Responsible Human-Machine Teaming Workshop

Joseph Early, Mohammad Divband Soorati, Gopal Ramchurn

Summary

The Responsible Human-Machine Teaming Workshop was a one day event that explored the areas of responsible artificial intelligence, explainable artificial intelligence, human-centred design, robotics, and human robot interaction. The workshop was co-organised by academics from UCL, University of Southampton, and University of Nottingham as well as the DSTL's AI Lab. It covered a mix of theoretical and practical research questions, and aimed to promote inter-disciplinary and multi-disciplinary approaches to responsible human-machine teaming. The event was attended by researchers from relevant fields, users and practitioners from diverse industries where humans and autonomous systems work in close partnership, and high level decision makers interested in the implications of human-machine teaming for their organisation. The main outcome of the workshop was the development of four key research areas:

- Human-Machine Teaming as an Engineering Discipline
- The Evolution of Human-Machine Teams
- Trust in Human-Machine Teams
- Explainability in Human-Machine Teams

Human-Machine Teaming as an Engineering Discipline

Due to the scope and multi-disciplinary nature of human-machine teaming, there exists a large, diverse body of literature from many disciplines (engineering, social sciences, and arts and humanities) that is relevant to human-machine teaming. A review of the current state of the art in each area is required from a human-machine teaming perspective. This will help establish the key questions in each research area, and, if coupled with applied case studies, will provide a concrete foundation from which further human-machine teaming research can occur as its own discipline.

Human-machine teaming is relevant to a wide range of audiences: the general public, policy makers, industry and applied domains, and research and academia. The relevant communication and information requirements differ for each group. By establishing human-machine teaming as its own discipline, resources can be provided that are tailored to the need of each group. There is then potential to push for more rigor in human-machine teaming, for example designing a curriculum for courses and teaching people human-machine teaming specific skills.

Evolution of Teams

Human-machine teams are not static entities – they develop over time as context, operators and goals change. There is a need for a high-level framework that articulates human machine teams and their purpose, components and development stages. This leads to two research questions:

- How can human-machine teams be evaluated through-out their lifecycle?
- Why and at what points do human-machine teams fail?

The former question needs to define metrics for team 'health' – an evaluation of the human elements, the machine elements and the team as a whole. Once the efficacy of a human-machine team can be evaluated, work on the latter question can begin. Experimental approaches to these questions could investigate the effect of changing the human component (e.g. inexperienced vs experienced operator), changing the machine implementation (e.g. using different algorithms), and changing the environmental context (e.g. how does the team adapt as the context changes over time). By analysing the evolution of human-machine teams over time, steps can be taken to develop more robust teams, and safety measures can be created that highlight if teams are likely to fail during operation.

Trust in Human-Machine Teams

Trust forms an important element of human-machine teams. Whilst trust is most often considered from the human perspective (i.e. if the humans trust the outcomes of the machine elements, especially if it is giving them commands), there is the contrasting aspect where the level of trust that a machine has in its human teammates can affect the outcome of its decisions. Trust encompasses many dimensions, for example temporal, subjective (self-reporting) and group dynamics. It also covers many disciplines such as neuroscience, engineering, human factors, human control interfaces and design.

To develop human-machine teams with trust as core feature, a rigorous framework for trust within human-machine teams must be developed. Trust is developed at different levels and over different timeframes, with some being more important than others. Often testing trust in these systems requires failure of the system, but how can this be done in a safe way? The balance between using simulations and real-world testing is an open question, and selecting the appropriate kind of experiments is dependent on the context in which the human-machine team operates.

Explainability in Human-Machine Teaming

Explanations are required in human-machine teams for several reasons; they can provide human team members with the reasoning behind algorithmic decisions, but also for post-hoc analysis of why certain decisions were made at certain times. Explanations must be appropriate in their level of detail and the time at which they are delivered. Explainability is closely related to trust – if humans can understand how a machine came to a decision, they will have more faith that the decision is correct, or at least be able to justify overriding the decision.

Two open research areas for explainability in human-machine teams were outlined at the workshop. The first is about providing explanations of dynamic environments. Human-machine teams are often deployed in changing environments, and human members of the team must be kept up to date with the current state of the environment. By providing sufficient explanations of the dynamic situation, human members can be kept up to date with what's going on. The second area is adaptive explainability. There is no "one size fits all" explanation that can be provided for every situation. Open questions revolve around providing the right information at the right time, and in the best form. Explainability is also tightly coupled with responsible innovation as it provides better transparency of systems.

Organizing committee & Participants

The workshop was held at The Alan Turing Institute in London. It comprised of a combination of talks from academia and industry, and café style discussion scenarios to establish key research directions and outline the future areas of work in human-machine teams. The workshop was led by the University of Southampton.

More than 40 participants with various backgrounds and expertise attended the workshop from organizations in UK and also worldwide including but not limited to: Spere, Dstl, King's College London, Thales, EPSRC UKRI, Air Force Research Laboratory, University of Oxford, Technical University of Crete, Liverpool, John Moores University, University of Nottingham, Science and Technology, lead; Mission Systems Europe. Archangel Imaging Ltd, Cranfield University, University of Birmingham.

Organisers included: Prof. Steve Meers (DSTL AI Lab), Dr. Joel Fischer (Nottingham), Dr. Enrico Costanza (UC), Prof. Sarvapali D. Ramchurn, Fiona Butcher (DSTL AI Lab)

Contact

Professor Sarvapali D. (Gopal) Ramchurn
University of Southampton
sdr1@soton.ac.uk
https://responsibleai.info/