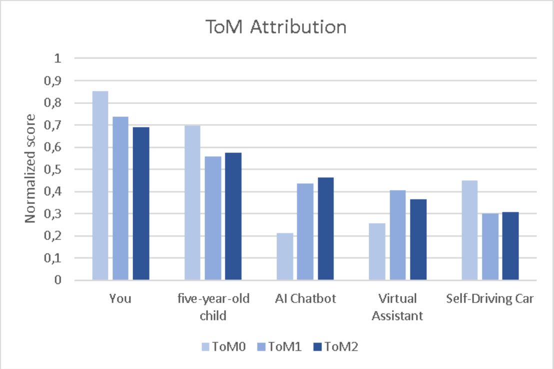

Figure 2. Mean of attributed ToM for each agent clustered by ToM order. The higher the assigned value, the more participants have chosen to attribute the abilities to the agents in the comparative questions.

4.2 Research Question 2

- Is there a difference between how people attribute Theory of Mind to technical artifacts and their preference for these artifacts to possess Theory of Mind?

STUDY 1

In order to investigate differences between the participants' attributed and preferred ToM abilities of the technical artifacts, paired sample T-tests were conducted: *AI-chatbot* (t(55) = -2.63. p = .011, Cohen's d = 0.05), *self-driving car* (t(55) = 3.28, p = .002, Cohen's d = 0.06).

The results revealed that participants exhibited a stronger preference for combined ToM in the *AI-chatbot* than they attributed to it. Conversely, for the *self-driving car*, the preferred ToM was found to be less than the attributed ToM.

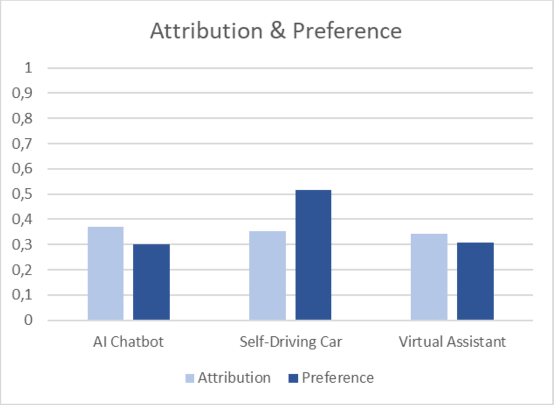

Figure 3. Mean of attributed ToM across all orders and mean of preferred ToM combined for all orders.

STUDY 2

In order to investigate differences between the participants' attributions and preferences of the ToM abilities of the technical artifacts, a paired sample T-test was conducted: *AI-chatbot* ($t(33) = 2.85$, $p = .008$, Cohen's $d = 0.09$), *self-driving car* ($t(33) = -6.51$, $p < .001$, Cohen's $d = 0.20$), *virtual assistant* ($t(33) = 2.11$, $p = .042$, Cohen's $d = 0.06$).

The findings indicated that participants demonstrated a stronger preference for combined ToM in *self-driving car* than they attributed to it. Conversely, for the *AI-chatbot* and *virtual assistant*, the preferred ToM was found to be less than the attributed ToM.

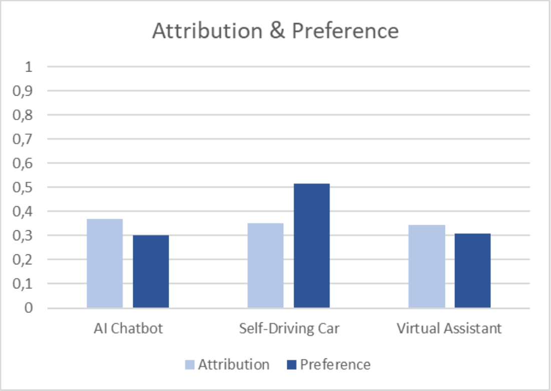

Figure 4. Mean of attributed ToM across all orders and Mean of preferred ToM across all orders.

4.3 Research Question 3

- Is there a correlation between attributing higher and lower orders of Theory of Mind?

STUDY 1

To determine the relationship between agents attributed ToM_0 and attribution of higher order ToM a Pearson's Correlation test was performed on each agent: *"you"* (Pearson's $r = .47$, $p < .001$), *AI-chatbot* (Pearson's $r = .36$, $p = .006$) *virtual assistant* (Pearson's $r = .33$, $p = .012$). The results showed weak correlation between ToM_0 and higher order of mind. There were no significant results for the other agents.

STUDY 2

To determine the relationship between agents attributed ToM_0 and attribution of higher order ToM a Pearson's Correlation test was performed on each agent: *self-driving car* (Pearson's $r = .49$, $p = .004$), *virtual assistant* (Pearson's $r = .54$, $p = .001$). The results showed moderate correlation between ToM_0 and higher order of mind. There were no significant results for the other agents.

5. Discussion

The results show differences between Study 1 and Study 2 although they were performed on the same research question and were processed by the same analysis methods. The differences between the two studies were the different participant groups and survey formulations, these factors should be viewed as potential influences on the observed outcomes.

The high proportion of excluded data from Study 1 insinuate an overall lack of validity. Because of this, no further conclusions will be drawn from the result of Study 1. Instead, the results will serve as a reference to compare the results with in Study 2, highlighting the importance of methodological choices when conducting research. Hence, the discussion of results will only concern the results of Study 2.

5.1 Research Question 1

- Is there a difference between how people attribute varying orders of Theory of Mind to human agents ("you" and five-year-old child) compared to technical artifacts with different degrees of implemented AI (virtual assistant, self-driving car and AI-chatbot)?

The findings show higher attribution of ToM abilities to human agents compared to the technical artifacts, where the agent *"you"* which referred to the participants themselves had the highest score, followed by the agent *five-year-old child*. The technical artifacts *AI-chatbot*, *self-driving car*, and *virtual assistant* received comparably lower ToM attributions, clustering at the lower end of the scale. Among these, the *AI-chatbot* was perceived slightly higher in terms of ToM attributes compared to the *self-driving car* and *virtual assistant*.

The work of Mou et al. (2020) emphasizes that there is a strong connection between trust and the attribution of high ToM. Their research underscores the importance of social abilities, including attributed ToM, in fostering perceived trustworthiness in human-robot interaction (HRI). In line with these insights, their findings could potentially explain the results of RQ1, where the participants attributed lower ToM to the technical artifact due to a perceived sense of trustworthiness. Consequently, further research is warranted to investigate humans' perceptions of these artifacts, as well as the underlying factors that influence trust and attribution of ToM.

Likewise, low preference to the agent *AI-chatbot* and *virtual assistant* could highlight potential problems in the interaction with these artifacts. Suchman (2006) claims that computational artifacts that display some degree of recognizable human abilities, increase the tendency of people to attribute them with even more capabilities than they possess. Both artifacts have the abilities to communicate by language, which would be a potential ability that strengthened the trust towards them. A further study that indicates that human-like attributes in AI-artifacts results in better interaction and trust is conducted by Troshani et al. (2020). These claims are not supported by the findings of this study, therefore the reason for the low preference should be discussed and acknowledged, since it could cause problems in interaction.

5.2 Research Question 2

- Is there a difference between how people attribute Theory of Mind to technical artifacts and their preference for these artifacts to possess Theory of Mind?

The findings indicate that participants expressed a preference for a lower capacity for ToM in the agents *AI-chatbot* and *virtual assistant* while preferring a higher capacity for ToM in the *self-driving car* agent.

Although self-driving cars are not yet fully integrated into traffic systems, their development necessitates an understanding of human perceptions in social interaction. If pedestrians and other drivers do not trust self-driving cars, it may lead to hindrances and challenges in traffic situations. What sets the agent self-driving car apart from the other technical artifacts in the study is its physical agency. The greater impact of the agent's abilities and decision-making processes may elicit a stronger preference for the agent to possess an understanding of others' intentions, beliefs, and desires. However, since the participants' motivations were not captured in the survey responses, no definitive conclusions can be drawn beyond these tentative suggestions, underscoring the need for further qualitative investigations.

Ivarsson and Lindwall (2013) found that AI-artifacts with human-like voices, a prominent feature of most virtual assistants, can decrease the reliability of interaction. This effect was attributed to users feeling deceived when the human-like entity they were interacting with was, in fact, not human. The results showing lower preference of ToM abilities in the virtual assistant agent support the theory presented in their study.

5.3 Research Question 3

- Is there a correlation between attributing higher and lower orders of Theory of Mind?

The results showed a moderate correlation for the agents *self-driving car* and *virtual assistant* (no significance for *AI-chatbot).* Looking at the results, a certain correlation can be discerned and support that the attribution of ToM_0 could serve as a foundational measure since it reveals participants attribution to the agent's capacities to hold beliefs and desires. These attributes can be considered prerequisites for higher order ToM tasks, thus the results support the research question's intention to confirm this proposition.

Disregarding the investigation of correlation, and instead observing the extent of attribution of ToM_0 amongst the agents an intriguing pattern is displayed. The formulation of the questions regarding ToM_0 were phrased by the following format; "Which of these agents is most likely to know things about their environment?". Analyzing the results, the agents who have the ability to perceive information about their physical environment showed a higher attribution of ToM_0 capacities (*"you", five-year-old child, self-driving car*). These results could indicate that the question's formulation created a bias towards attributing the capacities to agents that act on or or perceive their physical environment. The agents *AI-chatbot* and *virtual assistant* may have been overlooked due to the fact that they cannot act directly on their environment, or perceive their environment without voice or text input.

5.4 Methodological Discussion

Upon examining the outcomes of Study 1 and Study 2 and considering the inconsistencies observed, these issues primarily stem from the data collection process in Study 1 and subsequent adjustments made to the survey format in Study 2.

The rationale behind introducing an additional response option ("I don't understand the question") for Study 2 was to investigate potential reasons for the quality issues, which included speculation about question phrasing and high cognitive demands. The results observed in Study 2 support the notion that articulating and solidifying these ideas impose higher cognitive demands on individuals. To decrease the cognitive demands of the participants, the questions could be revised to a more contextualized form, in contrast to participants needing to imagine abstract mental scenarios in order to make comparison between agents. Additionally, it could have been beneficial to provide background on the concept of Theory of Mind, by explaining that the participants could have been better equipped to understand the nature and implications of the questions being asked.

Another potential improvement would be to adopt a qualitative approach, which could help address several of the issues identified in our study. By offering less rigidly structured questions, we could lessen biases from question formulation, allowing participants to more freely express their views.

Abstract inquiries could also be made less taxing, making the comprehension of ToM concepts more accessible. Furthermore, a qualitative approach could minimize self-reporting bias by encouraging comprehensive sharing of experiences and perspectives and help understand the underlying causes for the participants' attribution.

6. Conclusion

The results of the study reveal a pattern among the participants, wherein they attribute lower Theory of Mind abilities to the technical artifacts compared to the human agents involved in the study. Furthermore, the findings indicate a higher preference for ToM abilities in the agent *self-driving car*, despite attributing it with less ToM. In contrast, for the *AI-chatbot* and *virtual assistant* agents, participants expressed a lower preference for ToM abilities. A correlation emerged between the attribution of basic cognitive abilities, such as having beliefs and desires about the physical world, and the possession of higher-order ToM abilities. The observed patterns and correlations not only contribute to the understanding of how people perceive and interact with AI agents but also underscore the need for considering user attitudes and perceptions in the design and development of AI technologies.

References

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117*(1), 21-38. https://doi.org/10.1037/0033-2909.117.1.21

Ivarsson, J., & Lindwall, O. (2023). Suspicious Minds: the Problem of Trust and Conversational Agents. *Computer Supported Cooperative Work* (CSCW). https://doi.org/10.1007/s10606-023-09465-8

Leslie, A. M. (2001). Theory of Mind. *International Encyclopedia of the Social & Behavioral Science*, 15652-15656. https://doi.org/10.1016/B0-08-043076-7/01640-5

Lowry, L. (2016). Thinking about Thinking: How young children develop theory of mind. *The Hanen Centre.* https://www.hanen.org/MyHanen/Articles/Research/Thinking-about-Thinking--How-young-children-develo.aspx

Mou, W., Ruocco M., Zanatto D., & Cangelosi A. (2020). When Would You Trust a Robot? A Study on Trust and Theory of Mind in Human-Robot Interactions. *29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 956-962. https://doi.org/10.1109/RO-MAN47096.2020.9223551

Suchman, L. (2006). *Human-Machine Reconfigurations: Plans and Situated Actions* (2nd ed.). Cambridge University Press.

Troshani, I., Hill, S. R., Sherman, C., & Arthur, D. (2020). Do We Trust in AI? Role of Anthropomorphism and Intelligence.*Journal of Computer Information Systems, 61*(5), 481-491. https://doi.org/10.1080/08874417.2020.1788473

Wellman, H. M., Cross, D. R., & Watson, J. (2001). Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child Development, 72*(3), 655-684. https://doi.org/10.1111/1467-8624.00304

Supplementary materials