Explainable Robotic Systems

Maartje M.A. de Graaf
Brown University
maartje_de_graaf@brown.edu

Bertram F. Malle
Brown University
bertram_malle@brown.edu

Anca Dragan
UC Berkeley
anca@berkeley.edu

Tom Ziemke
Linköping University
tom.ziemke@liu.se

ABSTRACT

The increasing complexity of robotic systems are pressing the need for them to be transparent and trustworthy. When people interact with a robotic system, they will inevitably construct mental models to understand and predict its actions. However, people's mental models of robotic systems stem from their interactions with living beings, which induces the risk of establishing incorrect or inadequate mental models of robotic systems and may lead people to either under- and over-trust these systems. We need to understand the inferences that people make about robots from their behavior, and leverage this understanding to formulate and implement behaviors into robotic systems that support the formation of correct mental models of and fosters trust calibration. This way, people will be better able to predict the intentions of these systems, and thus more accurately estimate their capabilities, better understand their actions, and potentially correct their errors. The aim of this full-day workshop is to provide a forum for researchers and practitioners to share and learn about recent research on people's inferences of robot actions, as well as the implementation of transparent, predictable, and explainable behaviors into robotic systems.

KEYWORDS

Explainable robotics, behavior explanation, theory of mind, intentionality, transparency, trust calibration.

ACM Reference Format:
Maartje M.A. de Graaf, Bertram F. Malle, Anca Dragan, and Tom Ziemke. 2018. Explainable Robotic Systems. In *HRI '18 Companion: 2018 ACM/IEEE International Conference on Human-Robot Interaction Companion, March 5-8, 2018, Chicago, IL, USA.* ACM, New York, NY, USA, Article 4, 2 pages. https://doi.org/10.1145/3173386.3173568

1 TOPIC OVERVIEW

The call for Autonomous Intelligent Systems (AIS) to be transparent has recently become loud and clear 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 [1] [3]; other forms, such as explainability or intelligibility, are particularly important for ordinary people [2].

HRI '18 Companion, March 5-8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-5615-2/18/03…$15.00
https://doi.org/10.1145/3173386.3173568

As artificial agents, and especially socially interactive robots, enter human society, the demands for these agents to be transparent and explainable grow rapidly. When systems are able, for example, to explain how they made classifications or arrived at a decision, users are better able to judge the systems' accuracy and have more calibrated trust in them (e.g., [8] [10]).

More and more AI systems process vast amounts of information and make classifications or recommendations that humans use for financial, employment, medical, military, and political decisions. More precariously yet, autonomous social robots, by definition, make decisions reaching beyond direct commands and perform actions with social and moral significance for humans. The demand for these social robots to become transparent and explainable is particularly urgent. However, to make robots explainable, we need to understand how people interpret the behavior of such systems and what expectations they have of them.

In this workshop, we will address the topics of transparency and explainability, for robots in particular, from both the cognitive science perspective and the computer science and robotics perspective. Cognitive science elucidates how people interpret robot behavior; computer science and robotics elucidate how the computational mechanisms underlying robot behavior can be structured and communicated so as to be human-interpretable. The implementation and use of explainable robotic systems may prevent the potentially frightening confusion over why a robot is behaving the way it is. Moreover, explainable robot systems may allow people to better calibrate their expectations of the robot's capabilities and be less prone to treating robots as almost-humans.

2 BACKGROUND

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 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 [12]. Moreover, a long-term study [9] 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 given and initial beliefs were challenged with contradictory evidence.

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 [10]. 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 not only perceive the robot as creepy or unsettling [11], they also perceive such robots as erratic and untrustworthy even when they follow a clear decision-making process [5]. Indeed, when a robot is not transparent about its intentions (i.e., not providing any explanations for its behavior), people may even question its correct task performance and blame the robotic agent for its alleged errors [4]. 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 [7] or follow the lead of a potentially inept robot [6].

Thus, there should be little doubt about the value of making robotic systems perform easily interpretable actions and explain their own decisions and behaviors. This goal for explainable robots demands innovative work in computer science, robotics, and cognitive science, and our proposed workshop provides strong representation from all these fields. Additionally, we invite research that examines if and how such structures lead to better acceptance of robotic systems, to more accurate mental models of such systems, and to calibrated trust in these systems.

3 WORKSHOP ACTIVITIES

The aim of this full-day workshop is to provide a forum to share and learn about recent research on requirements for artificial agents' explanations as well as the implementation of transparent, predictable and explainable robotic systems. Extended time for discussions will highlight and document promising approaches and encourage further work. We welcome prospective participants to submit extended abstracts (max. 2 pages) that contribute to the discussion of people's interpretation of robot actions as well as the implementation of transparent, predictable, and explainable behaviors in robotic systems. After the conference, and with permission of the authors, we will provide online access to the workshop proceedings on a dedicated workshop website: explainableroboticsystems.wordpress.com. We encourage researchers to attend the workshop even without a paper submission, as a major portion of the workshop involves community engagement and discussion to foster uptake of concepts regarding explainable robotic systems within the field of HRI.

The morning session will cover invited talks from Rachid Alami, Joanna Bryson, Bradley Hayes, and Alessandra Sciutti, as well as short presentations by authors of accepted papers, followed by general discussion. The afternoon will be devoted to break-out sessions to discuss the next steps in research and development of explainable and transparent robotic systems. Groups will be composed of representatives of different disciplines in order to work on integrating the multiple necessary perspectives in this endeavor. To boost the discussions, we will ask presenters to prepare questions or raise pressing issues that provide starting points for the discussion groups.

4 ORGANIZERS

Maartje de Graaf is a postdoctoral research associate at the Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, USA. She has a Bachelor of Business Administration in Communication Management (2005), a Master of Science in Media Communication (2011), and a PhD in Communication Science and Human-Robot Interaction (2015). Her research research interest focuses on peopleâĂŹs social, emotional and cognitive responses to robots including the societal and ethical consequences of those responses.

Bertram Malle is Professor of Psychology in the Department of Cognitive, Linguistic, and Psychological Sciences at Brown University and Co-Director of the Humanity-Centered Robotics Initiative at Brown. He was trained in psychology, philosophy, and linguistics at the University of Graz, Austria, and received his Ph.D. in Psychology from Stanford University in 1995. His research focuses on social cognition, moral psychology, and human-robot interaction.

Anca Dragan is an Assistant Professor in Electrical Engineering and Computer Sciences at UC Berkeley. She received her PhD from the Robotics Institute at Carnegie Mellon University in 2015 and now runs a lab on algorithmic human-robot interaction. She has expertise in transparent and explainable robotics and AI, with applications in personal robots and autonomous cars.

Tom Ziemke is Professor of Cognitive Science at the University of Skövde and Professor of Cognitive Systems at Linköping University, Sweden. He received his Ph.D. in Computer Science from Sheffield University, UK, in 2000. His main research interests are embodied cognition and social interaction, and in recent years in particular people's interaction with different types of autonomous systems, ranging from social robots to automated vehicles.

REFERENCES

[1] Jane Cleland-Huang, Orlena Gotel, Andrea Zisman, et al. 2012. *Software and systems traceability*. Vol. 2. Springer.

[2] Maartje MA de Graaf and Bertram F Malle. 2017. How People Explain Action (and Autonomous Intelligent Systems Should Too). (2017).

[3] Michael Fisher, Louise Dennis, and Matt Webster. 2013. Verifying autonomous systems. *Commun. ACM* 56, 9 (2013), 84–93.

[4] Taemie Kim and Pamela Hinds. 2006. Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on*. IEEE, 80–85.

[5] Meghann Lomas, Robert Chevalier, Ernest Vincent Cross II, Robert Christopher Garrett, John Hoare, and Michael Kopack. 2012. Explaining robot actions. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, 187–188.

[6] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on*. IEEE, 101–108.

[7] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin 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 Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 141–148.

[8] Andreas Theodorou, Robert H Wortham, and Joanna J Bryson. 2016. 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.

[9] Joe Tullio, Anind K Dey, Jason Chalecki, and James Fogarty. 2007. How it works: a field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 31–40.

[10] Ning Wang, David V Pynadath, and Susan G Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 109–116.

[11] Tom Williams, Priscilla Briggs, and Matthias Scheutz. 2015. Covert Robot-Robot Communication: Human Perceptions and Implications for HRI. *Journal of Human-Robot Interaction* 4, 2 (2015), 23–49.

[12] Robert H Wortham and Andreas Theodorou. 2017. Robot transparency, trust and utility. *Connection Science* 29, 3 (2017), 242–248.