Humanness lies in unpredictability: Role of Theory of Mind on anthropomorphism in human-computer interactions

Julia Ayache
julia.ayache@ntu.ac.uk
Nottingham Trent University
Nottingham, United Kingdom

Andrew M Connor
Auckland University of Technology
Auckland, New-Zealand

Stefan Marks
Auckland University of Technology
Auckland, New-Zealand

Alexander Sumich
Nottingham Trent University
Nottingham, United Kingdom
Auckland University of Technology
Auckland, New-Zealand

Nadja Heym
Nottingham Trent University
Nottingham, United Kingdom

ABSTRACT

Predictability is a core aspect of human-computer interaction (HCI), but an excess of predictability can lead to stereotypical behaviors and decreases the perceived agentivity (i.e., anthropomorphism) of a virtual agent. Yet, it remains unclear if inter-individual variability in predicting the behavior of other is modulating tendencies for anthropomorphism. The present study investigated the interaction between Theory of Mind (ToM) abilities and agent predictability and their association with anthropomorphism. Participants (N=38) completed self-reports of Perspective Taking capabilities and played the Matching Pennies Game with a virtual agent. The agent's predictability was manipulated across four conditions, and participants reported their perception of the agent's humanness and predictability. Results revealed that perceived predictability rather than self-reported Perspective Taking capabilities were positively associated with ToM abilities in predicting virtual agent's behavior. However, increasing the virtual agent's predictability decreased tendencies for anthropomorphism, stressing the role of randomness in perceiving humanness.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative interaction**.

KEYWORDS

Human-Computer-Interaction, Out of the Loop, Empathy, Anthropomorphism

ACM Reference Format:
Julia Ayache, Andrew M Connor, Stefan Marks, Alexander Sumich, and Nadja Heym. 2022. Humanness lies in unpredictability: Role of Theory of Mind on anthropomorphism in human-computer interactions. In *Proceedings of the 10th International Conference on Human-Agent Interaction (HAI '22), December 5–8, 2022, Christchurch, New Zealand.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3527188.3563920

HAI '22, December 5–8, 2022, Christchurch, New Zealand

ACM ISBN 978-1-4503-9323-2/22/12.
https://doi.org/10.1145/3527188.3563920

1 INTRODUCTION

The capacity to predict the behavior of a virtual agent is a core aspect of human-computer interaction (HCI). On the one hand, predictability is desirable and prevents out-of-the-loop problems by allowing humans to monitor and predict the activity of autonomous agents [3]. On the other, predictability can be detrimental by decreasing the perceived agentivity of autonomous agents displaying stereotypical behaviors [1]. Furthermore, the capacity for predicting a virtual agent's behavior varies from one individual to another, rendering it difficult to disentangle the interaction between the virtual agent's behavior and inter-individual variability on capacities for predicting behavior in HCI.

According to the Computer-As-Social-Actor paradigm (CASA), humans tend to apply social heuristics when interacting with virtual agents [9]. Therefore, predicting a virtual agent's behavior relies on a mechanism similar to that used in predicting human behaviors, such as Theory of Mind abilities (ToM) [5]. Associated with executive functions, ToM refers to the ability to "read" other's minds, encompassing capacities to take the perspective of others and therefore predict their behavior accurately [10]. Yet, ToM remains an ill-defined concept, rendering it challenging to understand its exact contribution to HCI [2].

According to the three-factor model of anthropomorphism, the need for predictability is one of the main factors leading to attributing human-like behavior to virtual agents [8]. Therefore, it is the desire for decreasing environmental uncertainty that fosters tendencies to detect human agentivity [7]. Furthermore, unpredictability is a marker of humanness as humans are often judged more prone to make mistakes, and less trustworthy than virtual agents [6]. Yet, despite the critical association between predictability, ToM abilities and anthropomorphism, their interactions need to be clarified in the context of HCI.

This study investigated the interaction between inter-individual variability in ToM abilities and an agent's predictability. The aim was to identify the potential interactions between perceived predictability, humanness and ToM abilities in predicting the decisions of a virtual agent. ToM abilities of participants were assessed through self-reports of Perspective Taking capabilities [4] and the

ability to predict accurately a virtual agent's behavior in a decision-making task, the Matching Pennies Game (MPG) [11]. The virtual agent's predictability was manipulated across different conditions. After each condition, participants completed self-reports assessing their perception of humanness and predictability of the virtual agent. Based on the established literature, the following hypotheses were formulated:

- **H1**: A positive association between participants' self-reported Perspective Taking capabilities and effective ToM abilities in predicting virtual agent's behavior during MPG.
- **H2**: A negative association between virtual agents' predictability and tendencies for anthropomorphism.

Additionally, the interaction between participants' self-reported Perspective Taking capabilities and virtual agents' predictability on perceived predictability, humanness and MPG scores was explored.

2 METHODS

2.1 Population

Participants (N = 38, 15 men, 21 women and 2 non-binaries, mean age = 22.13, SD= 4.23 years) were recruited from a student population enrolled at Nottingham Trent University.

2.2 Empathy questionnaire

Participants completed the Perspective Taking sub-scale from the **Interpersonal Reactivity Index (IRI)**, composed of 7 items, scored on a 5-point Likert scale, assessing capabilities to take on the perspective of others [4].

2.3 Matching Pennies Game

Participants played the Matching Pennies Game (MPG) adapted from [11] by guessing the virtual agent's choice (i.e., picking a blue or red card). n, The bias of the virtual agent was manipulated across four conditions using a within-subject design, by increasing the tendency to choose a specific card colour from 50% to 80% in increments of 10%. The level of bias was randomized within-subject. For each correct answer (i.e., predicting the card accurately that the virtual agent will pick up), the participant won 1 point.

2.4 Anthropomorphism and predictability

After each condition, participants rated their perception of humanness and predictability of the virtual partner using visual analog scales ranging from 0 to 100.

2.5 Data analyses

Participants' ratings were standardized participant-wise for perceived humanness, predictability and MPG score, and variable-wise for scores obtained on the subscale of *Perspective Taking*. Spearman's correlations computed the association between the subscale *Perspective Taking*, perceived predictability, humanness and MPG scores. Mixed models tested the interaction of the experimental condition and the subscale *Perspective Taking*, on perceived predictability, humanness and MPG scores. Dataframe and script of analyses are available here: https://osf.io/wvs9p/

3 RESULTS

Perceived predictability displayed a positive and significant association with MPG scores (r (36)=.53, p < .001). The perceived predictability was higher in the 80 % bias condition (mean=0.37, SD= 1.02), with (b= 1.93, SE= 0.96, p= .05) - see **Figure 1 (left)**. The perceived humanness was lower in the 80 % bias condition (mean=-0.39, SD= 0.77), with (b= 1.93, SE= 0.96, p= .05) - see **Figure 1 (right)**. However, no significant associations or interactions with the experimental conditions were observed between the scores obtained on the sub-scale *Perspective Taking* and perceived predictability, humanness and MPG scores.

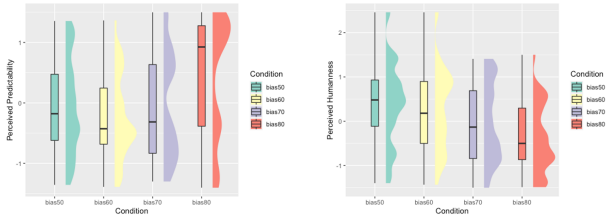

Figure 1: Raincloudplot for perceived predictability (left) and perceived humanness (right) and experimental conditions.

4 DISCUSSION

This study investigated the association and interaction between intra-individual ToM abilities and a virtual agent's predictability to disentangle their respective contribution to their perception of the virtual agent. Contrary to **H1**, self-reported Perspective Taking capabilities were not associated with effective ToM abilities in predicting the virtual agent's choice. Future investigations are required to clarify the exact contribution of Perspective Taking capabilities to ToM abilities, especially in HCI context. In line with **H2**, a higher degree of predictability decreased tendencies for anthropomorphism. In line with the three-factor model of anthropomorphism, this result stresses the role of uncertainty in attributing human agentivity. Interestingly, there was no association between perceived predictability and anthropomorphism, suggesting that participants might not consciously rely on this heuristic. Finally, there was no interaction between participants' self-reported Perspective Taking capabilities and virtual agents' predictability on perceived predictability, humanness and MPG scores.

5 CONCLUSION

Altogether, these findings suggest that perceived predictability rather than self-reports of Perspective Taking capabilities is a better marker of effective ToM abilities in predicting a virtual agent's behavior. Furthermore, this study reveals a paradoxical association between predictability and anthropomorphism, two core aspects in designing HCI. Future investigations are required to explore the role of ToM abilities in adopting specific game strategies and identify neurophysiological markers of perceived predictability associated with anthropomorphism.

ACKNOWLEDGMENTS

This research is supported by the Doctoral Alliance Training co-funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement [801604].

REFERENCES

[1] Jeremy N. Bailenson, Nick Yee, Kayur Patel, and Andrew C. Beall. 2008. Detecting digital chameleons. *Computers in Human Behavior* 24, 1 (2008), 66–87. https://doi.org/10.1016/j.chb.2007.01.015

[2] Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H. Beauchamp. 2020. Systematic Review and Inventory of Theory of Mind Measures for Young Children. *Frontiers in Psychology* 10 (2020), 2905. https://doi.org/10.3389/fpsyg.2019.02905

[3] Bruno Berberian, Bertille Somon, Aïsha Sahaï, and Jonas Gouraud. 2017. The out-of-the-loop Brain: A neuroergonomic approach of the human automation interaction. *Annual Reviews in Control* 44 (2017), 303–315. https://doi.org/10.1016/j.arcontrol.2017.09.010

[4] Mark H. Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology* 44 (1983), 113–126. https://doi.org/10.1037/0022-3514.44.1.113 Place: US Publisher: American Psychological Association.

[5] Sandra Devin and Rachid Alami. 2016. An implemented theory of mind to improve human-robot shared plans execution. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Christchurch, New Zealand, 319–326. https://doi.org/10.1109/HRI.2016.7451768

[6] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58, 6 (June 2003), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

[7] Michiel van Elk. 2013. Paranormal believers are more prone to illusory agency detection than skeptics. *Consciousness and Cognition* 22, 3 (Sept. 2013), 1041–1046. https://doi.org/10.1016/j.concog.2013.07.004

[8] Nicholas Epley, Adam Waytz, and John T. Cacioppo. 2007. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review* 114 (2007), 864–886. https://doi.org/10.1037/0033-295X.114.4.864 Place: US Publisher: American Psychological Association.

[9] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. Machinery, Boston Massachusetts USA, 72–78.

[10] Josef Perner and Birgit Lang. 2000. Theory of mind and executive function: Is there a developmental relationship? In *Understanding other minds: Perspectives from developmental cognitive neuroscience, 2nd ed.* Oxford University Press, New York, NY, US, 150–181.

[11] Peter T. Waade, Kenneth C. Enevoldsen, Arnault-Quentin Vermillet, Arndis Simonsen, and Riccardo Fusaroli. 2022. Introducing tomsup: Theory of mind simulations using Python. *Behavior Research Methods* 2022 (Aug. 2022), 1–35. https://doi.org/10.3758/s13428-022-01827-2