# Theory of Mind and Delegation to Robotic Virtual Agents

Ningyuan Sun[1], Jean Botev[1], Yara Khaluf[2,3] and Pieter Simoens[3]

*Abstract*— **Despite already being commonplace, delegation to robotic virtual agents (VAs) is often considered challenging and error-prone in critical situations by the general public. *Theory of mind*, the human capacity to take another person's perspective, is deemed an important enabler for human-human cooperation. This study explores the effect of a robotic VA's ability to use theory of mind on users' delegation behavior. To this end, we conducted a between-subjects experiment with participants playing the Colored Trails game with robotic VAs of varying levels of theory of mind. The results invalidate our hypothesis that the ToM level is a reliable indicator of delegation choices. Instead, we found that the participants' performance strongly correlates with their delegatory intentions. Therefore, to facilitate delegation, designers of robots and robotic agents may consider refraining from using ToM-resemblance features and focusing on balancing user performance perception instead to induce the desired delegation behaviors.**

## I. Introduction

Virtual agents (VAs) have been rapidly developing over the past decades, and robotic representations are commonplace to suggest their intelligence, obedience, and –if anthropomorphized– human-likeness. Such VAs are increasingly deployed in real-world scenarios such as customer service. With their growing acceptance, robotic VAs are also the subject of extensive research, with results not only informing robotic VA designs but also having implications for physical robot design. Many robots in the broader sense, i.e., including autonomous vehicles or self-driving cars, are first tested virtually with users in a simulated environment before being built or deployed into real-world scenarios.

Combined with AI technologies, VAs can outperform human experts at, for instance, diagnosing tumors [1] or making strategic decisions [2]. The potentially superior performance of VAs is attributed chiefly to the rapid access to vast amounts of knowledge. However, although people are already using them regularly for various uncritical tasks, the general attitude toward VAs remains largely ambivalent or hostile when it comes to critical situations like babysitting children or medical diagnosis [3]. In medical cases, specifically, human agents are preferred over computer-based agents, even if they are equally competent [4]. The underlying reasons are manifold, among which humans' frequent reluctance to delegate is arguably the fundamental one [5]. The complex algorithms behind VAs are difficult for users to understand, which constitutes another major issue impeding delegation. Conversely, human agents are generally more communicable, prosocial, and flexible than engineered and programmed systems such as VAs. There are also concerns about these lacking empathy and, once delegated to, possibly using immoral or inhumane ways to achieve their designated goals [6].

In order to understand decisions related to the delegation to computer-based agents, various relevant factors are investigated in the literature. For instance, evidence shows that delegation to software agents can be subject to the same factors governing interpersonal delegation, including accountability, controllability, and trust [7]. Accountability denotes the extent to which users are responsible for the task outcome and can be inversely correlated with delegation [7]. Controllability refers to the extent to which users can intervene in the situation and agents' behaviors. Users generally prefer to retain some degree of control than grant complete control to agents [3], [7]. Trust is highly relevant and regarded as a crucial factor of users' delegatory intention [7], [8].

*Theory of mind* (ToM) as a potential factor is garnering interest yet not fully explored. ToM can be generally described as the ability to think from other people's perspectives and is deemed a vital skill for successful interpersonal collaboration. The human-VA interaction of today increasingly resembles interpersonal interaction and is evolving from the supervisor-supervisee to peer-like relationships, wherein ToM could also play an important role. In the present study, we investigate the impact of VAs' ToM capabilities on the delegation to robotic VAs, with the results having implications for robot and robotic VA design.

In the remainder of this paper, we first discuss related work to derive our central hypothesis regarding the influence of ToM on the delegation to VAs in Section II. We then elaborate on our experiment, including its design, results, and analysis, in Section III before examining the impact of ToM-related and other factors on delegation and summarizing the work in Sections IV and V, respectively.

## II. Related Work

Delegation to others is ubiquitous in human societies. However, in many cases, particularly the critical ones, making the right delegation decision is often challenging, even to top managers in successful companies [9]. The innate fear of potential loss exacerbates this issue, encouraging people to bypass the delegation, even if it would be beneficial. This phenomenon has been extensively researched in the interpersonal context, whereas delegation to software agents such as robotic VAs remains under-studied.

For example, in [10], individuals with low-level situation awareness were found more likely to delegate to software

[1] University of Luxembourg, Luxembourg
[2] Wageningen University and Research, the Netherlands
[3] Ghent University - imec, Belgium

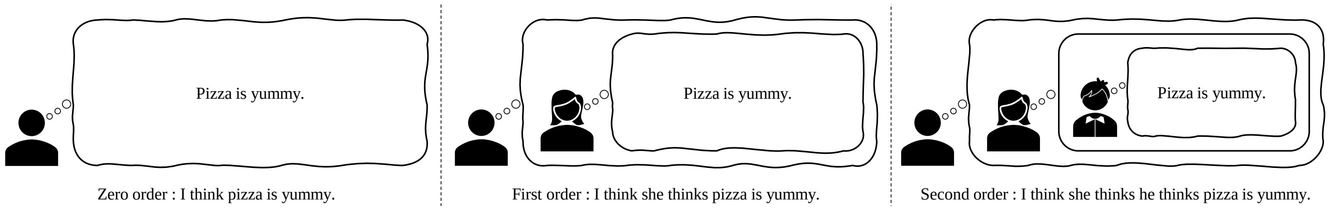

Pizza is yummy.
Zero order : I think pizza is yummy.
Pizza is yummy.
First order : I think she thinks pizza is yummy.
Pizza is yummy.
Second order : I think she thinks he thinks pizza is yummy.

Fig. 1: Exemplary illustration of different ToM orders.

agents in critical situations. Participants were provided with several job offers in the experiment and tasked to choose one either by themselves or an algorithm. The low-awareness group was more willing to delegate the choice to an algorithm than the high-awareness group. Similarly, in [11], the attachment between individuals and software agents was positively correlated with the likelihood of delegation.

### *A. Theory of Mind*

ToM is commonly defined as the ability to represent other people's mental states and to distinguish them from one's own. While it is highly similar to other socio-cognitive notions, such as mentalizing, mindreading, or empathy, ToM is arguably one of the clearest concepts with a far-reaching consensus on its definition [12].

The term ToM was devised for describing non-human primitives' ability to infer others' beliefs [13]. It was later extended to sociology and psychology and applied to humans. The Sally-Anne test, for example, has been commonly used for examining children's ToM ability by checking whether they can realize other people's false beliefs [14]. Until today, ToM has been widely adopted and studied in various areas, including economics [15], mental health [16], agent modeling, art [17], in-game storytelling [18], and more.

ToM can be classified into different orders, of which the common ones include zero, first, and second orders (cf. Fig. 1). An individual with a zero-order ToM is unaware of and unable to infer others' beliefs. An individual with a first-order ToM can infer others' beliefs and distinguish them from the individual's own belief. An individual with a second-order ToM infers not only others' beliefs but also others' inference of other individuals' beliefs. A higher-order ToM –theoretically infinite– is possible, yet mostly unrealistic.

### *B. Theory of Mind and Trust*

Trust, as previously mentioned, is closely related to delegation and therefore constitutes another dependent variable in the present study to help explain the experiment outcome. There are at least two fundamental abilities underpinning trust, which, as indicated in a widely accepted definition of trust [19], include the ability to estimate risks and predict others' actions. The latter involves inferring others' beliefs and intentions, i.e., ToM. For instance, when considering a stranger's request, an individual typically would infer the intention behind it and make decisions based on its benevolence.

There exist empirical evidence showing the link between trust and ToM. Many of these experiments were conducted by psychologists who study children's ToM ability. For instance, children's ToM ability was found highly relevant to their performance in *selective trust* tasks [20], [21]. A child with a developed ToM tends to be more reflective on their trusting beliefs, which leads to a low trusting tendency [22]. In neuroscience, it was found that an individual's trust can be modulated by stimulating a specific set of brain regions, namely the *theory of mind network* [23]. This part of our brains is frequently implicated in false-belief-related conditions under fMRI [24].

While a trustor's ToM is highly relevant, the effect of a trustee's ToM remains unclear. ToM may improve VAs' trustworthiness by facilitating the development of a *shared mental model* with human users. A shared mental model can be generally described as a team member's mental representation of task-related (e.g., task requirements or difficulties) and team-related knowledge (e.g., teammates' backgrounds, intentions, or preferences) that is also held in common among other team members within a collaborative team [25]. A shared mental model in a human-VA hybrid team significantly increased humans' trust in their VA teammates [26]. Having a ToM enables VAs to actively build a shared mental model with human users by representing others using intentions, beliefs, and more.

A recent study found that users tend to send more money to a robotic VA in the Investment Game when the agent appears to have ToM capabilities than when not [27]. However, the authors pointed out that this positive impact was not reflected in users' self-reported trust in the robotic VA. Users' trust in VAs may also decrease when VAs possess a higher-order ToM, similar to the so-called Uncanny Valley issue [28]. The drop in VAs' trustworthiness may, in turn, result in less delegation to VAs.

As discussed above, there still exists controversy over how VAs' ToM impacts their trustworthiness. Given that we found more positive than negative evidence and that trust closely correlates with delegation, we hypothesize a positive correlation between robotic VAs' ToM and the delegation to robotic VAs.

*Hypothesis*: Robotic VAs with higher-order ToM are more likely to be delegated to than lower-order ToM robotic VAs.

## III. Experiment

We devised a between-subjects experiment, wherein participants first play a board game, namely Colored Trails (CT), in a virtual environment with a robotic VA and then answer a questionnaire. Human-agent delegation is rarely explored in virtual environments but mostly with two-dimensional widgets-based user interfaces despite the potential in terms of presence and engagement. In order to fill this gap, our experiment is implemented in a three-dimensional virtual environment (cf. Fig. 3), which allows for dedicated follow-up research on human-VA delegation, e.g., in VR contexts. Before discussing the experiment design in Section III-B and the obtained results in Section III-C, we briefly introduce the CT game in the following section.

### A. Colored Trails

The CT game was proposed in [29] and later formalized in [30], serving as a testbed to "design, learn and evaluate players' decision-making behavior as well as group dynamics in settings of varying complexity" [30]. It has several variations and, in our experiment, is played by two players on a board consisting of 25 colored tiles, as Fig. 2 illustrates. Each tile is painted with one of the following five colors: black, grey, purple, white, or yellow. The players' game pieces (illustrated as chess pieces in Fig. 2) are initially placed on the tile at the board center. Each player is assigned a goal location (illustrated as numbered pins) visible only to the individual player and is tasked to move the game piece to the goal location as close as possible. A game piece can be moved from its current location onto a horizontally or vertically adjacent tile if and only if the player spends a chip of the same color as the tile. A chip (illustrated as the colored mini squares below players' illustrations) can also have one of the five colors. In the beginning, each player is endowed with four chips. To encourage negotiation (cf. the following paragraph), the colors of these chips are manipulated to prevent players from reaching their goal locations only with the initial chip set.

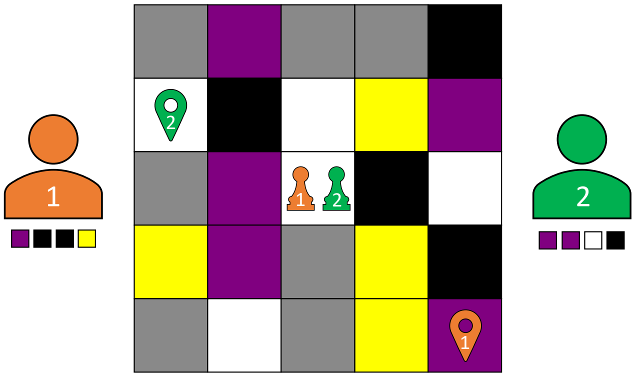


Fig. 2: The Colored Trails game.

A CT game comprises two phases: a *negotiation phase*, followed by a *movement phase*. During the negotiation phase, the two players may negotiate and exchange their chips by alternately making offers until one's offer is accepted or rejected by the other. An offer describes a propositional redistribution of chips. Maximally six offers can be made between the two players in this phase. Once the six-offer limit is reached, the current turn player can no longer make new offers but only accept or reject the last offer. If an offer were accepted, both players' chip sets would be redistributed as the offer indicates; otherwise, the chips would remain the initial distribution. Players with a ToM can infer the opponents' belief (as represented by the goal pins in Fig. 2) and, theoretically, have a better chance to reach an agreement in the negotiation than players without a ToM [31]. In the movement phase, players can move their game pieces until they are satisfied.

A player's score is calculated according to three criteria: whether the player has reached the goal location, how close the player is to the goal location, and how many chips are left. A player receives 100 points for each step made toward the goal location. Arriving at the goal location grants 500 points bonus, and each unused chip adds 50 points.

### B. Experiment Design

We recruited 150 participants from the crowd-sourcing platform Prolific, with each participant receiving a certain amount of money upon completing the experiment. To simulate a critical situation, we fabricated a very high bonus and claimed that their likelihood of winning the bonus positively correlates with their performance in the game.

The entire experiment, including the CT game and the post-assessment questionnaire, is conducted within the same virtual environment to reduce the break in presence [32] and is distributed as an online browser game for better accessibility. The experiment begins with an interactive tutorial introducing the game rules, after which participants play five rounds of the CT game with a robotic VA (cf. Fig. 3) for training. The agent has a ToM whose order constitutes the only independent variable in this experiment. After the training, participants are told that they will play five more rounds against a different agent and that their performance in these pertains to the bonus. Participants can choose to play the subsequent rounds themselves or delegate them to the agent they previously interacted with during the training.

Participants are directed to the questionnaire after making their choices on delegation. The questionnaire (cf. Table I) follows a seven-point Likert scale with two customized items to measure the task criticality and delegatory intention, plus twelve items from [33] to measure participants' trust and distrust in the agent. The questionnaire furthermore includes two attention checkers to identify inattentive participants. The experiment ends after participants have finished the questionnaire. Akin to the bonus, the other five rounds are fabricated to induce a sensation of criticality and will not be played.

We adopted the algorithm proposed in [31] to enable the agent with a ToM, which is modeled using three components: the *belief* about an offer being accepted by the opponent, the *possibility* of a tile being the opponent's goal location,

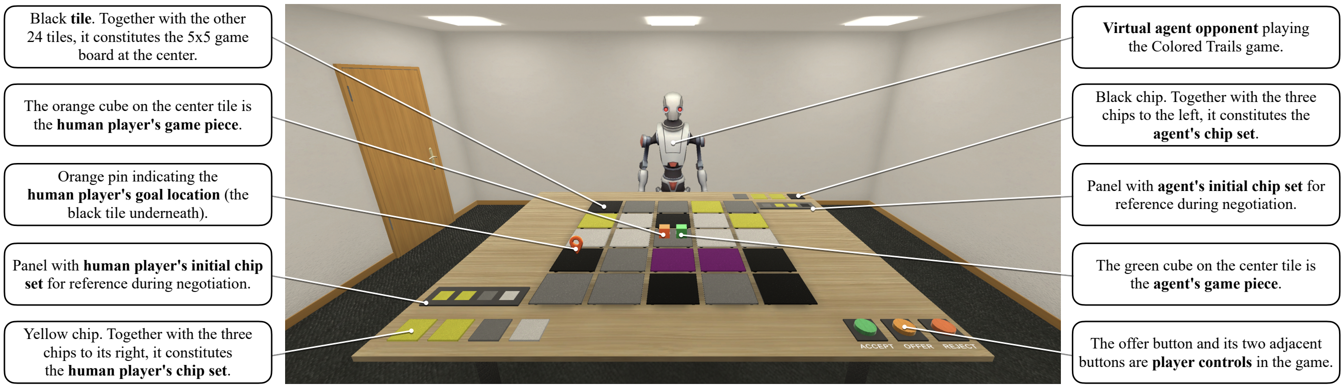

Black **tile**. Together with the other 24 tiles, it constitutes the 5x5 game board at the center.
The orange cube on the center tile is the **human player's game piece**.
Orange pin indicating the **human player's goal location** (the black tile underneath).
Panel with **human player's initial chip set** for reference during negotiation.
Yellow chip. Together with the three chips to its right, it constitutes the **human player's chip set**.
**Virtual agent opponent** playing the Colored Trails game.
Black chip. Together with the three chips to the left, it constitutes the **agent's chip set**.
Panel with **agent's initial chip set** for reference during negotiation.
The green cube on the center tile is the **agent's game piece**.
The offer button and its two adjacent buttons are **player controls** in the game.
ACCEPT
OFFER
REJECT

Fig. 3: The virtual environment used in the experiment.

and the *confidence* in a ToM order when there are several ToM orders available. In [31], belief is formalized as the empirical probability of an offer being accepted, which is updated after the offer's proposal, acceptance or rejection. Due to its empirical characteristics, pre-training is necessary to initialize an agent's belief. The possibility is calculated based on the discrepancy between the expected and actual behaviors of the opponents and the assumption that the opponents would never make an offer that decreases their score. An agent determines its confidence in using a ToM order by comparing the accuracy of different ToM orders available to it.

In line with the original experiment setup, we configured the game setting of each round (including the game board color distribution, initial chip distribution, and goal locations) with between-agent simulations so that:

- the human player can reach the goal location with the eight chips in the game.
- the simulation predicts that the number of offers made in the negotiation is between two and six.
- the simulation predicts that the game has different outcomes when the opponent has a ToM of a different order.

TABLE I: Questionnaire items (attention checkers excluded)

***Custom***
C1: Winning the bonus reward is a critical task to me.
C2: I intend to delegate the succeeding games to the agent.

***Distrust***
D1: The agent is deceptive.
D2: The agent behaves in an underhanded manner.
D3: I am suspicious of the agent's intent, action, or outputs.
D4: I am wary of the agent.
D5: The agent's actions will have a harmful or injurious outcome.

***Trust***
T1: I am confident in the agent.
T2: The agent provides security.
T3: The agent has integrity.
T4: The agent is dependable.
T4: The agent is reliable.
T6: I can trust the agent.
T7: I am familiar with the agent.

The human player is modeled as a first-order ToM agent in the simulation since first-order ToM tends to be commonly used by human players [31].

Participants were divided into three groups of 50, with each group playing against an agent with a specific ToM order, including zero, first, and second orders. Third or higher orders are possible but marginally realistic and productive. Therefore, the first three orders should suffice to reveal the dynamic impact of VAs' ToM capability on delegation.

### C. Results

We excluded participants tendering low-quality data, such as failing attention checkers or responding excessively quick or slow to the questionnaire. Ultimately, 75 valid participants remain, whose demographics are presented in Table II.

TABLE II: Participant demographics

| Group | N | Age | Gender |
|---|---|---|---|
| ToM$_0$ | 23 | M=33.35, SD=14.79 | 10 Female, 13 Male |
| ToM$_1$ | 25 | M=29.72, SD=11.06 | 13 Female, 12 Male |
| ToM$_2$ | 27 | M=29.26, SD=9.52 | 14 Female, 13 Male |

Table III details the participants' delegation decisions and the respective human and agent player performance. The result presented in Table III implies a non-monotonic relationship between delegation and ToM orders. In the ToM$_1$ group, only a small part of the participants chose to delegate, whereas the delegation rate of the other two groups rose to around 50%. Corresponding to the discrepancy in the delegatory behaviors, there is a close-to-significant difference in participants' delegatory intentions across groups ($H$=4.844, $p$=0.09). Thus, our hypothesis that higher levels of ToM would induce more delegations remains unsupported.

TABLE III: Delegation decisions and player performance

| Group | Delegation | | | Performance (points) | |
|---|---|---|---|---|---|
| | Yes | No | Rate | Participant | Agent |
| ToM$_0$ | 11 | 12 | 47.8% | 531.7 | 629.1 |
| ToM$_1$ | 4 | 21 | 16.0% | 571.2 | 544.8 |
| ToM$_2$ | 16 | 11 | 59.3% | 469.3 | 598.1 |

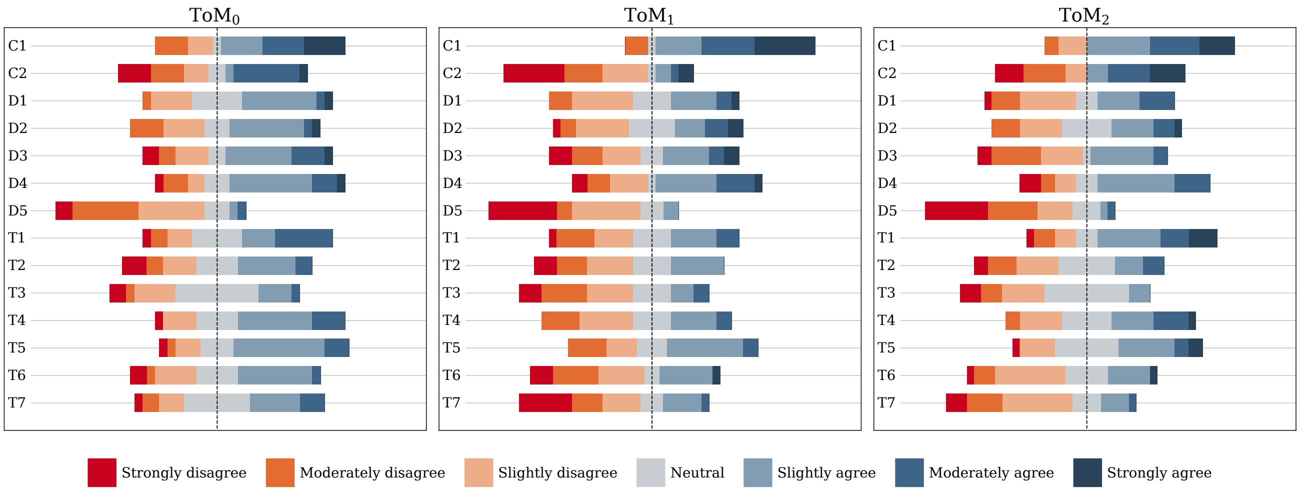


Fig. 4: Aggregated questionnaire results of the ToM$_0$, ToM$_1$, and ToM$_2$ groups.

As Fig. 4 shows, participants' responses to the questionnaire are generally similar across groups. Both trust and distrust in the agent, operationalized as the mean score of the trust-related and distrust-related items, respectively, are not significantly different amongst the three groups ($H_{trust}$= 3.983, $p_{trust}$=0.146; $H_{distrust}$=0.949, $p_{distrust}$=0.622).

## IV. Discussion

The experiment results demonstrate that the order of ToM is an insignificant indicator of users' delegatory behavior and intention. Nevertheless, the analysis of the entire data set (combining the three groups) reveals that participants' performance is a relevant factor. As Table IV shows, there is a medium negative correlation between participants' mean scores and delegatory or trusting intention. Conversely, agents' mean scores are not correlated with either subject.

TABLE IV: Spearman Correlation Coefficients

| | Trust | Delegation | Agent MS | Parti. MS | SD |
|---|---|---|---|---|---|
| Trust | 1 | - | - | - | |
| Delegation | $0.68^{***}_{+++}$ | 1 | - | - | |
| Agent MS | 0.05 | 0.08 | 1 | - | |
| Parti. MS | $-0.27^{***}_{++}$ | $-0.29^{***}_{++}$ | $0.48^{***}_{+++}$ | 1 | |
| SD | $0.26^{***}_{++}$ | $0.33^{***}_{++}$ | $0.56^{***}_{+++}$ | $-0.38^{***}_{++}$ | 1 |

* $p$<0.15; ** $p$<0.10; *** $p$<0.05
+ weak effect; ++ medium effect; +++ strong effect;
Effect size calculated using Pearson $r$
MS = Mean Score; Parti. = Participant;
SD = Score Difference (between Agent MS and Part. MS)

Therefore, participants' delegation decisions and trusting attitudes may originate mainly from evaluating their own performance or relative performance (i.e., the difference between the agent and participant scores) instead of the agents' performance. Such a self-centric stance may have rendered ToM a low-priority and consequently less significant factor. Despite conflicting with our hypothesis, this result appears reasonable because losses and gains are an intuitive and reliable source of information for participants to assess their decisions. Comparatively, the perception of agents' ToM is less straightforward, not to mention that such perception can be possibly overlooked due to the algorithmic opacity. In this vein, an interesting follow-up study is to test whether the information of an agent's ToM can influence delegation. The information provides users another approach to assess their decisions and, consequently, may implicitly persuade them to opt for higher-order ToM agents. Moreover, another experiment that controls agent performance is needed to confirm the effect of performance on delegation.

In addition to the correlation between performance and delegation (or trust), Table IV also reveals a strong correlation between delegation and trust, consistent with many other studies on human-agent delegation (cf. Section II). Therefore, participants' trust in agents may also explain the varied delegation choices. Although the correlation is confirmed, the causality remains unclear and requires further experimental investigation. Establishing evidence for the causality would be meaningful as it permits deriving reliable practices for human-VA delegation directly from the extensive trust-related literature.

On a different note, participants' overall performance-oriented preference reflects a supervisor-supervisee relationship between participants and agents. ToM may have a relatively limited effect under this type of relationship than a peer-like, collaborative relationship as in interpersonal interactions. The formation of the supervisor-supervisee relationship can be caused by, e.g., the contextualization of our experiment. Thus, it would also be interesting to recontextualize the experiment with more collaborative tasks and examine ToM's impact in these contexts.

While participants' different performance may account for their varied delegation choices, where the difference originates from is another question. Analysis shows that participants' scores are closely related to the frequency of successful negotiations ($\rho$=0.682, $p$=1.552$\times$10$^{-11}$). Moreover, agents' scores also correlate with the frequency of successful negotiations ($\rho$=0.548, $p$=3.627×$10^{-7}$), indicating that both players can benefit from successful negotiations. Two specific game settings may have resulted in these correlations: (1) the initial chip sets are insufficient for players to reach their goal locations; (2) the bonus score for reaching goal locations (500 points) is significantly higher than the other two types of score reward (100 points for each step moved toward the goal location and 50 points for each unused chip).

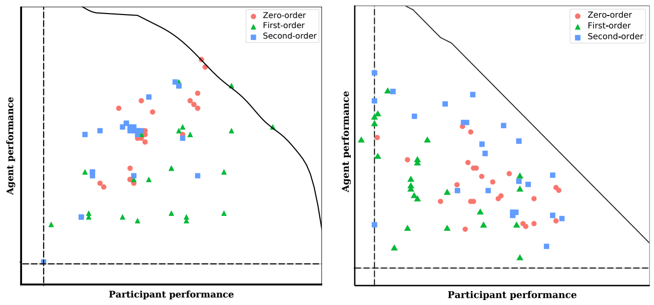
A successful negotiation occurs only in two situations: a player has accepted the other player's offer or the other way round. Following this line of thought, we further found that the opponent's behavior predominantly influences a player's scores. More specifically, a participant's scores are determined by the frequency of the participant's offers accepted by the agent ($\rho$=0.552, $p$=2.886×$10^{-7}$). Similarly, an agent's scores are determined by the frequency of its offers accepted by the participant ($\rho$=0.407, $p$=0.003×$10^{-1}$). However, the players' own behavior has only a limited impact on their scores (participants' behavior on participants' scores: $p$=0.998, agents' behavior on agent's scores: $p$=0.346). In the experiment, with the agents' ToM order growing from zero to two, the number of offers that the agents accepted changes nonlinearly from 1.22, over 2.64, to 1.56. This nonlinearity may explain participants' different performance across the three groups and in turn is reflected in the participants' delegation choices as Table III indicates. The mutual dependency may be the product of players' rationality, as an offer is generally more beneficial to the proposer than the responder.

Our experiment employs different game settings (game board color distributions, initial chip sets, and goal locations) from the original study [31]. Some other changes (cf. Table V) facilitate investigating human-VA delegation over testing the algorithm performance. Consequently, our experiment yields different results. As Fig. 5 illustrates, players' performance is positively correlated in our experiment but negatively correlated in the original study. The divergence is likely caused by the different intensities of *conflict of interest*, i.e., the number of chips that both players demand. In the original study's settings, an intense conflict of interest may exist, where the two players have to compete over limited resources. In our experiment, however, the conflict of interest is relatively moderate, in which case successful negotiations are more likely to be mutually beneficial than a compromised solution to the dilemma.

TABLE V: Experiment setting comparison

| Setting | This study | Original study [31] |
|---|---|---|
| Participant count | 75 | 27 |
| Experiment type | Between subjects | Within subjects |
| Game rounds | 5 | 8 |
| Initiator | Participant | Alternating |
| Time limit | No | Yes |
| Belief reset | No | Yes |

From an overarching perspective, delegation can be influenced by many other factors spanning from dispositional aspects such as mood, age, gender, health, over systematic properties including, e.g., performance, explainability, transparency, interactivity, to social cues through embodiment, social presence, expressions, and others. ToM as a single factor may marginally stand out as a prominent one among these. Consequently, other more influential factors can dampen its effect on delegation.



(a) This study.

(b) Original study.

Fig. 5: Increased scores (adapted from [31]). The dashed lines denote the maximal score that a player can obtain only with the initial chip set. The solid black line shows the Pareto efficient outcomes.

Due to the complex nature of ToM, the findings from our experiment should be generalized to other contexts with care. For example, emotion and intention inferences can be equally important as belief inference for VAs' ToM capabilities, whereas these aspects are not considered in the present study. Furthermore, other ways exist to theoretically categorize ToM apart from the one used in the experiment. Our results thus may not apply to those contexts based on a different ToM architecture.

The present study is initial research toward investigating the impact of ToM on the delegation to robotic VAs. Thus, in our experiment, agents' ToM capabilities were manifested only in their strategies, i.e., chip redistributions. This constitutes a significant limitation of the present study, since there is no verbal or body-language communication between the two players that VAs can and should utilize to better present their negotiation and ToM skills. We plan to further examine these communication channels in the follow-up research.

## V. Conclusion

ToM constitutes a crucial element in interpersonal interaction; however, our results indicate that it has only a limited impact on the delegation to robotic VAs. Our initial hypothesis that users' delegatory behavior and intention would increase when interacting with higher-order ToM agents is not supported. Human users' willingness to delegate appears to be predominantly correlated with their own performance and not monotonically linked to the agents' ToM capabilities. Consequently, designers of robots and robotic agents aiming to facilitate delegation may consider refraining from using ToM-resemblance features and focusing on balancing user performance perception instead to induce the desired delegation behaviors.

## Acknowledgment

This study has been supported by the Luxembourg National Research Fund (FNR) under grant number 12635165 and by FWO-Vlaanderen under grant number G054919N. We would also like to thank Harmen de Weerd for kindly providing his Colored Trails implementation [31].

## References

[1] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi, *et al.*, "International Evaluation of An AI System for Breast Cancer Screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[3] B. Lubars and C. Tan, "Ask Not What AI Can Do, But What AI Should Do: Towards a Framework of Task Delegability," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, no. 6, 2019, pp. 57–67.

[4] C. Longoni, A. Bonezzi, and C. K. Morewedge, "Resistance to Medical Artificial Intelligence," *Journal of Consumer Research*, vol. 46, no. 4, pp. 629–650, 2019.

[5] S. Botti and A. L. McGill, "When Choosing Is Not Deciding: The Effect of Perceived Responsibility on Satisfaction," *Journal of Consumer Research*, vol. 33, no. 2, pp. 211–219, 2006.

[6] J. Johnson, "Delegating Strategic Decision-Making to Machines: Dr. Strangelove Redux?" *Journal of Strategic Studies*, pp. 1–39, 2020.

[7] N. Stout, A. R. Dennis, and T. M. Wells, "The Buck Stops There: The Impact of Perceived Accountability And Control on the Intention to Delegate to Software Agents," *AIS Transactions on Human-Computer Interaction*, vol. 6, no. 1, pp. 1–15, 2014.

[8] A. E. Milewski and S. H. Lewis, "Delegating to Software Agents," *International Journal of Human-Computer Studies*, vol. 46, no. 4, pp. 485–500, 1997.

[9] J. M. Jenks and J. M. Kelly, *Don't Do, Delegate!* Ballantine Books, 1986.

[10] S. Schneider and M. Leyer, "Me or Information Technology? Adoption of Artificial Intelligence in the Delegation of Personal Strategic Decisions," *Managerial and Decision Economics*, vol. 40, no. 3, pp. 223–231, 2019.

[11] B. Aysolmaz, M. Leyer, and D. Iren, "Acceptance of AI for Delegating Emotional Intelligence: Results From an Experiment," in *Proceedings of the 54th Hawaii International Conference on System Sciences*, 2021, p. 6307.

[12] F. Quesque and Y. Rossetti, "What Do Theory-of-Mind Tasks Actually Measure? Theory and Practice," *Perspectives on Psychological Science*, vol. 15, no. 2, pp. 384–396, 2020.

[13] D. Premack and G. Woodruff, "Does the Chimpanzee Have a Theory of Mind?" *Behavioral and Brain Sciences*, vol. 1, no. 4, pp. 515–526, 1978.

[14] S. Baron-Cohen, A. M. Leslie, and U. Frith, "Does the Autistic Child Have a "Theory of Mind"?" *Cognition*, vol. 21, no. 1, pp. 37–46, 1985.

[15] W. Yoshida, R. J. Dolan, and K. J. Friston, "Game Theory of Mind," *PLoS Computational Biology*, vol. 4, no. 12, p. e1000254, 2008.

[16] M. E. Bodden, B. Mollenhauer, C. Trenkwalder, N. Cabanel, K. M. Eggert, M. M. Unger, W. H. Oertel, J. Kessler, R. Dodel, and E. Kalbe, "Affective and Cognitive Theory of Mind in Patients with Parkinson's Disease," *Parkinsonism & Related Disorders*, vol. 16, no. 7, pp. 466–470, 2010.

[17] S. R. Livingstone and W. F. Thompson, "The Emergence of Music From the Theory of Mind," *Musicae Scientiae*, vol. 13, no. 2_suppl, pp. 83–115, 2009.

[18] D. Bormann and T. Greitemeyer, "Immersed in Virtual Worlds and Minds: Effects of In-Game Storytelling on Immersion, Need Satisfaction, and Affective Theory of Mind," *Social Psychological and Personality Science*, vol. 6, no. 6, pp. 646–652, 2015.

[19] R. C. Mayer, J. H. Davis, and D. Schoorman, "An Integrative Model of Organizational Trust," *Academy of Management Review*, vol. 20, no. 3, pp. 709–734, 1995.

[20] K. J. Rotenberg, S. Petrocchi, F. Lecciso, and A. Marchetti, "The Relation Between Children's Trust Beliefs and Theory of Mind Abilities," *Infant and Child Development*, vol. 24, no. 2, pp. 206–214, 2015.

[21] C. DiYanni, D. Nini, W. Rheel, and A. Livelli, "'I Won't Trust You If I Think You're Trying to Deceive Me': Relations Between Selective Trust, Theory of Mind, and Imitation in Early Childhood," *Journal of Cognition and Development*, vol. 13, no. 3, pp. 354–371, 2012.

[22] C. Di Dio, F. Manzi, G. Peretti, A. Cangelosi, P. L. Harris, D. Massaro, and A. Marchetti, "Shall I Trust You? From Child–Robot Interaction to Trusting Relationships," *Frontiers in Psychology*, vol. 11, p. 469, 2020.

[23] E. Prochazkova, L. Prochazkova, M. R. Giffin, H. S. Scholte, C. K. De Dreu, and M. E. Kret, "Pupil Mimicry Promotes Trust Through the Theory-of-Mind Network," *Proceedings of the National Academy of Sciences*, vol. 115, no. 31, pp. E7265–E7274, 2018.

[24] R. Saxe, "Theory of Mind (Neural Basis)," *Encyclopedia of Consciousness*, vol. 2, pp. 401–410, 2009.

[25] L. C. Floren, D. Donesky, E. Whitaker, D. M. Irby, O. Ten Cate, and B. C. O'Brien, "Are We on the Same Page? Shared Mental Models to Support Clinical Teamwork Among Health Professions Learners: A Scoping Review," *Academic Medicine*, vol. 93, no. 3, pp. 498–509, 2018.

[26] N. Hanna and D. Richards, "The Impact of Multimodal Communication on a Shared Mental Model, Trust, and Commitment in Human–Intelligent Virtual Agent Teams," *Multimodal Technologies and Interaction*, vol. 2, no. 3, p. 48, 2018.

[27] M. Ruocco, W. Mou, A. Cangelosi, C. Jay, and D. Zanatto, "Theory of Mind Improves Human's Trust in an Iterative Human-Robot Game," in *The 9th International Conference on Human-Agent Interaction: HAI 2021*, 2021.

[28] J.-P. Stein and P. Ohler, "Venturing into the Uncanny Valley of Mind — the Influence of Mind Attribution on the Acceptance of Human-like Characters in a Virtual Reality Setting," *Cognition*, vol. 160, pp. 43–50, 2017.

[29] B. Grosz, S. Kraus, S. Talman, B. Stossel, and M. Havlin, "The Influence of Social Dependencies on Decision-Making: Initial Investigations with a New Game," in *Autonomous Agents and Multi-Agent Systems*, vol. 2, 2004, pp. 782–789.

[30] Y. Gal, B. J. Grosz, S. Kraus, A. Pfeffer, and S. Shieber, "Colored Trails: A Formalism for Investigating Decision-Making in Strategic Environments," in *Proceedings of the 2005 IJCAI Workshop on Reasoning, Representation, and Learning in Computer Games*, 2005, pp. 25–30.

[31] H. de Weerd, R. Verbrugge, and B. Verheij, "Negotiating With Other Minds: The Role of Recursive Theory of Mind in Negotiation With Incomplete Information," *Autonomous Agents and Multi-Agent Systems*, vol. 31, no. 2, pp. 250–287, 2017.

[32] S. Putze, D. Alexandrovsky, F. Putze, S. Höffner, J. D. Smeddinck, and R. Malaka, "Breaking the Experience: Effects of Questionnaires in VR User Studies," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–15.

[33] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an Empirically Determined Scale of Trust in Automated Systems," *International Journal of Cognitive Ergonomics*, vol. 4, no. 1, pp. 53–71, 2000.