MULTITTRUST: 2nd Workshop on Multidisciplinary Perspectives on Human-AI Team Trust

Nicolo' Brandizzi[1], Carolina Centeio Jorge[2], Roberto Cipollone[1], Francesco Frattolillo[1], Luca Iocchi[1] and Anna-Sophie Ulfert-Blank[3]

[1]*Sapienza University of Rome, Via Ariosto, 25, 00185 Roma RM, Italy*
[2]*Delft University of Technology, Van Mourik Broekmanweg, 6, 2628 XE Delft, Netherlands*
[3]*Eindhoven University of Technology, De Zaale, Atlas 7.404, 5600 MB Eindhoven, Netherlands*

Abstract

In the evolving field of Artificial Intelligence (AI), research is transitioning from focusing on individual autonomous agents to exploring the dynamics of agent teams. This shift entails moving from agents with uniform capabilities (homogeneous) to those exhibiting diverse skills and functions (heterogeneous). At this phase, research on mixed human-AI teams is the natural extension of this evolution, promising to extend the application of AI beyond its traditional, highly controlled environments. However, this advancement introduces new challenges to the learning system, such as trustworthiness and explainability. These qualities are critical in ensuring effective collaboration and decision-making in mixed teams, where mutual cooperation and decentralized control are fundamental. Reinforcement Learning emerges as a flexible learning framework that well adapts to semi-structured environments and interactions, such as those under consideration in this work.

This paper aims to contribute to bridging the gap between Multi-Agent Reinforcement Learning (MARL) and other disciplines that focus on human presence in teams or examine human-AI interactions in depth. We explore how MARL frameworks can be adapted to human-AI teams, highlight some of the necessary modeling choices, discuss key modeling decisions, and highlight the primary challenges and constraints. Our goal is to establish a unified framework for mixed-learning teams, encouraging cross-disciplinary contributions to refine MARL for complex settings.

Keywords

Human-AI Team Trust, Multidisciplinary Perspectives, Computational Trust Estimation, Human-Robot Interaction

1. Overview

Our workshop[1] emerges from the need to connect the multidisciplinary research community that concentrates on examining the different aspects of trust in human-AI teams. With the rapid growth of these teams across varied industries, there is an increasing call for careful consideration of the challenges that come with it. Trust, a vital construct within mixed human-robot teams, has been studied extensively across disciplines, particularly in human-computer

HAI '23: International Conference on Human-Agent Interaction, December 4–7, 2023, Gothenburg, Sweden
brandizzi@diag.uniroma1.it (N. Brandizzi); c.jorge@tudelft.nl (C. C. Jorge); cipollone@diag.uniroma1.it (R. Cipollone); frattolillo@diag.uniroma1.it (F. Frattolillo); iocchi@diag.uniroma1.it (L. Iocchi); a.s.ulfert.blank@tue.nl (A. Ulfert-Blank)
0000-0002-3191-6623 (N. Brandizzi); 0000-0002-6937-5359 (C. C. Jorge); 0000-0002-0421-5792 (R. Cipollone); 0000-0002-2040-3355 (F. Frattolillo); 0000-0001-9057-8946 (L. Iocchi); 0000-0001-6293-4173 (A. Ulfert-Blank)

[1]The workshop's website is available at https://multittrust.github.io/2ed/.

interaction and psychology. However, considering the complex dynamics and diverse team compositions, a comprehensive understanding of trust in human-AI teams remains elusive.

This workshop is a second edition following its initial launch at the HHAI 2023 Conference in Munich (multittrust.github.io). Building on the successes and learnings of the first edition, where nearly 30 participants explored the aspects of human-AI trust, we observed key themes emerging:

- **Intention Communication**: Effective human-AI cooperation requires that AI systems clearly communicate their intentions, boosting trust and collaboration.
- **Trust Calibration**: There's a recurring challenge of overtrust, particularly in crisis situations, which necessitates methodologies to achieve a balanced trust calibration between humans and AI entities.
- **Team Dynamics**: Several papers indicated the need to consider AI as part of a team, emphasizing 'teamness' and the role of AI-enabled decision support systems.
- **Ethical Considerations**: With AI playing such a crucial role in decision-making, ethical considerations around their deployment, especially in high-risk situations, came to the forefront.

This second edition stems from the high enthusiasm registered during the first one. Notably, the exploration of practical methodologies to assess trust in mixed human-AI teams emerged as a focal point of discussions, sparking the interest of numerous potential contributors for future submissions. This edition focuses on integrating knowledge across fields with the goal of enhancing computational methods to estimate trust in human-AI teams. We aim to facilitate meaningful conversations and collaborations among researchers from various disciplines, including psychology, sociology, cognitive science, computer science, artificial intelligence, robotics, human-computer interaction, and human-robot interaction.

2. Goals and Objectives

The primary goal of this workshop is to explore and identify computational approaches that can accurately estimate trust in human-AI teams. We aim to:

- Build a comprehensive understanding of trust in human-AI teams, leveraging knowledge from various disciplines.
- Promote the development and application of computational methods for trust estimation.
- Encourage collaboration among researchers from various fields.

The previous edition of the workshop appeared at the intersection of Interactive Intelligence and Organizational Psychology. In addition, this edition includes new members in the organizing team with complementary backgrounds, including researchers specialized in Multi-Agent Reinforcement Learning. As such, in this edition of the workshop, we aim to extend the community and keep connecting new fields.

3. Workshop/Tutorial Structure

The schedule begins with an introduction, followed by keynote sessions, lightning talks, breaks, and wraps up with group and round table discussions before the closing. In the following sections, we explain the structure of each of these activities.

Keynote Speakers

This workshop will spotlight two keynote talks that explore the core topics of the workshop.

1. **Alan Wagner, Pennsylvania State University**: Wagner will discuss human responses to robot guidance in simulated emergencies, highlighting tendencies for overtrust and the influence of anthropomorphism. He will also touch upon ethical considerations for evacuation robots.
2. **Karinne Ramirez-Amaro, Chalmers University of Technology**: Ramirez-Amaro's research focuses on the intricacies of human-agent collaboration and communication. Her talk will expand on these themes, offering an in-depth exploration of collaborative strategies and trust metrics.

Together, these keynotes offer a comprehensive perspective on trust in human-AI relationships, bridging foundational concepts with practical implications.

3.1. Lightning talks

We believe it is important to have brief presentations where the participants who previously submitted a short paper can present their work. The main purpose of these talks is not to give details about the contributions. Instead, we want to introduce some of the researchers to the community, connect different research groups and start discussions (that will be continued in the afternoon). After each presentation, of approximately 10 min., we will proceed with few questions/answers and ask participants to write down the rest of the questions, which will be very welcome in the afternoon.

3.2. Small-groups and round-table discussions

After the keynote and lightning talks, our participants will have a long list of questions and ideas. They will be suggested to join a group (previously decided by the organizers based on submitted contributions) to discuss these questions and ideas further. Naturally, they can choose another group they prefer to join. Besides the fruitful discussions, we hope that the participants find time here to network and get to form some connections within the community.

Finally, we would like to close the workshop by discussing with everyone the reflections of the day. The organizers will summarize the outcome of the small group discussions and will incentivize further discussions. The organizers will moderate this discussion.

4. Expected Outcomes

This workshop is designed to address key challenges and opportunities related to trust in human-AI interactions. The anticipated outcomes include:

1. **Promotion of Multi-Disciplinary Dialogues**: Engage researchers across various disciplines to promote a comprehensive understanding of trust dynamics in human-AI systems, which will improve current frameworks and solutions.
2. **Computational Trust Evaluation**: Identify and improve computational models for trust assessment. This involves both critiquing existing models and proposing refined or new approaches to better capture the meaning and different levels of trust.
3. **Trust in Human-Robot Interactions**: Examine the practical applications and challenges of trust metrics in mixed human-robot teams. This includes current scenarios and anticipated future developments.

By merging insights from different fields and emphasizing computational methods, the workshop aims to advance research and practical implementations in the domain of trust for human-AI collaborations.

5. Expected Audience and Call for Papers Plan

This workshop calls for contribution and/or participation from several disciplines, including psychology, sociology, cognitive science, computer science, artificial intelligence, robotics, human-computer interaction, and human-robot interaction. Topics related to this workshop include:

- Dynamics of trust between humans and AI in teamwork.
- Computational measures of team trust and evaluation methods of trustworthiness in human-AI teams.
- Human's trust and trustworthiness in human-AI teams.
- Experimental settings for trust dynamics in human-AI teams.
- Design of systems that take into account trust dynamics in human-AI teams.

The call for paper and the review process will be carried out with the main goal of gathering diverse researchers working on topics of interest that can contribute to the discussion towards the identified goals.

5.1. Program committee

Hebert Azevedo-Sa, Military Institute of Engineering, BR; *Piercosma Bisconti*, DEXAI – Artificial Ethics, IT; *Angelo Cangelosi*, University of Manchester, UK; *Filippo Cantucci*, ISTC-CNR, IT; *Cristiano Castelfranchi*, ISTC-CNR, IT; *Antonio Chella*, University of Palermo, IT; *Filipa Correia*, ITI Larsys, PT; *Rino Falcone*, ISTC-CNR, IT; *Eleni Georganta*, University of Amsterdam, NL; *Glenda Hannibal*, Ulm University, DE; *Jundi Liu*, Iowa State University, US; *Federico Manzi*, Universita' Cattolica del Sacro Cuore di Milano, IT; *Antonella Marchetti*, Universita' Cattolica del

Sacro Cuore di Milano, IT; *Siddharth Mehrotra*, Delft University of Technology, NL; *Alessandro Sapienza*, ISTC-CNR, IT; *Beau Schelble*, Clemson University, US; *Samuele Vinanzi*, Sheffield Hallam University, UK; *Michelle Zhao*, Carnegie Mellon University, US.

ACKNOWLEDGMENTS

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA8655-23-1-7257, and supported by TU Delft AI Initiative, under the AI*MAN lab, ERC-ADG White-Mech (No. 834228), EU ICT-48 2020 project TAILOR (No. 952215), PNRR MUR FAIR (No. PE0000013).