Introduction to the Special Issue on Explainable Robotic Systems

MAARTJE M. A. DE GRAAF, Utrecht University
ANCA DRAGAN, University of California, Berkeley
BERTRAM F. MALLE, Brown University
TOM ZIEMKE, Linkoping University

Additional Key Words and Phrases: Robotic systems, human-robot interaction, explainability, transparency

ACM Reference format:
Maartje M. A. de Graaf, Anca Dragan, Bertram F. Malle, and Tom Ziemke. 2021. Introduction to the Special Issue on Explainable Robotic Systems. *ACM Trans. Hum.-Robot Interact.* 10, 3, Article 22 (July 2021), 4 pages.
https://doi.org/10.1145/3461597

1 TOPIC RELEVANCE

Robotic systems are likely to become increasingly ubiquitous but at the same time also increasingly complex. With this will come the need for them to be transparent and trustworthy for a broad range of users: people have to understand enough about a robot's inner workings to assess when such systems can be trusted. The call for **autonomous intelligent systems (AIS)** to be transparent has recently become loud and clear (e.g., [19]) and currently is a pressing funding and research agenda. Some forms of transparency, such as traceability and verification, are particularly important for software and hardware engineers [2, 5]; other forms, such as explainability or intelligibility, are particularly important for ordinary people [3]. As artificial agents, and especially socially interactive robots, enter human society, the demands for such systems to be transparent and explainable grow rapidly. When people interact with a robotic system, they construct mental models to understand and predict its actions. However, people's mental models of robots stem at least to some degree from their interactions with living beings. Thus, people easily run the risk of establishing incorrect or inadequate models of robotic systems, which may result in self-deception or even harm [23]. Moreover, a long-term study [18] showed that initially established (incorrect) mental models of an intelligent information system remained robust over time, even when details of the system's implementation were explained and initial beliefs were challenged with contradictory evidence. This can easily result in people either under-trusting or over-trusting robotic systems.

Incorrect mental models of AIS can have significant consequences for trust in such systems and, as a result, for acceptance of and collaboration with these systems [20]. Several studies indicate that people distrust a robotic system when they are unable to understand its actions. When a robot fails to communicate its intentions, people perceive the robot not only as creepy or unsettling [22] but also as erratic and untrustworthy even when it follows a clear decision-making process [9]. Indeed, when a robot is not transparent about its intentions (i.e., not providing any explanations

2573-9522/2021/07-ART22 $15.00
https://doi.org/10.1145/3461597

for its behavior), people may even question correct task performance and blame the robotic agent for its alleged errors [8]. In addition to such cases of distrust, incorrect mental models of AIS can also lead to the opposite situation. People sometimes over-trust artificial agents, such as when they comply with a faulty robot's unusual requests [16] or follow the lead of a potentially inept robot [14]. When a system is able to explain, however, how it made classifications or arrived at a judgment, users will be able to understand the system's behavior [6]. Furthermore, when a robot provides explanations of its own actions, people gather objectively more reliable information about the robot's abilities and mental states [7, 11]. Consequently, people are able to build more accurate mental models of robots [23] and establish appropriate levels of trust in robots [17, 20].

This special issue features seven articles on the topics of transparency, predictability, and explainability in the context of **human-robot interaction (HRI)**.

2 OVERVIEW OF INCLUDED ARTICLES

In "Building the Foundation of Robot Explanation Generation Using Behavior Trees," Han et al. aim to fill a crucial technical gap in the plan explanation literature by extending a promising and expressive planning technique into the toolbox of explainable methods. A series of algorithms is presented for enabling plan explanation within the context of behavior tree-based robot controllers.

Three articles present empirical research to address a variety of issues with the goal to increase explainability in robotic systems. In "Back-Off: Evaluation of Robot Motion Strategies to Facilitate Human-Robot Spatial Interaction," Reinhardt et al. evaluate strategies for mobile robots to convey their intention of yielding to pedestrians at bottlenecks. A video-based study explored the effectiveness of a robot's back-off strategy, compared to three other motion strategies, to signal its yielding intention. The authors then investigated the efficiency of this back-off strategy in real-world human-robot encounters. The results from the video-based study indicate that participants found the back-off strategy more legible, and the lab study with real encounters revealed that a short back-off strategy enables more efficient pedestrian motion.

Moon et al., in "Design of Hesitation Gestures for Nonverbal Human-Robot Negotiation of Conflicts," focus on robot movements as transparent indicators of hesitation in human-robot conflict negotiations. Based on a laboratory study identifying trajectory patterns observed in human negotiating hesitation, the authors developed an artificial "Negotiative Hesitation Generator" for a robotic system. An online video study validated these hesitation behaviors and concluded that people perceive robot movements produced by the Negotiative Hesitation Generator as more hesitant, animate, and anthropomorphic than they perceive smooth stopping behaviors.

In "I See What You Did There: Understanding People's Social Perception of a Robot and Its Predictability," Schadenberg et al. report on the development of a method for measuring people's perception of robot behaviors as predictable. An online video-based experiment tested several responsive actions of a robot. A lack of explanation or visibility of the robot's behavior increased participants' perception of the robot as incompetent, and the participants' intolerance of uncertainty was associated with their perception of the robot as uncomfortable.

Three further articles provide theoretical insights into how to make robotic systems more explainable. In the first article, "Explaining in Time: Meeting Interactive Standards of Explanation for Robotic Systems," Arnold et al. identify several criteria for designing explainable interactive robots. The article introduces a formal model of how to represent a robot's tasks and actions such that they can be explained, with a special focus on temporal aspects and interpretability.

In "Explainable Embodied Agents Through Social Cues: A Review," Wallkötter et al. provide a systematic literature review identifying four main motivations behind research on explainability, three categories of social cues that can be used to attain explainability, a number of algorithms used for implementation, and three main measures of the effects of explainability on interaction with robotic agents.

Finally, "The Perceptual Belief Problem: Why Explainability Is a Tough Challenge in Social Robotics," by Thellman and Ziemke, identifies the challenge to facilitate the attribution of appropriate intentional states, such as beliefs and desires, to robots as one fundamental mechanism in people's ability to explain and predict robot behavior. The authors analyze the challenges involved in forming mental models of a robot's perceptual beliefs, outline a general approach to studying such models empirically, and discuss potential solutions to the challenge of understanding a robot's intentional states.

3 FUTURE DIRECTIONS

After the shift in the role of robots from tools to partners in collaboration [1, 13, 21], the next step is a deeper integration of robots into society [12, 15]. HRI research can aid in this endeavor by developing, applying, and evaluating knowledge about HRI as it unfolds in complex everyday contexts, involving a diversity of users. In such an endeavor, many social, ethical, and technical challenges must be acknowledged and addressed, and we see three main challenges and thereby opportunities for future work.

First, the central goals of intelligibility and justified trust will continue to occupy HRI researchers and designers. However, we must also probe the limitations of designing explainable (as well as explaining) systems to achieve understanding and trust. Some aspects of artificial agents cannot and need not be made understandable to everyday users (e.g., deep technical implementations), and in some settings explanations can hurt understanding (because they confuse) or damage trust (because they are disruptive, shallow, or superfluous). We must investigate when and for what purposes people demand explanations [6, 10] from robotic systems and when and by which means explanations engender trust. Further, human-human trust is often achieved by other means than explanations, so insights into these additional bases of trust will help assign explainability the proper role in successful HRIs.

Second, ethical questions of how people should design, deploy, and treat robots arise from the rapid deployment of autonomous systems in sensitive human environments, from eldercare and education to security and law enforcement. However, ethical discussions of autonomy, deception, or intimacy sometimes reveal divides among communities and cultures. Mere theoretical discussions will not resolve these divisions, but ethical questions can be combined with empirical research from the social, behavioral, and cultural sciences. As a result, we may better understand the norms and values of specific communities and domains of application, enabling the design of robots that are socially and ethically acceptable to those humans they support and collaborate with.

Third, integrating empirical science with computational and engineering efforts will be paramount in developing technical solutions that are human centered—that is, responsive to human needs, expectations, and abilities. Extensive empirical testing throughout the design and deployment process while including potential end users would also provide important short-term and long-term evaluations of the impact of robotic technologies [4], on human-robot as well as human-human interactions.

In closing, we would like to thank the authors, reviewers, and production staff who have contributed to this special issue. The present articles illustrate that empirical, theoretical, and technical investigations are all needed to help elucidate the deeply intertwined relations between technical solutions, psychological research, and ethical questions, which together will inform our current and future interactions with robots. It is our hope that the articles featured in this special issue will inspire further ground-breaking and interdisciplinary research on explainable, trustworthy, and generally human-centered robot systems.

REFERENCES

[1] S. Bankins and P. Formosa. 2020. When AI meets PC: Exploring the implications of workplace social robots and a human-robot psychological contract. *European Journal of Work and Organizational Psychology* 29, 2 (2020), 215–229. https://doi.org/10.1080/1359432X.2019.1620328

[2] Jane Cleland-Huang, Orlena Gotel, and Andrea Zisman (Eds.). 2012. *Software and Systems Traceability*. Springer, London, UK. https://doi.org/10.1007/978-1-4471-2239-5

[3] Maartje de Graaf and Bertram F. Malle. 2017. How people explain action (and autonomous intelligent systems should too). In *Proceedings of the 2017 AAAI Fall Symposium Series Technical Reports*. 19–26.

[4] Maartje M. A. de Graaf, Somaya Ben Allouch, and Jan A. G. M. van Dijk. 2019. Why would I use this in my home? A model of domestic social robot acceptance. *Human–Computer Interaction* 34, 2 (2019), 115–173.

[5] M. Fisher, L. Dennis, and M. Webster. 2013. Verifying autonomous systems. *Communications of the ACM* 56, 9 (2013), 84–93. https://doi.org/10.1145/2494558

[6] S. R. Haynes, M. A. Cohen, and F. E. Ritter. 2009. Designs for explaining intelligent agents. *International Journal of Human-Computer Studies* 67, 1 (2009), 90–110.

[7] S. Kiesler. 2005. Fostering common ground in human-robot interaction. In *Proceedings of the International Workshop on Robot and Human Interactive Communication (RO-MAN'05)*. 729–734. https://doi.org/10.1109/ROMAN.2005.1513866

[8] T. Kim and P. Hinds. 2006. Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *Proceedings of the International Symposium on Robot and Human Interactive Communication (RO-MAN'06)*. IEEE, Los Alamitos, CA, 80–85. https://doi.org/10.1109/ROMAN.2006.314398

[9] M. Lomas, R. Chevalier, E. V. Cross, R .C. Garrett, J. Hoare, and M. Kopack. 2012. Explaining robot actions. In *Proceedings of the International Conference on Human-Robot Interaction (HRI'12)*. ACM, New York, NY, 187–188. https://doi.org/10.1145/2157689.2157748

[10] Bertram F. Malle and Joshua Knobe. 1997. Which behaviors do people explain? A basic actor-observer asymmetry. *Journal of Personality and Social Psychology* 72, 2 (1997), 288–304. https://doi.org/10.1037/0022-3514.72.2.288

[11] E. T. Mueller. 2016. *Transparent Computers: Designing Understandable Intelligent Systems*. CreateSpace Independent Publishing Platform.

[12] I. R. Nourbakhsh. 2013. *Robot Futures*. MIT Press, Cambridge, MA.

[13] E. Phillips, S. Ososky, J. Grove, and F. Jentsch. 2011. From tools to teammates: Toward the development of appropriate mental models for intelligent robots. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting (HFES'11)*, Vol. 55. 1491–1495.

[14] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *Proceedings of the International Conference on Human-Robot Interaction (HRI'16)*. 101–108. https://doi.org/10.1109/HRI.2016.7451740

[15] S. Šabanović. 2010. Robots in society, society in robots. *International Journal of Social Robotics* 2, 4 (2010), 439–450. https://doi.org/10.1007/s12369-010-0066-7

[16] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn. 2015. Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the International Conference on Human-Robot Interaction (HRI'15)*. 1–8. https://doi.org/10.1145/2696454.2696497

[17] A. Theodorou, R. H. Wortham, and J. J. Bryson. 2016. Why is my robot behaving like that? Designing transparency for real time inspection of autonomous robots. In *Proceedings of the AISB Workshop on Principles of Robotics*.

[18] J. Tullio, A. K. Dey, J. Chalecki, and J. Fogarty. 2007. How it works: A field study of non-technical users interacting with an intelligent system. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'07)*. ACM, New York, NY, 31–40. https://doi.org/10.1145/1240624.1240630

[19] S. Wachter, B. Mittelstadt, and L. Floridi. 2017. Transparent, explainable, and accountable AI for robotics. *Science Robotics* 2, 6 (2017), 3aan6080.

[20] N. Wang, D. V. Pynadath, and S. G. Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *Proceedings of the International Conference on Human Robot Interaction (HRI'16)*. IEEE, Los Alamitos, CA, 109–116.

[21] S. F. Warta, K. A. Kapalo, A. Best, and S. M. Fiore. 2016. Similarity, complementarity, and agency in HRI: Theoretical issues in shifting the perception of robots from tools to teammates. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting (HFES'16)*, Vol. 60. 1230–1234.

[22] T. Williams, P. Briggs, and M. Scheutz. 2015. Covert robot-robot communication: Human perceptions and implications for human-robot interaction. *Journal of Human-Robot Interaction* 4, 2 (2015), 24–49. https://doi.org/10.5898/JHRI.4.2.Williams

[23] R. H. Wortham and A. Theodorou. 2017. Robot transparency, trust and utility. *Connection Science* 29, 3 (2017), 242–248.