Socially Assistive Robot Decision Making: Transparency, Motivations, and Intentions

***NEED TO DETERMINE ORDERING HERE

LYNNE BAILLIE, Heriot-Watt University, UK

EMILYANN NAULT, Heriot-Watt University & University of Edinburgh, UK

CARL BETTOSI, Heriot-Watt University & University of Edinburgh, UK

RONNIE SMITH, Heriot-Watt University & University of Edinburgh, UK

SCOTT MACLEOD, Heriot-Watt University, UK

MAJA MATARIĆ, University of Southern California, USA

MANFRED TSCHELIGI, University of Salzburg, AUT

FABIO PATERNÒ, Research Council of Italy - Institute of Information Science and Technologies, ITA

VIVEK NALLUR, University College Dublin, IRL

Socially assistive robots (SARs) have already become a pervasive presence in our daily lives, fulfilling a range of roles that were previously filled only by humans. As the complexity and capability of such agents grow, they will be expected to take on higher degrees of responsibility and execute greater levels of autonomous decision-making. Therefore, it is imperative that the Human-Robot Interaction (HRI) and greater Human-Computer Interaction (HCI) community seriously consider how those agents communicate about their role and the motivations and intentions behind these decisions. The proposed workshop will address challenges with respect to SAR decision making, current approaches to these challenges, and develop ideas and strategies for how the community should move forward in this area.

CCS Concepts: • Computer systems organization → Robotic autonomy; • Human-centered computing → *Accessibility*. Additional Key

Words and Phrases: Socially assistive robots, Transparency, Robotic autonomy, Machine ethics

ACM Reference Format:

***Need to determine ordering here, Lynne Baillie, Emilyann Nault, Carl Bettosi, Ronnie Smith, Scott MacLeod, Maja Matarić, Manfred

Tscheligi, Fabio Paternò, and Vivek Nallur. 2018. Socially Assistive Robot Decision Making: Transparency, Motivations, and Intentions.

In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY.* ACM, New York, NY, USA, 7 pages.

https://doi.org/XXXXXXX.XXXXXXX

1 MOTIVATION

Socially Assistive Robots (SARs) are becoming increasingly more relevant in the human-robot interaction space with promising applications in healthcare, education, assistive living, physical coaching, and beyond. They can allow for more seamless integration into the existing social dynamic of the environment through intelligent means, as opposed to forcing the user to adapt to the limitations of some traditional technology intervention [3]. Furthermore, SARs' social

Manuscript submitted to ACM

capabilities have been seen to drive greater engagement over the likes of virtual agents in assistive scenarios where they have been deployed as exercise coaches [2] [5]. Beyond the benefits for the end-user, real-world SAR interventions have the potential to reduce workload for assistive staff, yet, to reach this level of usefulness will require a higher level of autonomy thus handing over more serious decision-making responsibilities to the robot agent. This shift towards SARs as key decision-makers creates new challenges for the HRI community that we aim to shed light on during this workshop.

One differentiating factor of social robots is their exposure to users who are not knowledgeable in their motivations or intentions, something that is not necessarily obvious at face-value. In addition to this, in human-human interactions, assistive staff are considered experts in their area of practice and are therefore trusted in their decisions, or otherwise when prompted can reason as to why decisions were made. As we hand over greater decision-making responsibilities to SARs, a key consideration will be how to build transparent and trustworthy interactions in which the robot can communicate not only its motivation and intentions, but justification and reasoning around its individual choices. Such transparency is becoming an ever-growing challenge as personalization and adaption are now recognized as key factors in achieving greater engagement with these systems [6]. Because of this, research is progressing from simple rule-based decision systems to complex decision-making mechanisms such as machine learning approaches. These techniques have provided promising results, however, are notoriously difficult for even humans to explain to one another.

To complicate this further, research has shown that human trust can be significantly impacted by errors during HRI [4]. Moreover, SARs are typically seen to operate within sensitive populations and challenging tasks. Although frameworks have been introduced in recent years to consider ethical and cultural differences when designing these systems [1], other characteristics of the interaction such as physical or cognitive impairments create new considerations for how these agents should communicate their decisions.

2 CHALLENGES TO ADDRESS

There are various challenges that arise as SARs increasingly take on decision-making responsibilities that historically would have been in the hands of humans. Challenges are exacerbated particularly in interactive scenarios where the end-user is vulnerable and/or has some form of physical/cognitive impairment. The objective of this workshop is to bring together researchers and practitioners within HCI and HRI to discuss the following challenges:

- Determining which decisions should be made by the robot can be difficult. Where do we draw the line, and who decides this?
- What to do when underlying reasoning is based on complex processes that are not always straightforwardly interpretable or explainable (i.e. black-box processes)?
- How to strike a balance between transparency and explanations which are understandable to all users?
- How do SARs smoothly integrate into existing relationships in assistive scenarios to build trust with its users?
- How do SARs recover from unintended/incorrect actions through communication with the user and ensure this does not negatively affect trust? That is, assuming the SAR learns via some mechanism that its previous action(s) were incorrect.
- How should a SAR's decision-making approach be adapted for specific populations, (e.g., informed consent for individuals with cognitive impairments)?

- How do we limit the negative real-world consequences of decisions made by SARs?

3 ORGANIZERS

Lynne Baillie, Heriot-Watt University and University of Edinburgh (UK) Point of Contact. Prof Lynne Baillie has been involved in user centred design of home, mobile and rehabilitation technologies for over fifteen years. She has had several full papers published at the SIGCHI conference and has also previously run a workshop at CHI. She is currently the Director of the Interactive and Trustworthy Technologies Research Group at Heriot-Watt University. Her research work has been funded by RCUK and FFG, international companies (Orange, Telecom Austria, Alcatel-Lucent, Siemens, Microsoft), charities (Heritage Lottery Fund, CHSS, Paths to Health, Calman), and Governments (local, national and EU).

Emilyann Nault, Heriot-Watt University and University of Edinburgh (UK) Point of Contact. Emilyann Nault is a fourth-year PhD student researching how socially assistive robots and sensory feedback can be used to foster engagement with cognitive activities for older adults. She has integrated Participatory Design and user-centered design methodologies to engage end-users and relevant stakeholders throughout the research process.

Carl Bettosi, Heriot-Watt University and University of Edinburgh (UK). Carl Bettosi is a second-year PhD student whose research focuses on adaptive socially assistive robots for upper-limb rehabilitation. Specifically, Carl is interested in the use of machine learning techniques to help drive better engagement through more adaptive and personalised behavioural policies in the agent.

Ronnie Smith, Heriot-Watt University and University of Edinburgh (UK). Ronnie Smith is a PhD student researching how to bring humans 'in-the-loop' of their own assistive technology solutions, ultimately enabling pro-active robotic assistance during daily life. His research bridges themes in artificial intelligence and human-robot interaction, with recent work on activity recognition, active learning, conversational agents, and human-robot collaboration during activities of daily living.

Scott MacLeod, Heriot-Watt University (UK). Scott MacLeod is a PhD student whose research focus is using pervasive sensing, telepresence and robotic technology, to facilitate remote assessment, and automate continuous cognitive assessment of people with mild cognitive impairment and/or at risk of developing dementia. To enable the provision of assistance always in tune with people's stages in life and care needs.

Maja Matarić, University of Southern California (USA). Maja Matarić is the Chan Soon-Shiong Distinguished Professor of Computer Science, Neuroscience, and Pediatrics at the University of Southern California (USC), and founding director of the USC Robotics and Autonomous Systems Center. Her PhD and MS are from MIT, BS from Kansas University. She is Fellow of AAAS, IEEE, AAAI, and ACM, recipient of the Presidential Award for Excellence in Science, Mathematics & Engineering Mentoring, Anita Borg Institute Women of Vision for Innovation, NSF Career, MIT TR35 Innovation, and IEEE RAS Early Career Awards, and authored "The Robotics Primer" (MIT Press). A pioneer of the field of socially assistive robotics, her research is developing human-machine interaction methods for personalized support in convalescence, rehabilitation, training, and education for autism spectrum disorders, stroke, dementia, anxiety, and other major health and wellness challenges.

Manfred Tscheligi, University of Salzburg (AUT). Waiting on bio

Fabio Paternò, National Research Council of Italy - Institute of Information Science and Technologies (ITA). Fabio Paternò is a Research Director at the CNR-ISTI. His main research interests are in Interactive Smart Spaces, Human-Robot Interaction, Accessibility, End-User Development, and Human-centered Artificial Intelligence.

Vivek Nallur, University College Dublin (IRL). Works in the area of Machine Ethics. He is interested in how to implement and verify ethics in autonomous machines. Questions such as what kinds of ethics would autonomous machines agree to among themselves, how would we ensure that individually ethical machines don't combine to produce un-ethical behaviour, are interesting to pose and answer computationally. On the Organizing Committee for AAAI 2021 Spring Symposium Series on Implementing AI Ethics [22-24 March 2021] and a Program Committee member for the 1st Computational Machine Ethics Workshop at KR2 2021. He is a member of the IEEE P7008 Standards committee for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems. Multi-Agent Systems (MAS) are his preferred tool for approaching such problems and simulating possible futures with different ethical frameworks.

Sara Copper, PAL Robotics (ESP).

- Email: sara.cooper@pal-robotics.com
- Google Scholar

4 WEBSITE

https://sites.google.com/view/sar-decision-making. (May change host prior to submission).

5 PRE-WORKSHOP PLANS

Interested parties will be invited to submit a short 4-page paper which engages with the overarching theme of decision making around SARs. Prior to the workshop, a dedicated Discord server will be used to encourage discussion and begin building a sense of community. Discord has been chosen based on prior positive experiences with the platform, however, we will ask participants for their opinions and will adapt the platform if necessary to better serve the group. Participants will be asked to share sources to relevant research, demonstrations, news articles, etc. This will assist the organizers in determining how to best structure the discussion at the workshop in order to highlight the communities interests and priorities. We will further reach out via Discord asking to let the organizers know privately (via Discord or email) if they have any accessibility requirements, which we will then organize through communication with the conference accessibility chairs.

6 HYBRID APPROACH

In order to include globally diverse participants, we plan to organize the workshop so participants may attend in person or online. To support this hybrid setup, we will question participants prior to the workshop so we can adapt our approach to include and support all in attendance. The synchronous discussion will be held in person and remotely via the Discord server. The asynchronous discussion will be facilitated through Discord. Certain organizers will be assigned the role of Support Chair to facilitate and assist remote participants.

7 ASYNCHRONOUS ENGAGEMENT

Resources will be provided to before, during, and after the workshop in order to facilitate discussion and community building.

(1) Before the workshop, the accepted papers will be accessible through the website. Also, there will be a dedicated Discord channel for participants to introduce themselves.
(2) The workshop discussions will be recorded and an online whiteboard tool (i.e.„ Miro) will be used to virtually collaborate across synchronous and asynchronous participation.
(3) After the workshop, relevant resources will be posted to Discord and the workshop website to encourage further asynchronous communication.

8 WORKSHOP STRUCTURE

The workshop will bring together experts in HRI, HCI, and ethics from around the world who have encountered the challenges presented in this workshop around SAR decision-making. They will contribute their experiences, methodologies, and practices to build a holistic view of these challenges from a variety of disciplines. With this foundation, we will discuss conflicts within the ethical boundaries of these challenges, current approaches, and how they can be improved and better integrated going forward.

This will be a one-day workshop split into three sessions (Table 1). We expect 20-30 participants to be in attendance in person and will accept 15-25 short (4 pages) position or research papers for presentation. These presentations will take place on the day of the workshop and will vary in format (e.g., short talk, poster, demos, videos) in order to maximize engagement and collaboration between participants. For the workshop activity, participants will be split into groups, each of which will receive a scenario in which a SAR needs to make a decision. They will be asked to come up with how the robot in their given scenario will make decisions and how they will be communicated to the user with respect to the challenges of the workshop. They can do so through written descriptions or sketches. Online participants will engage with the activity through the Miro boards and Discord channel, and an organizer will be available to help facilitate. The goal of this activity is to develop concrete strategies towards SAR decision-making regarding what information and how it is communicated to the user. It will also provide a practical means to discuss these overarching challenges. The final session of the workshop will consist of an open discussion of themes that have been identified throughout the day and how we can move forward in the space of SAR decision-making. The proposed schedule is as follows:

Table 1. Workshop Schedule

9:00 - 9:30	Introduction & Welcome
9:30 - 12:00	Session 1 (Presentations in various formats)
12:00 - 13:00	Lunch
13:00 - 15:00	Session 2: Workshop Activity
15:00 - 17:00	Session 3: Discussion

Expected Outcomes.

- Build a community across the fields of HRI, HCI and ethics who are currently working in or are interested in the area of SAR decision-making.
- Establish challenges around how SARs make decisions, how and in what way these decisions are communicated to their users, and the surrounding ethical implications.
- Derive concrete strategies and practices to address these challenges.
- Disseminate the wider results and outcomes from this workshop through a publication and follow-up workshop.

9 ACCESSIBILITY

To ensure the inclusion of all participants, we will work with those with accessibility needs and ascertain the support they require through collaboration with the accessibility chairs. At least one workshop organizer will be given the role of Accessibility Chair to ensure all needs are met. Further, we will require all paper submissions to follow the SIGCHI Guide for an Accessible Submission, and all video submissions must contain subtitles/captioning.

10 POST-WORKSHOP PLANS

Accepted workshop papers will be submitted to post on arxiv, with links accessible through the workshop website. The organizers will create a publication summarizing the challenges discussed surrounding SAR decision-making and the outcomes of the workshop. A poster version will also be created in order to share this work with the broader community. A subsequent workshop will also be proposed to discover how the outcomes have been integrated into HRI research. Finally, a mailing list will be created to facilitate future collaboration.

11 CALL FOR PARTICIPATION

Socially Assistive Robot Decision Making: Transparency, Motivations, and Intentions is a one-day hybrid workshop which aims to discuss challenges, current practices, and ethical implications of Socially Assistive Robot (SAR) decision making. We welcome 4-page position or research contribution papers from researchers in the areas of HCI and HRI who would like to enrich our collective understanding of the potential practical and societal impact of social agents that make decisions about humans. The papers should be in the CHI extended abstract format and be submitted through Easychair. Papers should ideally address one or more of the challenges listed below. However, we welcome works that present material related to other potential challenges within the theme of SARs that make decisions.

- Which decisions should be made by the SAR and who decides this?
- How should we provide reasoning for complex underlying processes?
- How do we strike a balance between transparency and explanations which are understandable to all users?
- How do SARs smoothly integrate into existing relationships in assistive scenarios to build trust with its users?
- How do SARs recover from unintended/incorrect actions through communication with the user?
- How should SARs approaches be adapted for specific populations?
- How do we limit the negative real-world consequences of decisions made by SARs?

Key Dates.

- Submission Deadline: 11:59pm February 1st, 2023
- Notification of Acceptance: 11:59pm February 28th, 2023
- Camera Ready: 11:59pm March 8th, 2023

REFERENCES

[1] Linda Battistuzzi, Antonio Sgorbissa, Chris Papadopoulos, Irena Papadopoulos, and Christina Koulouglioti. 2018. Embedding ethics in the design of culturally competent socially assistive robots. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1996–2001.

[2] Juan Fasola and Maja J Matarić. 2013. A socially assistive robot exercise coach for the elderly. *Journal of Human-Robot Interaction* 2, 2 (2013), 3–32.

[3] Allison Langer, Ronit Feingold-Polak, Oliver Mueller, Philipp Kellmeyer, and Shelly Levy-Tzedek. 2019. Trust in socially assistive robots: Considerations for use in rehabilitation. *Neuroscience & Biobehavioral Reviews* 104 (2019), 231–239.

[4] Birthe Nesset, David A Robb, José Lopes, and Helen Hastie. 2021. Transparency in hri: Trust and decision making in the face of robot errors. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 313–317.

[5] Valentina Vasco, Cesco Willemse, Pauline Chevalier, Davide De Tommaso, Valerio Gower, Furio Gramatica, Vadim Tikhanoff, Ugo Pattacini, Giorgio Metta, and Agnieszka Wykowska. 2019. Train with me: a study comparing a socially assistive robot and a virtual agent for a rehabilitation task. In *International Conference on Social Robotics*. Springer, 453–463.

[6] Katie Winkle, Praminda Caleb-Solly, Ailie Turton, and Paul Bremner. 2018. Social robots for engagement in rehabilitative therapies: Design implications from a study with therapists. In *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 289–297.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009