CHI 2021 Workshop: Towards Trustworthy and Explainable Autonomous Physical Systems

Workshop Date: May 7, 13:00 UK time.

The safe deployment of autonomous physical systems in real-world scenarios requires them to be explainable and trustworthy, especially in critical domains. In contrast with 'black-box' systems, explainable and trustworthy autonomous physical systems will lend themselves to easy assessments by system designers and regulators. This promises to pave ways for easy improvements that can lead to enhanced performance, and as well, increased public trust. In the one-day virtual workshop, we gathered a globally distributed group of researchers and practitioners to discuss the opportunities and social challenges in the design, implementation, and deployment of explainable and trustworthy autonomous physical systems, especially in a post-pandemic era. Interactions were fostered by panel discussions and a series of spotlight talks. We conducted a pre-workshop survey (before the workshop) which examined the public perception of the trustworthiness of autonomous physical systems. This document serves as a summary report that provides details about the pre-workshop survey and as well as the identified challenges from the workshop.

Workshop Speakers

Check link for information about the speakers:
https://etapsworkshop.github.io/#speakers

Workshop Schedule (BST Timing)

Time	***Event***	***Speaker***	***Time***	***Event***	***Speaker***
13:00	Welcome	Organiser 1	15:50	Keynote 6	Erik Vinkhuyzen
13:05	Keynote 1	Tim Miller	16:15	Panel discussion 1	Invited speakers
13:30	Keynote 2	Paul Luff	16:40	Coffee break 2	
13:55	Keynote 3	Alun Preece	16:55	Keynote 7	Ehud Sharlin
14:20	Spotlights	Paper authors	17:20	Keynote 8	Kati Driggs-Campbell
14:45	Coffee break 1		17:45	Panel discussion 2	Invited speakers
15:00	Keynote 4	Masoumeh Mansouri	18:10	Wrap-up	Organiser 2
15:25	Keynote 5	Bastian Pfleging	18:15	Virtual drinks (optional)	Everyone!

Workshop Notes

Below are the details of the speakers' talks including their slides (where available).

Speaker	***Title***	***Abstract***
Tim Miller	Explainable artificial intelligence: beware the inmates running the asylum. Or How I learnt to stop worrying and love the social and behavioural sciences)	In his seminal book *The Inmates are Running the Asylum: Why High-Tech Products Drive Us Crazy And How To Restore The Sanity*, Alan Cooper argues that a major reason why software is often poorly designed (from a user perspective) is that programmers are in charge. As a result, programmers design software that works for themselves, rather than for their target audience; a phenomenon he refers to as the 'inmates running the asylum'. In this talk, I argue that explainable AI risks a similar fate if AI researchers and practitioners do not take a multi-disciplinary approach to explainable AI. I further assert that to do this, we must understand, adopt, implement, and improve models from the vast and valuable bodies of research in philosophy, psychology, and cognitive science; and focus evaluation on people instead of just technology. I paint a picture of what I think the future of explainable AI will look like if we went down this path.
Paul Luff	Planning and Situating Actions: challenges for explanation and trustworthiness in autonomous systems	Lucy Suchman's seminal work 'Plans and Situated Actions' (1987) contrasted different ways of viewing actions and interactions with technology, particularly with 'intelligent' systems. With related work at the time it led to a radical shift In how we consider the nature of interaction with and the capabilities of technologies. In this talk, we will briefly revisit this work and discuss its relevance to contemporary studies of explainable systems and trustworthy technologies. We will draw on examples of explanations and everyday behaviour to consider the nature of explanations and trust and discuss research currently underway in two research projects. The first, THuMP considers how techniques derived from planning, provenance, argumentation and visualisation might help to develop explainable systems. The second concerning trustworthy autonomous systems brings together researchers from different disciplines to develop novel methods and approaches to investigate trustworthiness. We will suggest that a concern with 'situated action' and a particular characterisation of 'context' remain pertinent when considering such contemporary

		developments in trustworthy autonomous systems
Alun Preece	"If it walks like a duck and quacks like a duck...": Coherent Multimodal Explanations for Trustable Machine Teammates	Our work focuses on the problem of coalition situational understanding where humans and machine agents need to form ad hoc teams to deal with rapidly-changing situations, particularly in 'front line' settings. Quickly building trust between humans and machine agents is a key issue in ad hoc team formation, because the team members may have limited prior experience of working together. We are particularly interested in exploiting machine agents based on deep neural networks, where explainability and interpretability are active areas of research; we are also interested in hybrid neuro-symbolic machine agents. Moreover, coalition situational understanding commonly involves decision-making based on data collected from sensors of multiple modalities, for example video and audio. In this talk I will describe how we employ multimodal explainability techniques together with 'tellability' (the ability for humans to impart knowledge to machine teammates) to support rapid calibration of trust in ad hoc coalition teams.
Masoumeh Mansouri	Trustworthy and Explainable Autonomous Robotic Systems: Requirements and Solutions	
Bastian Pfleging	HCI challenges of automated vehicles – how can we trust and understand what they do?	
Erik Vinkhuyzen	Normal Traffic Assumptions	Foundational to movement on public roads is mutual trust between road users. This trust is that other people on the road are on the whole benevolent, that they see the world in the same way, that they are self-aware and understand that their own actions will communicate to others, just as they will interpret other road users actions as essentially communicative of their own intentions, and that they act deliberately, and can always fall back on direct interaction in case mutually recognized miscommunication occurs. In short, people proceed under what I have called

		“Normal Traffic Assumptions”. The implications for Autonomous Vehicles will be discussed."
Ehud Sharlin	Autonomous Vehicles and Pedestrians: From Obstacle Avoidance to Interaction	
Kati Driggs-Campbell	Building Trust in Autonomous Systems through Communication and Validation	Abstract: Autonomous robots are becoming tangible technologies that will soon impact the human experience. However, the desirable impacts of autonomy are only achievable if the underlying algorithms can handle the unique challenges humans present: People tend to defy expected behaviors and do not conform to many of the standard assumptions made in robotics. In this talk, we'll discuss tools for designing safe, interactive autonomy that is effective in (near) failure modes by analyzing critical states. Specifically, we'll focus on methods for communicating and engaging with the user to integrate human insights in the engineering design process and to correct or intervene at runtime. Using tools from decision-making under uncertainty and explainable AI, we'll demonstrate how to achieve smooth interaction and stable adaptation that results in improved safety and trust in autonomous vehicles.

Survey Report

We conducted a survey around some topics in autonomous physical systems. We used the feedback from the survey to drive conversations in the workshop panel discussions. View survey responses

Paper Presentations

Check this link to view presented papers: https://etapsworkshop.github.io/#presentations