To err is robotic: Understanding, Preventing, and Resolving robots' failures in HRI

Alessandra Rossi[1], Kheng Lee Koay[2] and Kerstin S. Haring[3]

***Abstract*— Robots placed in human-oriented dynamic environments, such as private homes, shopping malls, healthcare facilities, are likely to exhibit occasional behaviours which are perceived by people as unexpected, failures, or actual errors. Robots' errors can negatively affect people's perception of the robotic behaviours, in terms of usefulness, functionalities and capabilities, trustworthiness and acceptability. This session will focus on examine how people from perceive robot's failures in short- and long-term interactions, and highlight how different failures influence their perceptions and emotions toward the robots. This will help with subject related to classifying of robot errors from the aspect of human (e.g. intentional, voluntarily, perceived vs. real errors – recklessness, forgetfulness, poor motivation –) and robot dimensions (e.g. actual error - algorithms, sensors, actuators). This session will explore how different techniques can be used to enhancing natural human-robot communication (such as inner speech, legibility, predictability and transparency of robotic behaviours, explicit and non-explicit strategies) can be used to help people understand the implications, risks, and goals of robots' behaviours and to mitigate the perception of the failures. In this session, we also want to explore strategies both to prevent robots from exhibiting unintended behaviours, and to mitigate the effects of robot errors on human-robot interaction.**

I. Title

To err is robotic: Understanding, preventing and resolving robots' failures in HRI

II. Aim & Scope

Robots might be placed in unpredictable human-oriented dynamic environments. This includes environments outside of controlled or structured laboratories, in private homes, shopping malls, or healthcare facilities and it is likely that at times they exhibit behaviours which might be perceived as disruptive, erratic, faulty, or useless. In addition, robotics in dynamic environments do not exhibit behaviors in a vacuum: their behaviours may be influenced by unexpected people's behaviours. Unexpected human behaviors are unforeseen environmental variables that affect the robots' sensor readings, algorithms, and mechanical limitations. A robot's decision-making abilities may also be limited, so while trying to carry on the correct course of actions, robots might still mistakenly make the wrong decision. Moreover, people tend to attribute any unexpected and incoherent robot behaviours, perceived robot failures or actual failures as robot errors [1]–[3]. For example, a robot that navigates too slowly might be considered having faulty behaviours. People's perception of the robotic behaviours, in terms of usefulness, functionalities and capabilities, trustworthiness and acceptability, is influenced by the nature of the error. People's perception of the negative effects of robots' errors may be principally differentiated by consequences of the error [4], the timing in which this may happen [5], the repetition of such errors over time, the probability of a negative outcome, and so on. A first step in identifying how robotic behaviours are perceived as failures were to divide them in two categories [3]: technical and interaction failures. Technical failures are considered errors produced by hardware or software problems, which can depend on an erroneous design, communication or processing. In contrast, interaction failures are related to social norm violations, organisational and mental-model based faults (e.g., expectation) in the interaction within a particular context between people and robots. However, robotics errors and how these are perceived by people do not only depend on robot self, but they are also a consequence of other factors. For example, they may be a consequence of human errors, or an unclear and non-transparent communication, or they may depend on a misunderstanding and miscommunication of the social, psychological and cognitive conventions expected by people. Moreover, there are also some cases in which these behaviours may be perceived as if robots intentionally deceive or cheat people. As a consequence, these may result in people wrongly interpreting and predicting the robots' intents and behaviours, and negatively affecting Human-Robot Interaction (HRI) [6].

In this session, we aim to further understand how robot's failures can be categorised, how the failures are perceived by people, in short- and long-term interactions, and how different failures influence people's perceptions and emotions toward the robots, by exploring aspects from both the robot and human dimensions. In this direction, we will explore how different techniques used for enhancing natural human-robot communication, such as inner speech, legible, predictable and transparent robotic behaviours, explicit and non-explicit strategies, can be used to allow people to understand the implications, risks, and goals of robots' behaviours to mitigate the perception of the failures.

While dealing with robotic failures, it is fundamental to consider two different strategies. The first one is oriented to prevent robots from exhibiting unintended behaviours, which in some extreme cases may even endanger people's, pets' safety, or break objects. For example, roboticists [7] have been borrowed strategies from agent-oriented theory, such

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

[2]Kheng Lee Koay is with the University of Hertfordshire, UK k.l.koay@herts.ac.uk

[2]Kerstin Sophie Haring is with the University of Denver, USA b.d.researcher@ieee.org

Proposal 9 submitted to 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). Received January 25, 2023.

as Software Verification Methods, to develop robots that are able to adapt and recover from any erroneous behaviour by analysing their own internal state, the state of the other agents and the environmental context, or by ensuring that newly created or learned behaviours do not collide with existing ones [8]. Other approaches may include in endowing the robots with the appropriate robot etiquette [9] and Theory of Mind [10], to ensure that robots may meet people's social expectations. A second strategy is, instead, focused on mitigating the effects of robot errors on human-robot interaction by endowing the robot with human-like expected behaviours [4], [11], such as apologies, promises or additional reasons that can explain or justify the erroneous behaviours. Another inner ability of people is to be able to understand and predict others' behaviours, therefore, robots also need to be able to show legible and predictable behaviours to enhance people's feeling of safety, comfort, efficiency and ability of the robot itself.

Nonetheless, it is not clear how effective these strategies might be with different types of errors (real or perceived), or how these vary in long-term interactions. In this session, we want to investigate which are the most appropriate strategies for preventing, and recovering people's positive perception in a robot (i.e., trust, acceptance, reliance, and so on) in case the robot may still exhibit erroneous behaviours.

The topics covered in this special session are in line with the main theme of the conference (i.e., "Design New Bridge for H-R-I"). In particular, we want to start by fostering the **[R] Robotic Recovery and Reconnection** to allow **[I] Intelligent Interface and Interaction** for the **[H] Human Health, Happiness and Hope**. 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
- Modelling Trust and Acceptance in HRI

III. 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:

- Antonio Chella, University of Palermo (Italy)
- Eduard Fosch-Villaronga, Leiden University (Netherlands)
- Georgios Angelopoulos, University of Naples Federico II (Italy)
- Mohamed Chetouani, CNRS UMR7222, Sorbonne University (France)
- Pourya Aliasghari, University of Waterloo (Canada)
- Caroline L. van Straten, University of Amsterdam (Netherlands)
- Alessandra Rossi, University of Naples Federico II (Italy)
- HeeSun Choi, Texas Tech University (USA)
- P.A. Hancock, University of Central Florida (USA)
- Rachid Alami, LAAS-CNRS (France)
- Kheng Lee Koay, University of Hertfordshire (UK)
- Helen Hastie, Heriot-Watt University (UK)
- Shelly Levy-Tzedek, Ben-Gurion University of the Negev (Israel)
- Kerstin Dautenhahn, University of Waterloo (Canada)
- Nicole Robinson, Monash University (Australia)
- Severin Lemaignan, PAL Robotics (Spain)

ACKNOWLEDGMENT

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

IV. Biographies

Dr Alessandra Rossi 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 (https://secure-robots.eu/) 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 is Publicity chair at IEEE RO-MAN 2023. She has been Publicity chair at IEEE RO-MAN 2022, Virtual Organizing Chair of IEEE RO-MAN 2021, Registration Chair and Social Media Responsible of IEEE RO-MAN 2020. She is the team leader of RoboCup team "Bold Hearts" at the University of Hertfordshire (UK), and TC member of the RoboCup Humanoid League since 2021. 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 2018-2021, workshop TRAITS at HRI 2021 and 2022, special sessions at RO-MAN 2019, 2020, 2021 and 2022, special issues at the Interaction Studies (IS), International Journal of Social Robotics (IJSR), and Paladyn Journal of Behavioral Robotics.

Dr Kheng Lee Koay received his B.Sc. degree in Robotics and Automated Systems and Ph.D. from the University of Plymouth, Plymouth, U.K. in 1997 and 2003 respectively. He was a Senior Research Fellow at the Adaptive Systems Research Group, University of Hertfordshire, Hertfordshire, U.K. from 2003 and joined the Department of Computer Science as a Senior Lecturer from 2016.

He was involved in European projects COGNIRON (Cognitive Robot Companion), LIREC (Living with Robots and Interactive Companions), ACCOMPANY (Acceptable Robotics Companions for Ageing Years), SECURE (Safety Enables Cooperation in Uncertain Robotic Environments), as well as the UK Engineering and Physical Sciences Research Council (EPSRC) funded project Trustworthy Robotic Assistants and Innovate UK funded Knowledge Transfer Partnership (KTP) project CuPick (Cucumber Picking System).

His research interests include Mobile Robotics, Robotic Home Companions and Human-Robot Interaction, in particular, aspects of human-centred socially acceptable interactions, personalisation, trust, experimental design and evaluation methodologies. He has experience in working with numerous robotic platforms and systems, including coordinating, developing and integrating the University of Hertfordshire Robot House with different robotic platforms. He has designed and built two robots for the study of robotic home companions which provide cognitive and physical assistance using social behaviours, inspired by hearing dogs, to visually communicate intention.

Dr Kerstin Sophie Haring is an Assistant Professor at the Ritchie School of Engineering and Computer Science at University of Denver, USA. She directs the Humane Robot Technology Laboratory (HuRoT) that envisions interdisciplinary research in robotics with the goal of improving human lives through the promotion of better technology. Her research interests include socially intelligent robots, robot ethics, robot design, robot theory of mind, and explainable, effective, and efficient interactions with robots. Before her appointment at the University of Denver, she researched Human-Machine-Teaming at the U.S. Air Force Academy, completed her PhD in 2015 in Human Robot Interaction at the University of Tokyo in Japan, and has a graduate degree in Computer Science from the University of Freiburg in Germany. Dr. Haring serves as Associate Editor of the IEEE Robotics and Automation Letters and as Editor of the Springer "Women in Robotics" book.

REFERENCES

[1] E. Short, J. Hart, M. Vu, and B. Scassellati, "No fair!! an interaction with a cheating robot," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, March 2010, pp. 219–226.

[2] S. Lemaignan, J. Fink, F. Mondada, and P. Dillenbourg, "You're doing it wrong! studying unexpected behaviors in child-robot interaction," in *Social Robotics*, A. Tapus, E. André, J.-C. Martin, F. Ferland, and M. Ammi, Eds. Cham: Springer International Publishing, 2015, pp. 390–400.

[3] S. Honig and T. Oron-Gilad, "Understanding and resolving failures in human-robot interaction: Literature review and model development," *Frontiers in Psychology*, vol. 9, 2018.

[4] A. Rossi, K. Dautenhahn, K. L. Koay, and M. L. Walters, "The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario." *Paladyn Journal of Behavioral Robotics*, vol. 9, 2018.

[5] A. Rossi, K. Dautenhahn, K. L. Koay, M. L. Walters, and P. Holthaus, "Evaluating people's perceptions of trust in a robot in a repeated interactions study," in *Social Robotics: 12th International Conference, ICSR 2020, Golden, CO, USA, November 14–18, 2020, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 453–465.

[6] K. Haring, K. Nye, R. Darby, E. Phillips, E. de Visser, and C. Tossell, "I'm not playing anymore! a study comparing perceptions of robot and human cheating behavior," in *Social Robotics*, M. A. Salichs, S. S. Ge, E. I. Barakova, J.-J. Cabibihan, A. R. Wagner, Á. Castro-González, and H. He, Eds. Cham: Springer International Publishing, 2019, pp. 410–419.

[7] S. Costantini, "Ensuring trustworthy and ethical behaviour in intelligent logical agents," *Journal of Logic and Computation*, vol. 32, 01 2022.

[8] K. L. Koay, M. Webster, C. Dixon, P. Gainer, D. Syrdal, M. Fisher, and K. Dautenhahn, "Use and usability of software verification methods to detect behaviour interference when teaching an assistive home companion robot: A proof-of-concept study," *Paladyn, Journal of Behavioral Robotics*, vol. 12, no. 1, pp. 402–422, 2021. [Online]. Available: https://doi.org/10.1515/pjbr-2021-0028

[9] K. Koay, M. Walters, A. May, A. Dumitriu, B. Christianson, N. Burke, and K. Dautenhahn, "Exploring robot etiquette: Refining a hri home companion scenario based on feedback from two artists who lived with robots in the uh robot house," in *Social Robotics*, ser. Lecture Notes in Computer Science. Springer, Dec. 2013, pp. 290–300, 5th Int Conf on Social Robotics, ICSR 2013 ; Conference date: 27-10-2013 Through 29-10-2013.

[10] A. Rossi, A. Andriella, S. Rossi, C. Torras, and G. Alenyà, "Evaluating the effect of theory of mind on people's trust in a faulty robot," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2022, pp. 477–482.

[11] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *Proceeding HRI '16 The Eleventh ACM/IEEE International Conference on Human Robot Interation*. IEEE Press Piscataway, 2016, pp. 101–108.

Proposal 9 submitted to 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). Received January 25, 2023.