Proposal for a Special Session at IEEE RO-MAN 2024
Mind Attribution in HRI: Determinants and Consequences

Maartje de Graaf[1], Alessandra Rossi[2], Mariacarla Staffa[3] and Angelo Cangelosi[4]

***Abstract*— Mind attribution refers to "the cognitive capacity to reflect upon one's own and other persons' mental states such as beliefs, desires, feelings and intentions". A growing group of Human-Robot Interaction (HRI) research recently focuses on investigating whether people form mental models towards robots, and how robots are able to mentalize. However, there is currently no consensus about what types of data qualify as proof of mind attribution to robots, and the ability to reason about false beliefs can improve the quality of the HRI, including transparency of robots' behaviors, naturalness of the communication, etc. To tackle down these challenges, in this special session, we want to start by (1) fostering a shared terminology used to denote mind attribution, (2) unraveling determinants and consequences of mind attribution in HRI, and (3) sharing best practices in methods used to study mind attribution to robots in terms of stimulus materials and measures.**

I. Title

Mind Attribution in HRI: Determinants and Consequences.

II. Aim and Scope of the Special Session

Mind attribution refers to "the cognitive capacity to reflect upon one's own and other persons' mental states such as beliefs, desires, feelings and intentions" [1]. Social robots invite people to attribute mind [2] because these robots exhibit more complex social cues [3], suggesting autonomy (e.g., approaching, pointing, and asking inquiries) and mental activity (e.g., eye movements for attention, memory retrieval, and task-planning). These social cues trigger people to perceive robots as intentional social agents [4]. How useful mind attribution is in the context of human-robot interaction is still largely unknown. Why, rather than underlying computational or physical facts, do we occasionally attribute robot behavior to mental states? How can we trust ascribed mental states to predict a robot's behavior? There is currently no consensus about what types of data qualify as proof of mind attribution to robots and, consequently, researcher employ a wide variety of methods, resulting in contradictory findings in the literature [5]. Even though this early stage of research is likely to benefit from the interdisciplinary diversity, to access and build upon each other's work, researchers are finding it ever more necessary to establish a common language and basic assumptions about the phenomenon they are studying. In this session, we aim at looking at identifying and further discussing such requirements to define and attribute mental models to robots.

According to several authors (e.g., [6], [7]), social robots are often designed as deceptive. On one side, these argue that any technique allowing robots to have human-like or social behaviors is a form of deception. On another side, researchers are also investigating different ways to design deceptive behaviors for social robots to help in behavioural changes [8] and compliance with instructions, such as medical and health prescriptions [9]. In particular, in such cases, an important aspect to consider is the "false belief understanding", which is people's ability to infer when others have beliefs that contradict the reality [10]. Previous studies have shown that people who are not able to recognize the robot's attempt of deceiving them, do not believe that the robot could or have intention to deceive them [11]. In HRI, the presence of possible false beliefs can result in communication breakdowns or misunderstandings regarding the actions of the robot, and ultimately a mismatch between user expectations and robot behaviour can alter the perception of trust, jeopardizing the success of the interaction. In this session, we aim to look at how the ability to reason about false beliefs, including the understanding of the robot's intentionality of deceiving people, may reduce the miscommunication, and which are the implications on people's trust and acceptance of robots' ability to look after their wellbeing.

A contemporary literature review on mental state attribution to robots [5] indicate a research gap in the (behavioral) consequences of such attributions with only 15% of published research papers reporting such findings. Investigating the relationship between people's comprehension of a robot's mind and how mind attributions affect subsequent human-robot interactions is crucial to improve people's understanding of robots and establish beneficial interaction outcomes. Indeed, as such insights may be utilized to improve human-robot interactions through robot design that facilitates the attribution of "correct" mental states, investigating how mind attribution influences people's interactions with robots is essential. Moreover, the majority of HRI research examining mental state attribution use survey-based methods [5], [12]. However, people are frequently unconscious of these automatic inclinations with artificial agents [13] and with other humans [14]. Mentalizing tendencies originate from subconscious processes and implicit social-cognitive inferences [15]. People may assert, for instance, that they do not believe robots to have minds, even though behavioral measurements suggest otherwise [12]. Thus, merely depending on explicit

[1]Maartje de Graaf is with Utrecht University, Netherlands m.m.a.degraaf@uu.nl

[2]Alessandra Rossi is with the University of Naples Federico II, Italy alessandra.rossi@unina.it

[3]Mariacarla Staffa is with the University of Naples Parthenope, Italy mariacarla.staffa@uniparthenope.it

[4]Angelo Cangelosi is with the University of Manchester, UK angelo.cangelosi@manchester.ac.uk

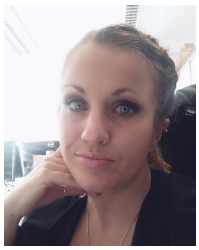
Proposal 7 submitted to 2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). Received January 19, 2024.

survey-based metrics would not be appropriate when examining the attribution of mental states to robots [16], [17]. In this session, we aim to cover (behavioral) consequences of mind attribution to robots given that people's predictions and explanations of robot behavior –as well as their choices about how to interact with them– are driven by the attribution of particular mental states (beliefs, desires, intentions, and so forth).

The topics covered in this special session are in line with the main theme of the conference (i.e., "Embracing the human-centered HRI"). In particular, we want to start by (1) fostering a shared terminology used to denote mind attribution, (2) unraveling determinants and consequences of mind attribution in HRI, and (3) sharing best practices in methods used to study mind attribution to robots in terms of stimulus materials and measures. Notably, accepted topics include, but are not limited to:

- Explainable AI (XAI) in HRI
- Multi-modal situation awareness and spatial cognition
- Social intelligence for robots in interactive and non-interactive tasks
- Verifications Methods for autonomous agents
- Legibility, Predictability and Transparency in HRI
- Cognitive robotics
- Deception in HRI
- Robot cheating in HRI
- Theory of Mind, Mental models in HRI
- Robot etiquette and social norms
- Modelling Trust and Acceptance in HRI

III. Organisers



Alessandra Rossi, Assistant Professor

Affiliation: University of Naples Federico II, Italy.

Email: alessandra.rossi@unina.it

Phone: +39 081679961

Bio: Alessandra is Assistant Professor at the University of Naples "Federico II", Italy. Her PhD thesis was part of the Marie Sklodowska-Curie Research ETN SECURE project at the University of Hertfordshire (UK). She is also a Visiting Lecturer at University of Hertfordshire. Her research interests include Human–(Multi) Robot Interaction, social robotics, trust, XAI, multi-agent systems and user profiling. Alessandra is Project Manager of Marie Sklodowska-Curie Research ETN PERSEO. She has been Publicity chair at IEEE RO-MAN 2022 and 2023, Virtual Organizing Chair of IEEE RO-MAN 2021, Registration Chair and Social Media Responsible for IEEE RO-MAN 2020. Alessandra has great experience in organising scientific events has main organsiser and in collaboration with her peers, some examples are the workshops SCRITA at RO-MAN 2018-2023, workshop TRAITS at HRI 2021 and 2022, special sessions at RO-MAN 2019-2023, special issues at international journals (e.g., International Journal of Social Robotics, and Paladyn Journal of Behavioral Robotics).

Maartje de Graaf, Assistant Professor

Affiliation: Utrecht University, Netherlands.

Email: m.m.a.degraaf@uu.nl

Phone: +31 648512017

Bio: Maartje is Assistant Professor of Human-Computer Interaction at Utrecht University, Netherlands. Her research focuses on people's affective, behavioural, and cognitive responses to robots, including topics of moral agency, norm violations, explainability, and communication strategies for trust-repair. She obtained her PhD in Communication Science and Human-Robot Interaction (2015, Twente University, Netherlands) investigating the long-term acceptance of social robots in home environments. She is a member of the HRI Steering Committee, has been Workshop Chair at HRI 2020 and ICSR 2021, Program Committee Track Chair at HRI 2023-2025, Publicity Chair at HRI 2025, is an Associate Editor of THRI, and has co-organized 20+ workshops at HRI, RO-MAN, and ICSR, including workshops on Mental Models in HRI at HRI 2020 as well as Explainability in Robotics at HRI 2019 and 2024 and RO-MAN 2023.

Mariacarla Staffa, Assistant Professor

Affiliation: University of Naples Parthenope.

Email: mariacarla.staffa@uniparthenope.it

Phone: +31 0815476580

Bio: Mariacarla Staffa (F) is an Assistant Professor in Human-Computer/Robot Interaction, Artificial Intelligence and Cognitive Robotics at the Department of Science and Technologies of the University of Naples Parthenope, Italy. She received the M.Sc. degree in Computer Science from the University Federico II with honors, in 2008. She got a Ph.D. in Computer Science and Automation Engineering from the University Federico II in 2011. She was a visiting researcher at the "Institute de Système Intelligentes et de Robotique" at the University of Paris "Pierre et Marie Curie". She is a senior member of the Institute of Electrical and Electronics Engineers (IEEE) and of the EUCognition - European Society for Cognitive Systems (ID: 2037). She is part of the IEEE RAS Technical Committee for Cognitive Robotics. She serves as Expert Reviewer for the European Commission Framework Programme for Research and Innovation (Area: AI and Robotics). She is the Coordinator of the BRAIN Laboratory (the "Bioinspired Robotics and Artificial intelligence Networking Lab" of the University of Naples Parthenope) working in the fields of Cognitive Robotics, Artificial Intelligence and Social and Assistive Robotics. She authored several works on Social Assistive Robots, Adaptive Human Robot Interaction, Human Behavior and Emotion interpretation, etc. She is the Principal Scientific Coordinator

of the Project RESTART - Robot Enhanced Social abilities based on Theory of mind for Acceptance of Robot in assistive Treatments and Unit Scientifc Coordinator of the SPECTRA Project (Supporting schizophrenia PatiEnts' Care wiTh Robotics and Artificial Intelligence) both funded by the Italian Ministry of University and Research.

Angelo Cangelosi, Full Professor

Affiliation: University of Manchester, United Kingdom.

Email: angelo.cangelosi@manchester.ac.uk

Phone: +44 (0)7766134202

Bio: Angelo Cangelosi is Professor of Machine Learning and Robotics at the University of Manchester (UK) and co-director and founder of the Manchester Centre for Robotics and AI. He was selected for the award of the European Research Council (ERC) Advanced grant (UKRI funded). His research interests are in cognitive and developmental robotics, neural networks, language grounding, human robot-interaction and trust, and robot companions for health and social care. Overall, he has secured over £40m of research grants as coordinator/PI, including the ERC Advanced eTALK, the UKRI TAS Trust Node and CRADLE Prosperity, the US AFRL project THRIVE++, and numerous Horizon and MSCAs grants. Cangelosi has produced more than 300 scientific publications. He is Editor-in-Chief of the journals Interaction Studies and IET Cognitive Computation and Systems, and in 2015 was Editor-in-Chief of IEEE Transactions on Autonomous Development. He has chaired numerous international conferences, including ICANN2022 Bristol, and ICDL2021 Beijing. His book "Developmental Robotics: From Babies to Robots" (MIT Press) was published in January 2015, and translated in Chinese and Japanese. His latest book "Cognitive Robotics" (MIT Press), coedited with Minoru Asada, was recently published in 2022.

IV. Tentative Speakers

We commit to promote and increase the visibility of the session through the most popular used channels to reach the appropriate audience, such as robotics mailing-lists and directly inviting leading researchers in the fields. We expect submissions from experts in the fields of cognitive and behavioural robotics, autonomous agents systems, and social HRI. In particular, we prospect submissions from representatives of the above-mentioned fields such as the following:

- Agnieszka Wykowska, Italian Institute of Technology, Italy
- Barbara Müller, Radboud University Nijmegen, The Netherlands
- Bertram Malle, Brown University, USA
- Elizabeth Phillips, George Mason University, USA
- Emily Cross, ETH Zurich, Switzerland
- Eva Wiese, Berlin Institute of Technology, Germany
- Jaime Banks, Syracuse University, USA
- Matthew Rueben, University of Portland, USA
- Sam Thellman, Linköping University, Sweden
- Brian Scassellati, Yale University
- Kerstin Fisher, University of Southern Denmark
- Helen Hastie, Heriot-Watt University, UK
- Alessandra Rossi, University of Naples Federico II, Italy
- Rachid Alami, CNRS, Univ de Toulouse, LAAS, France
- Antonio Andriella - Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Spain

ACKNOWLEDGMENT

This work has been supported by Italian PON R&I 2014-2020 - REACT-EU Azione IV.4 (CUP E65F21002920003).

References

[1] M. Brüne, M. Abdel-Hamid, C. Lehmkämper, and C. Sonntag, "Mental state attribution, neurocognitive functioning, and psychopathology: what predicts poor social competence in schizophrenia best?" *Schizophrenia research*, vol. 92, no. 1-3, pp. 151–159, 2007.

[2] M. Kwon, M. F. Jung, and R. A. Knepper, "Human expectations of social robots," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 463–464.

[3] S. C. Johnson, "The recognition of mentalistic agents in infancy," *Trends in Cognitive Sciences*, vol. 4, no. 1, pp. 22–28, 2000.

[4] L. Floridi and J. W. Sanders, "On the morality of artificial agents," *Minds and machines*, vol. 14, pp. 349–379, 2004.

[5] S. Thellman, M. de Graaf, and T. Ziemke, "Mental state attribution to robots: A systematic review of conceptions, methods, and findings," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 11, no. 4, pp. 1–51, 2022.

[6] A. MATTHIAS, "Robot lies in health care : when is deception morally permissible?" *Kennedy Institute of Ethics Journal*, vol. 25, no. 2, pp. 169–192, Jun. 2015.

[7] R. Sparrow and L. Sparrow, "In the hands of machines? the future of aged care," *Minds and Machines*, vol. 16, pp. 141–161, 05 2006.

[8] S. Rossi, S. J. Santini, D. Di Genova, G. Maggi, A. Verrotti, G. Farello, R. Romualdi, A. Alisi, A. E. Tozzi, and C. Balsano, "Using the social robot nao for emotional support to children at a pediatric emergency department: Randomized clinical trial," *J Med Internet Res*, vol. 24, no. 1, p. e29656, Jan 2022.

[9] N. Lee, J. Kim, E. Kim, and O. Kwon, "The influence of politeness behavior on user compliance with social robots in a healthcare service setting," *International Journal of Social Robotics*, vol. 9, no. 5, pp. 727–743, 2017.

[10] 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.

[11] A. Rossi and S. Rossi, "Evaluating people's perception of trust of a deceptive robot with theory of mind in an assistive gaming scenario," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2023, pp. 18–23.

[12] J. Banks, "Theory of mind in social robots: Replication of five established human tests," *International Journal of Social Robotics*, vol. 12, no. 2, pp. 403–414, 2020.

[13] C. Nass and Y. Moon, "Machines and mindlessness: Social responses to computers," *Journal of Social Issues*, vol. 56, no. 1, pp. 81–103, 2000.

[14] J. Low and J. Perner, "Implicit and explicit theory of mind: State of the art," pp. 1–13, 2012.

[15] A. G. Greenwald and M. R. Banaji, "Implicit social cognition: attitudes, self-esteem, and stereotypes," *Psychological Review*, vol. 102, no. 1, p. 4, 1995.

[16] M. M. A. de Graaf, S. B. Allouch, and S. Lutfi, "What are people's associations of domestic robots? comparing implicit and explicit measures," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016, pp. 1077–1083.

[17] S. Thellman, A. Giagtzidou, A. Silvervarg, and T. Ziemke, "An implicit, non-verbal measure of belief attribution to robots," in *2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2020, pp. 473–475.