The challenges of first- and second-order belief reasoning in explainable human-robot interaction

Sam Thellman[1], and Maartje M.A. de Graaf[2]

Abstract— **Current approaches to implement eXplainable Autonomous Robots (XAR) are dominantly based on Reinforcement Learning (RL), which are suitable for modelling and correcting people's first-order mental state attributions to robots. Our recent findings show that people also rely on attributing second-order beliefs (i.e., beliefs about beliefs) to robots to interpret their behavior. However, robots arguably form and act primarily on first-order beliefs and desires (about things in the environment) and do not have a functional "theory of mind". Moreover, RL models may be incapable to appropriately address second-order belief attribution errors. This paper aims to open a discussion of what our recent findings on second-order mental state attribution to robots imply for current approaches to XAR.**

I. Introduction

Robots are progressively spreading in everyday society, incorporating roles in healthcare, education, and personal and public services. To benefit human users, a robot's decisions, recommendations, and actions need to be intuitively interpretable and understandable. Psychological research indicates that people cannot help but infer mind and intentional agency in other agents (e.g., [25]), including robots [14], [26], [42], [57]. Since people have learned to construct mental models based on their interaction with other living beings [25], they run the risk of establishing incorrect mental models of robots [56]. As a result, people perceive robot behaviors as ambiguous [14], [27], wrongfully blame robots for alleged errors [27], improperly trust robots' abilities [29], [41], and collaborate ineffectively with robots [12], [54]. The increasingly complex but unintelligible algorithms that govern the behavior of robots invoke the need for interpretable robot behaviors that offer meaningful insights into their decisions, recommendations, and actions [18], [44], [52]. However, current research on the explainability of intelligent autonomous systems is not only limited [28] but lacks theoretical foundations and rigorous research methods [9], [34], [50]. In our view, the challenge is to design human interactions with such systems in such a way that people attribute appropriate intentional (mental) states to them [50] –a challenge that prompts researchers across several scientific disciplines to advocate for a multidisciplinary approach on explainability [9], [34], [37]. This paper provides an overview of psychological findings on how people perceive robot minds and how these insights can benefit the design of explainable human-robot interactions (HRIs).

II. How People Perceive Robot Minds

A vast body of research in philosophy [43], psychology [20], and social cognition [30] illuminates how people define, generate, select, evaluate, and present explanations of behavior in general. People understand and explain their own and others' behaviors by the ascription of intentional states to agents in social contexts [11]. When explaining behavior, people use a sophisticated set of distinctions, which follow directly from their folk concept of intentionality and specify the constituent components of (un)intentional behavior [31].

People seek behavior explanations for two reasons [32]: (1) to find meanings in behavior when they themselves wonder why the behavior occurred, and (2) to manage social interaction when they expect others to wonder why the behavior occurred. This "wondering why" may be implicit. When explaining any given behavior, be it our own or that of someone or something else, people choose from a number of explanatory tools that occur at three levels [30]: *mode* (cause, reasons, causal-history of reasons, enabling factors), *type* (either type of reason, e.g., belief versus desire, or type of cause, e.g., trait versus non-trait), and *linguistic form* (e.g., reasons with or without mental state markers).

Work in numerous disciplines has shown that explanations of human behavior are fundamentally grounded in a conceptual framework of intentional agency and mind, typically referred to as *theory of mind* [3], [38] or *folk psychology* [19], [23]. This framework guides people's explanations and predictions of behavior as well as inferences about specific mental states, such as specific beliefs, desires, or intentions [8]. A substantial body of research indicates that people ascribe mental states to robots [14], [26], [42], [57] (for a review, see [48]), and further evidence suggests that people spontaneously apply the same explanatory tools of folk psychology when explaining robot behavior as they apply to human behavior [10], [49]. To increase the predictability and explainability of robots, human-robot interactions must be designed in a way that facilitates the attribution of behavior-congruent intentional states to robotic systems (i.e., design that helps people make useful inferences about what robots *want*, *know*, *intend*, etc.) [47], [50].

III. Folk-Psychological Approaches to Explainability

Current approaches toward explainable autonomous intelligent systems lack theoretical foundations and rigorous methodological evaluations [1]. They mainly focus on making the robot's algorithms more transparent (e.g., [27], [36], [51], [56]), often based on researchers' own intuition of

[1]Department of Computer and Information Science, Linköping University, SE-581 83 Linköping, Sweden sam.thellman@liu.se,

[2]Department of Information and Computing Sciences, Utrecht University, 3584 CC Utrecht, The Netherlands m.m.a.degraaf@uu.nl,

what constitutes transparency or interpretability [34]. More importantly, these approaches typically fail to consider that human interpretations of behavior are shaped within a folk-conceptual framework and that explanations must be phrased in the language of this framework to create the meaning, understanding, and trust that people seek within interactions [9], [34], [50]. Given that previous research indicates that people readily apply the concepts and linguistic tools of folk psychology to explain robot behavior [10], [49], it stands to reason that people will be comfortable when robots explain their own behavior using this framework. Therefore, we have previously proposed that robots should explain their own intentional actions (planned or already performed) by reference to reasons for their actions and using the language of beliefs and desires that is so strongly present in human interactions [9], [10], [50]. Here, we highlight several open challenges for explainable HRIs mainly based on our previous work:

- *The Perceptual Belief Problem:* Robots perceive the world differently from humans and people may not readily understand their perspective on the world [50]. References to objects and events in the shared physical space must therefore be made in terms that make referents identifiable to humans when they occur as part of behavior explanations.
- *The Language of Explanation:* The behavior explanations people expect of robots will have to be phrased in the language familiar to people as communicators and audiences of ordinary explanations [9] (as opposed to, for example, being communicated in the form of an algorithmic decision tree or other types of information structures that are important for software and hardware engineers in terms of traceability and verification [7], [15]). Research shows that people prefer explanations that are compatible with the interpretative framework of folk psychology –that is, explanations referring to beliefs, desires and other mental states that motivated their decisions.
- *Distinct Classes of Behavior, Distinct Explanations:* Any robot that explains its behavior the manner that people expect it to should possess two cognitive abilities that are fundamental to folk psychological reasoning. First, the robot must be able to distinguish intentional actions from accidental events (at least in itself but, ideally, also in other agents). Second, the robot must be able to explain each of these classes of behaviors in reference to the expected *type* of explanation – unintentional behaviors to (mere) causes, intentional behaviors to reasons.
- *Selecting Relevant Explanations:* People do not want to hear a complete account of all the beliefs, goals, sub-goals, or rejected actions that a system tracked [39], [40]. This shows that in addition to expecting explanations to address the relevant level in the hierarchy of capabilities, people will also choose and assess explanations based on some sort of relevance criterion. In psychology, this is sometimes called the "causal selection problem" –the difficulty of selecting a small number of causes/reasons that sufficiently explain a particular event. For example, when people ask about why a certain action was taken, they often actually mean to ask why some other action was not [35]. This counterfactual is typically not explicitly articulated, yet understood by humans. The type of explanation that people would like to hear from a robot might further depend on whether the behavior was expected or not [39]. Together, these features enable a *common ground* on the basis of which explanations can be formulated (see below).
- *Explanation Timing:* Explanations may serve different pragmatic goals: audience design [5], [22], which is tailoring to the assumptions, knowledge, and specific interests that an audience has when decoding the explanation [16], [24]; and impression management [4], [17], where the goal of the explainer is not merely to optimize communication but also to influence the audience's perceptions and evaluations [33]. When to provide an explanation may depend on the communicative purpose of the given explanation, and wrongly timing the provided explanation may negatively impact people's evaluations of the robot [46].
- *From Structure to Content:* Simply knowing that an agent has some desires that could (possibly) be satisfied by a particular action does not adequately explain that action; understanding an action critically involves being able to identify what specific desire gave rise to that action. A broad knowledge structure of associative, social, and causal linkages is necessary for the right interaction between the contents of beliefs, desires, and behaviors –the dreaded "common sense" [13], [45].
- *Common Ground of Explanation:* Although people in interactions typically use implicit means of negotiating what knowledge, skills, and information needs they have and what their partners can rely on [6], it remains unclear how common ground (i.e., the mutual knowledge, beliefs, and assumptions that partners in a conversation rely on in order to communicate and interact efficiently) may be reached in HRIs and, as a result, how the criteria for a successful explanation can be satisfied given that these implicit mechanisms only partially work in HRIs.

IV. Implications of Second-Order Belief Reasoning for Explainable Robotics

The scientific literature on people's perceptions of robot minds has so far focused primarily on attributions of first-order mental states, i.e., mental states, such as beliefs and desires, that are oriented toward physical states of affairs in the world (e.g., that Pepper *believes* the ball is under the red cup (see Figure 1) In attributing first-order mental states to robots, people must consider what robots are likely to want and know given their physical environment. Current approaches to implement XAR are dominantly based on RL [55], which are for the most part suitable for modelling and correcting people's first-order mental state attributions

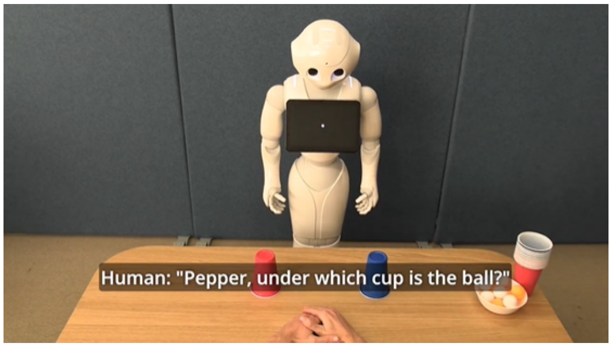

Fig. 1. Location-change task employed in our recent experiments.

to robots (which is instrumental to address many of the challenges listed above). However, our recent findings show that people also attribute second-order mental states to robots, that is, mental states that are about the mental states of others. Reasoning about the second-order mental states of robots require people to consider what robots know about the mental states of themselves or others (e.g., whether a robot *understands* that a human *wishes* to deceive it).

We recently tested whether people attribute second-order beliefs to social robots and how this affects the way in which they interact with such robots. In an online survey experiment, participants (n = 155) watched a video in which a human attempted to deceive a Pepper robot in a location change task (i.e., by moving a ball from underneath a blue cup to a red cup when the robot is purposely distracted and looking the other way, Figure 1). Our results show that 19% of the participants attributed the robot's behavior to a second-order belief about being deceived by the human. This finding contributes to the preexisting literature on mental state attribution to robots [10], [14], [26], [42], [49], [57], which so far only focused on the attribution of first-order mental states. It is an important contribution considering that robots arguably form and act primarily on first-order beliefs and desires (about things in the environment) and do not have a functional "theory of mind". Hence, the results suggest that, with respect to holding second-order mental states, people may expect more of robots than they can "deliver". We also found that participants who ascribed the second-order belief to the robot (about being intentionally deceived) were less willing to accept offers made by the robot that were considered as unfair in the context of a resource allocation negotiation task. Importantly, this finding suggests a link between people's second-order mental state attributions and how they interact with robots.

These recent results suggest that robots need to provide explanations or other types of interaction-managing interventions to support people in attributing second-order mental states to robots. However, the dominantly used RL techniques in XAR [55] may be unsuitable to address people's second-order belief attribution errors. Although some attempts of computational cognitive modeling of second-order false belief reasoning have been presented in the literature (e.g., [21], [53]), Arslan et al. [2] advocate the use of Instance-Based Learning (IBL) models to update incorrect mental models of the second-order kind instead. Hence, RL-based robotic systems might not be able to cope with human misinterpretation of their second-order beliefs and thus might not be able to provide relevant behavior explanations (either proactively or reactively) to support such reasoning. Although both RL and IBL models strengthen or modify their techniques based on experience and the feedback "Correct/Wrong" without further explanation, the way each handle this feedback differentiates them from one another [2]. An IBL model adds an instance of a different strategy whereas a RL model penalizes the methods that result in an incorrect response, given that the strategy selection is explicit in the IBL model while it is implicit in the RL model. Additionally, because it explicitly raises the level of mental state attribution to a higher level in response to feedback that includes additional explanations, the IBL model is more likely to employ a second-order mental state attribution strategy. The RL model, on the other hand, seems helpless in the face of such further explanations.

V. Conclusion

The increasingly complex but unintelligible algorithms that underlie the behavior of robots require us to build interpretable robot behaviors that offer meaningful insights into their decisions, recommendations, and actions [18], [44], [52]. The challenge is to design human-robot interactions that facilitate the attribution of behavior-congruent intentional states to such systems [47], [50], which prompts numerous researchers to advocate for a multidisciplinary approach on explainability [9], [34], [37]. Our recent findings suggest that people attribute not only first-order but also second-order beliefs to social robots, which, in turn, affects the way in which people interact with robots. More importantly, these findings imply that the dominantly used RL models to address XAR [55] may be unsuitable as they are argued to be incapable to appropriately address second-order belief attribution errors [2]. This paper aims to open a discussion about the implications of these findings for XAR.

References

[1] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," in *AAMAS 2019, Montreal, Canada, May 13–17*, 2019, pp. 1078–1088.

[2] B. Arslan, N. A. Taatgen, and R. Verbrugge, "Five-year-olds' systematic errors in second-order false belief tasks are due to first-order theory of mind strategy selection: A computational modeling study," *Frontiers in Psychology*, vol. 8, p. 275, 2017.

[3] S. Baron-Cohen, "Without a theory of mind one cannot participate in a conversation," *Cognition*, vol. 29, no. 1, pp. 83–84, 1988.

[4] H. Barrett, *Maintaining the Self in Communication: Concept and Guidebook*. Alpha & Omega Book Pub, 1998.

[5] S. Bromberger, "An approach to explanation," *Analytical Philosophy*, vol. 2, pp. 72–105, 1965.

[6] H. H. Clark, *Using language*. Cambridge University Press, 1996.

[7] J. Cleland-Huang, O. Gotel, A. Zisman *et al.*, *Software and systems traceability*. Springer, 2012, vol. 2, no. 3.

[8] R. d'Andrade, "A folk model of the mind," *Cultural models in language and thought*, 1987.

[9] M. M. A. De Graaf and B. F. Malle, "How people explain action (and ais should too)," in *2017 AAAI Fall Symposium Series*, 2017.

[10] ——, "People's explanations of robot behavior subtly reveal mental state inferences," in *HRI 2019*. IEEE, 2019, pp. 239–248.

[11] D. C. Dennett, *The intentional stance*. MIT press, 1989.

[12] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *HRI 2013*. IEEE, 2013, pp. 301–308.

[13] H. L. Dreyfus, *What computers still can't do: A critique of artificial reason*. MIT Press, 1992.

[14] F. Eyssel, D. Kuchenbrandt, and S. Bobinger, "Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism," in *HRI 2011*, 2011, pp. 61–68.

[15] M. Fisher, L. Dennis, and M. Webster, "Verifying autonomous systems," *Communications of the ACM*, vol. 56, no. 9, pp. 84–93, 2013.

[16] S. R. Fussell and R. M. Krauss, "Coordination of knowledge in communication: Effects of speakers' assumptions about what others know." *Journal of Personality and Social Psychology*, vol. 62, no. 3, p. 378, 1992.

[17] E. Goffman *et al.*, *The presentation of self in everyday life*. Harmondsworth London, 1959.

[18] B. Goodman and S. Flaxman, "Eu regulations on algorithmic decision-making and a "right to explanation"," in *ICML workshop on human interpretability in machine learning (WHI 2016), New York, NY. http://arxiv. org/abs/1606.08813 v1*, 2016.

[19] J. D. Greenwood, *The future of folk psychology: Intentionality and cognitive science*. Cambridge University Press, 1991.

[20] F. Heider, *The psychology of interpersonal relations*. Wiley, 1958.

[21] L. M. Hiatt and J. G. Trafton, "Understanding second-order theory of mind," in *HRI 2015 Companion*, 2015, pp. 167–168.

[22] D. J. Hilton, "Conversational processes and causal explanation." *Psychological Bulletin*, vol. 107, no. 1, p. 65, 1990.

[23] T. Horgan and J. Woodward, "Folk psychology is here to stay," *The Philosophical Review*, vol. 94, no. 2, pp. 197–226, 1985.

[24] W. S. Horton and R. J. Gerrig, "Speakers' experiences and audience design: Knowing when and knowing how to adjust utterances to addressees," *Journal of Memory and Language*, vol. 47, no. 4, pp. 589–606, 2002.

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

[26] S. Kiesler and J. Goetz, "Mental models of robotic assistants," in *CHI 2002 Extended Abstracts*, 2002, pp. 576–577.

[27] T. Kim and P. Hinds, "Who should i blame? effects of autonomy and transparency on attributions in human-robot interaction," in *ROMAN 2006*. IEEE, 2006, pp. 80–85.

[28] P. Langley, B. Meadows, M. Sridharan, and D. Choi, "Explainable agency for intelligent autonomous systems," in *Twenty-Ninth IAAI Conference*. AAAI, 2017, pp. 4762–4764.

[29] M. Lomas, R. Chevalier, E. V. Cross, R. C. Garrett, J. Hoare, and M. Kopack, "Explaining robot actions," in *HRI 2012*, 2012, pp. 187–188.

[30] B. F. Malle, "How people explain behavior: A new theoretical framework," *Personality and Social Psychology Review*, vol. 3, no. 1, pp. 23–48, 1999.

[31] B. F. Malle and J. Knobe, "The folk concept of intentionality," *Journal of Experimental Social Psychology*, vol. 33, no. 2, pp. 101–121, 1997.

[32] ——, "Which behaviors do people explain? a basic actor–observer asymmetry." *Journal of Personality and Social Psychology*, vol. 72, no. 2, p. 288, 1997.

[33] B. F. Malle, J. Knobe, M. J. O'Laughlin, G. E. Pearce, and S. E. Nelson, "Conceptual structure and social functions of behavior explanations: Beyond person–situation attributions." *Journal of Personality and Social Psychology*, vol. 79, no. 3, pp. 309–326, 2000.

[34] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[35] ——, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[36] E. T. Mueller, "Transparent computers: Designing understandable intelligent systems," *Erik T. Mueller, San Bernardino, CA*, 2016.

[37] G. Papagni and S. Koeszegi, "Understandable and trustworthy explainable robots: a sensemaking perspective," *Paladyn, Journal of Behavioral Robotics*, vol. 12, no. 1, pp. 13–30, 2021.

[38] 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.

[39] M. Riveiro and S. Thill, ""that's (not) the output i expected!" on the role of end user expectations in creating explanations of ai systems," *Artificial Intelligence*, vol. 298, 2021.

[40] ——, "The challenges of providing explanations of ai systems when they do not behave like users expect," in *UMAP 2022*. New York, NY, USA: ACM, 2022, p. 110–120.

[41] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *HRI 2016*. IEEE, 2016, pp. 101–108.

[42] A. Sciutti, A. Bisio, F. Nori, G. Metta, L. Fadiga, and G. Sandini, "Robots can be perceived as goal-oriented agents," *Interaction Studies*, vol. 14, no. 3, pp. 329–350, 2013.

[43] J. R. Searle, S. Willis *et al.*, *Intentionality: An essay in the philosophy of mind*. Cambridge University Press, 1983.

[44] A. Selbst and J. Powles, ""meaningful information" and the right to explanation," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 48–48.

[45] M. Shanahan, M. Crosby, B. Beyret, and L. Cheke, "Artificial intelligence and the common sense of animals," *Trends in Cognitive Sciences*, vol. 24, no. 11, pp. 862–872, 2020.

[46] S. Stange and S. Kopp, "Explaining before or after acting? how the timing of self-explanations affects user perception of robot behavior," in *ICSR 2021, Singapore, Singapore, November 10–13, 2021*. Springer, 2021, pp. 142–153.

[47] S. Thellman, "Social robots as intentional agents," Ph.D. dissertation, Linköping University Electronic Press, 2021.

[48] 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*, vol. 11, no. 4, pp. 1–51, 2022.

[49] S. Thellman, A. Silvervarg, and T. Ziemke, "Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots," *Frontiers in Psychology*, vol. 8, p. 1962, 2017.

[50] S. Thellman and T. Ziemke, "The perceptual belief problem: Why explainability is a tough challenge in social robotics," *ACM Transactions on Human-Robot Interaction*, vol. 10, no. 3, pp. 1–15, 2021.

[51] A. Theodorou, R. H. Wortham, and J. J. Bryson, "Why is my robot behaving like that? designing transparency for real time inspection of autonomous robots," in *AISB Workshop on Principles of Robotics*. University of Bath, 2016.

[52] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[53] S. Wahl and H. Spada, "Children's reasoning about intentions, beliefs and behaviour." *Cognitive Science Quarterly*, 2000.

[54] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," in *HRI 2016*. IEEE, 2016, pp. 109–116.

[55] L. Wells and T. Bednarz, "Explainable ai and reinforcement learning—a systematic review of current approaches and trends," *Frontiers in Artificial Intelligence*, vol. 4, p. 550030, 2021.

[56] R. H. Wortham and A. Theodorou, "Robot transparency, trust and utility," *Connection Science*, vol. 29, no. 3, pp. 242–248, 2017.

[57] A. Wykowska, T. Chaminade, and G. Cheng, "Embodied artificial agents for understanding human social cognition," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 371, no. 1693, p. 20150375, 2016.