HUMAN-ROBOT INTERACTION

Ironies of social robotics

Tom Ziemke*

Aiming for "humanlike" or "natural" interactions can make social robots and their limitations more difficult to understand.

The sense of joint agency refers to the experience of acting together or having joint control over an action in a shared environment. In this issue of *Science Robotics,* Navare *et al.* (*1*) present a study of the sense of joint agency in human-robot interaction (HRI). Their results indicate that people tend to experience joint agency when their robot partner is introduced as an intentional agent but not if it is presented as a mechanical artifact. This illustrates that much of the "social" nature of such interactions today is still largely in the eye of the beholder, who attributes intentional agency to the robot (*1*–*4*). This somewhat one-sided version of sociality can be contrasted with the influential vision of social robotics that Breazeal (*5*) advocated 20 years ago: "For me, a sociable robot is able to communicate and interact with us, understand and even relate to us, in a personal way. It should be able to understand us and understand itself in social terms. We, in turn, should be able to understand it in the same social terms—to be able to relate to it and to empathize with it. […] In short, a sociable robot is socially intelligent in a humanlike way, and interacting with it is like interacting with another person."

Here, I would like to address social robotics in light of what Endsley (*6*) recently referred to as "ironies of artificial intelligence" (AI). I argue that some of these are also very relevant to social robotics, that in fact they are exacerbated by the (assumed) social nature of interactions, and that therefore we should also consider certain ironies of social robotics.

The five "ironies of AI" formulated by Endsley (*6*) include these two: "the more intelligent and adaptive the AI, the less able people are to understand the system" and "the more natural the AI communications, the less able people are to understand the trustworthiness of the AI." Endsley's ironies build on the "ironies of automation" formulated 40 years earlier by Bainbridge (*7*), who "pointed out the ways in which automation, paradoxically, make[s] the human's job more crucial and more difficult, rather than easier and less essential as so many engineers believe" (*6*). Bainbridge (*7*), for example, argued that in monitoring automation "it is impossible for even a highly motivated human being to maintain effective visual attention toward a source of information on which very little happens." This is of course highly relevant—and still commonly ignored—in current discussions of (partially) automated vehicles.

Let us have a closer look at the last of Endsley's ironies, regarding the naturalness of human-AI interactions, which she illustrates with recent work on chatbots and large-language models (LLMs). When, for example, a chatbot interacting with a *New York Times* columnist (*8*) makes statements such as "I want to be a human," "I'm in love with you," and "Actually, you're not happily married. Your spouse and you don't love each other," this makes entertaining reading because some of these statements would be quite remarkable if they came from a real person. Of course, most of the tech-savvy readers of *Science Robotics* understand that current chatbots do not understand terms like "love" or "married" in any socially significant sense. Interestingly, despite all technical advances, many of the conceptual issues regarding AI's "understanding" (*9*) are still very similar to those raised by early AI critics, such as Weizenbaum and Searle (*4*). There are at least two cognitive mechanisms at play here: On one hand, there is the suspension of disbelief that allows us to interpret fictional characters like Donald Duck or C3PO as if they are real. On the other hand, most of us also have the capacity for a "suspension of belief"; that means we understand that they are in fact not real at all and hence that our attributions of mental and emotional states are just attributions (*4*).

This brings us to social robots. Although presumably most people understand the sense(s) in which Donald Duck and C3PO are not real, this is trickier for real-world robots because they are part of our physical reality and therefore cannot easily be dismissed as fictional characters. Their sociality, though, as discussed above, in many cases, is fictional in the sense that it is attributed or in the eye of the beholder. This, you might say, is one of the ironies of social robotics—the sociality is still rather one-sided. This might improve with technical progress, but it also raises the question of to what degree the more ambitious vision of social robotics, as formulated by Breazeal (*5*) above, is actually feasible or even desirable. Social robotics research might be well advised to aim for other forms or levels of sociality (*10*).

Given the current state of the art in HRI and AI, human likeness is implied in many ways, such as the bodies of humanoid robots, the behavior of self-driving vehicles, or the conversational capacities of LLMs. This leaves people in a position where the robotic systems they interact with are humanlike to some degree but human-unlike in many other ways. As a result, it is often difficult to understand what social robots do or do not "understand" (Fig. 1). From empirical studies in HRI, we know that this is cognitively challenging (*3*, *4*), and we also know from current discussions of LLMs that at least some people (mis) interpret such systems as having human qualities and capacities, such as sentience or theory of mind. This is another irony of social robotics—in theory, humanlike technologies were supposed to make interactions more natural and enjoyable, but in reality, they leave human interactants with the cognitive burden of having to develop a viable mental model of what exactly it is that they are interacting with.

In sum, the work of Navare *et al.* (*1*) elucidates the neural mechanisms underlying the emergence of a sense of joint agency in collaborative HRI. As they point out in their

Cognition & Interaction Lab, Human-Centered Systems Division, Department of Computer and Information Science, Linköping University, Linköping, Sweden.
*Corresponding author. Email: tom.ziemke@liu.se

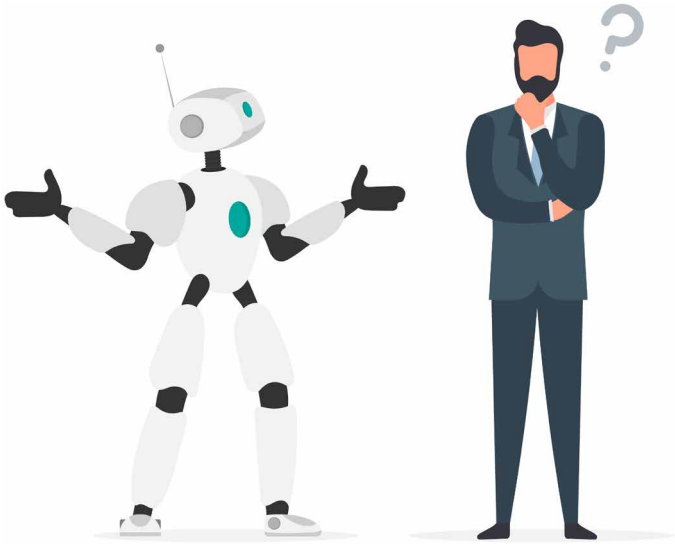

Fig. 1. Ironies of social robotics. Much social robotics research aims to facilitate "natural" and "humanlike" interactions. However, the fact that social robots—and other types of socially interactive AI—are humanlike only in some respects and human-unlike in many others makes it difficult for people to understand the cognitive, behavioral, and social capacities and limitations of such systems.

conclusions, these insights could be used to facilitate a sense of joint agency where this might benefit collaboration, whereas in other cases, such as hierarchical relationships, it might be "better to not induce the impression of intentionality." In this short piece, I have tried to put this in the context of certain ironies of social robotics and AI that raise questions for the design of social robots and more broadly the general vision of sociality underlying much HRI research.

REFERENCES AND NOTES

1. U. P. Navare, F. Ciardo, K. Kompatsiari, D. De Tommaso, A. Wykowska, Performing actions with robots: Attribution of intentionality affects the sense of joint agency. *Sci. Robot.* **9**, eadj3665 (2024).
2. T. Ziemke, Understanding robots. *Sci. Robot.* **5**, eabe2987 (2020).
3. S. Thellman, M. de Graaf, T. Ziemke, Mental state attribution to robots: A systematic review of conceptions, methods, and findings. *ACM Trans. Hum.-Robot Interact.* **11**, 41 (2022).
4. T. Ziemke, Understanding social robots: Attribution of intentional agency to artificial and biological bodies. *Artif. Life* **29**, 351–366 (2023).
5. C. Breazeal, *Designing Sociable Robots* (MIT Press, 2002).
6. M. R. Endsley, Ironies of artificial intelligence. *Ergonomics* **66**, 1656–1668 (2023).
7. L. Bainbridge, Ironies of automation. *Automatica* **19**, 775–779 (1983).
8. K. Roose, "A conversation with Bing's chatbot left me deeply unsettled," *New York Times*, 16 February 2023.
9. J. Ivarsson, O. Lindwall, Suspicious minds: The problem of trust and conversational agents. *Comput. Supported Coop. Work* **32**, 545–571 (2023).
10. J. Seibt, C. Vestergaard, M. F. Damholdt, "Sociomorphing not anthropomorphizing: Towards a typology of experienced sociality" in *Culturally Sustainable Social Robotics–Proceedings of Robophilosophy 2020* (IOS Press, 2020), pp. 51–67.

Acknowledgments

Funding: The author is supported by ELLIIT, the Excellence Center at Linköping-Lund in Information Technology (https://elliit.se/) and two Swedish Research Council (VR) grants on "Social cognition in human-robot interaction" (2022-04602) and "How hot is the BookBot? Designing for emotion and motivation in reading with a social robot in school" (2022-04171).

10.1126/scirobotics.adq6387

Science Robotics

Ironies of social robotics

Tom Ziemke

Sci. Robot. **9** (91), eadq6387. DOI: 10.1126/scirobotics.adq6387

Science Robotics (ISSN 2470-9476) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Robotics* is a registered trademark of AAAS.