

Towards Explainable and Trustworthy Autonomous Physical Systems

Daniel Omeiza
University of Oxford
Oxford, UK
daniel.omeiza@cs.ox.ac.uk

Sule Anjomshoae
Umeå University
Sweden
sule.anjomshoae@umu.se

Konrad Kollnig
University of Oxford, UK
konrad.kollnig@cs.ox.ac.uk

Oana-Maria Camburu
University of Oxford and the Alan
Turing Institute, UK
ocamburu@turing.ac.uk

Kary Främling
Umeå University
Sweden
kary.framling@umu.se

Lars Kunze
Oxford Robotics Institute, University
of Oxford
UK

ABSTRACT

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 this one-day virtual workshop, we aim to gather 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 will be fostered through panel discussions and a series of spotlight talks. To ensure lasting impact of the workshop, we will conduct a pre-workshop survey which will examine the public perception of the trustworthiness of autonomous physical systems. Further, we will publish a summary report providing details about the survey as well as the identified challenges resulting from the workshop’s panel discussions.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative interaction**.

KEYWORDS

Explainability, trust, collaboration, human-machine interaction

ACM Reference Format:

Daniel Omeiza, Sule Anjomshoae, Konrad Kollnig, Oana-Maria Camburu, Kary Främling, and Lars Kunze. 2021. Towards Explainable and Trustworthy Autonomous Physical Systems. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI ’21 Extended Abstracts)*, May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3411763.3441338>

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI ’21 Extended Abstracts, May 8–13, 2021, Yokohama, Japan

© 2021 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-8095-9/21/05.

<https://doi.org/10.1145/3411763.3441338>

1 BACKGROUND

Autonomous systems with a high degree of autonomy are gaining increasing influence on business, science, and society in general. They are particularly playing an important role by performing specific responsibilities in critical domains [1] e.g., health-care; especially in the current COVID-19 pandemic era. Users’ attitude towards such systems is crucial for the public acceptance and adoption of these systems. This is particularly important in human-robot interaction [5, 7, 15]. If the behaviour of the given autonomous system is not readily predictable, the user tries to make sense of the action which may not necessarily reflect the actual rationale of the system [11]. This does not only degrade the quality of the interaction but may also leads to serious consequences (e.g., pedestrian misreading the action of a self-driving car) which could affect the systems’ trustworthiness. In this workshop we would be concentrating on autonomous physical systems, that is, embodied autonomous systems (e.g. mobile robots and autonomous vehicles). It is necessary to ensure that this class of systems are capable of explaining and justifying their behaviour and actions in order to increase users’ trust, confidence, trust and public acceptance [3, 10, 14].

It is important to investigate ways to enable these autonomous physical systems to communicate with humans and other systems, provide reasons behind their actions and also communicate unstable conditions or adversarial attacks [4, 8, 9]. This improves performance in the context of human-systems collaboration, since a common understanding is key for successful collaboration [2, 6, 13]. Furthermore, autonomous physical systems with such explanation capabilities will facilitate error tracing in order to correct faulty behaviours. Thus, contributing to the evolution of autonomous physical systems [12]. In order to achieve explainable and trustworthy autonomous physical systems, it is necessary to understand how people interpret their behaviour and what people expect of them.

This workshop will offer the possibility to explore ways of developing solutions to issues relating to explainability, collaboration, and acceptance, with a focus on the interaction between human and autonomous physical systems. Therefore, the objectives of this workshop are to:

- Establish common ground for the study and development of human-centered explainable and trustworthy autonomous physical systems.

- Explore the components of trustworthy autonomous physical systems.
- Assess the impact of explanations on user confidence, understandability, and trust.
- Discuss applications and contributions towards overcoming the lack of explainability and trust.
- Discuss alternative means for the next generation of explainable and trustworthy autonomous physical systems.

2 ORGANISERS

We believe our organisational team is well-suited to conduct this workshop, given both the diversity of our disciplines as well as a common interest grounded in autonomous system applications. Our team comes from the AI accountability and trust, data privacy, cognitive robotics, explainable autonomous driving, and explainable AI backgrounds. We have experience organising successful workshops in the past as discussed below.

Daniel Omeiza (daniel.omeiza@cs.ox.ac.uk) is a PhD student at the University of Oxford working on explainability and trust in autonomous vehicles. He is also a research candidate in the mobile robotics group of the Oxford Robotics Institute. He obtained a masters degree from Carnegie Mellon University and has worked briefly with IBM Research. He has experience co-organising the NeurIPS workshop on Machine Learning for Autonomous Driving. Daniel is also part of the programme committee for the AI in Africa for Sustainable Economic Development workshop. He has also served as a volunteer in the Black in AI (BAI) workshops co-located with the NeurIPS conference in 2018 and 2019.

Lars Kunze is a Departmental Lecturer in Robotics in the Oxford Robotics Institute (ORI) and the Department of Engineering Science at the University of Oxford. He is also the lead of the Cognitive Robotics Group in the ORI. Lars has a PhD in Cognitive Science, and has organised several workshops on autonomous robotic systems at top-tier conferences including ICRA, ITSC and IV.

Sule Anjomshoae is a PhD student in the Explainable AI (XAI) research team at the Umeå University. Her research focuses on generating and presenting human-understandable explanations for the predictions made by black-box algorithms. During her master studies, she worked on pattern recognition and image processing areas. After graduation, she worked as a research assistant in VicubeLab at the University of Technology Malaysia focusing on natural user interfaces using 3D modeling and simulation.

Kary Främling is the head of the Explainable AI (XAI) team at the University of Umeå in Sweden. He has been an active researcher in Artificial Intelligence focusing on topics such as neural network learning, multiple criteria decision support and reinforcement learning. Kary is also the founder and head of the Adaptive Systems of Intelligent Agents team at Aalto University, Finland. He has organised several workshops on IoT, intelligent products and most recently on XAI (EXTRAAMAS).

Konrad Kollnig is a PhD student at the University of Oxford. With a background in computer science and mathematics, his research tries to analyse how to increase user participation in the design of our day-to-day technological architecture. He did his MSc at the University of Oxford on analysing and promoting GDPR compliance, under the supervision of Professor Max Van Kleek. As

part of this, he developed TrackerControl, a widely used privacy app that emphasises user participation in its design.

Oana-Maria Camburu is a post-doctoral researcher at the University of Oxford and co-investigator at the Alan Turing Institute on the project of Natural Language Explanations for Deep Neural Networks. Oana's research focuses mainly on explainability and natural language processing. Oana completed her PhD at the University of Oxford, working on explainability and natural language processing.

3 LINK TO WEBSITE

Information about the workshop which includes call for abstract, and talks are provided on the website below.

- <https://etapsworkshop.github.io/>

4 PRE-WORKSHOP PLANS

We currently have 11 confirmed speakers for our workshop. These speakers have previously spoken in related workshops such as XAI, XAI@IJCAI, and Explainable Robotic Systems. They were selected based on our search for previous participants of the CHI conference, and from our extended networks from academia and the industry.

We will circulate a call for participation via relevant mailing lists, forums, and social media. To make the workshop's presentations and panel discussion sessions relevant and productive, we will send out a survey along with the call for participation. This survey will help us identify critical issues around the workshop theme. The feedback from the survey will help our multidisciplinary group of speakers appropriately position their presentations.

A website will be set up to provide information about the workshop. This includes call for extended abstract with topics of interest, submission guidelines, and links to related material.

5 WORKSHOP STRUCTURE

We propose a one-day virtual workshop consisting of eleven keynotes, two panel discussions, and a session of spotlight talks. The keynotes will be given by invited speakers from a range of different research areas, including social science, computer science, robotics, and human-computer interaction. During the workshop, we will have two panel discussions, which will keep the workshop lively and engaging. The panel discussions will be guided by the responses received from the pre-workshop survey and some pertinent points from the workshop talks. Generally, this will be around current trends and trajectories for explainable and trustworthy autonomous physical systems. Interactions between participants and the audience will be further fostered through a session of spotlight talks of submitted papers and dedicated chat rooms. To make this workshop widely accessible across all continents and to maximise participation, we have chosen a time frame which allows people from different parts of the world to join. See Table 1 for the tentative schedule (including times in GMT). We are considering using Microsoft Teams or Zoom for the virtual conference, and Slack for paper discussions. Keynote speakers will be provided with the options of either providing a video of their talk or making a live presentation. We will ask the authors of contributed papers to provide a short video describing their work. These videos will be made

available on the workshop’s website. All presenters will be required to attend the virtual workshop to engage with the audience during the panel discussions and the virtual chat rooms’ discussions.

Table 1: Tentative workshop schedule (time is in GMT)

Time	Event	Time	Event
13:00	Welcome	16:00	Coffee break 2
13:10	Keynote 1	16:15	Keynote 7
13:30	Keynote 2	16:35	Keynote 8
13:50	Keynote 3	16:55	Keynote 9
14:10	Spotlight talks	17:15	Coffee break 3
14:30	Coffee break 1	17:20	Keynote 10
14:40	Keynote 4	16:40	Keynote 11
15:00	Keynote 5	18:00	Panel discussion 2
15:20	Keynote 6	18:20	Wrap-up
15:40	Panel discussion 1	18:30	Virtual drinks (optional)

6 POST-WORKSHOP PLANS

The workshop’s outcome will be communicated to a larger audience by producing a technical report, which will be based on the pre-workshop survey, and will include recommendations and conclusions resulting from the workshop. The report will be submitted to ACM Interactions. The report and the video records of the presentations will be made available on the workshop’s website.

7 CALL FOR PARTICIPATION

We propose a one-day virtual workshop, titled ‘Towards Explainable and Trustworthy Autonomous Physical Systems (ETAPS)’, as a venue for social scientists, AI researchers, and practitioners to discuss issues surrounding the design, implementation, and deployment of autonomous physical systems in society. The goal of this workshop is to identify the current and impending issues surrounding the general acceptance of the current state-of-the-art autonomous physical systems in key domains such as transportation, healthcare, and education. In the current pandemic, guidelines to enforce restricted movements are being set by many governments with the aim of preventing the spread of COVID-19. This has led to new considerations for the deployment of autonomous physical systems to perform certain tasks in the absence of human experts. The risk of deploying these systems as well as measures to make them explainable and trustworthy will be addressed in the workshop.

Presentations and extended abstracts will be around the following topics of interest:

- Accountability and trust in autonomous systems
- Algorithmic transparency and accountability
- Human factors in explanation generation and presentation
- AI ethics

- Insights on explainability from the social sciences
- Explainable and expressive autonomous systems
- Explainable planning
- Explicability, readability, legibility
- Context-aware and situation-aware explanations
- Interaction design and explainable autonomous systems
- Personalised explanations
- Explanations for non-experts
- And other closely related topics.

ACKNOWLEDGMENTS

The organisers acknowledge the support by the UK’s Engineering and Physical Sciences Research Council (EPSRC) through project RoboTIPS: Developing Responsible Robots for the Digital Economy, grant reference EP/S005099/1. They also thank the Assuring Autonomy International Programme, a partnership between Lloyd’s Register Foundation and the University of York.

REFERENCES

- [1] T Barfoot, J Burgner-Kahrs, E Diller, A Garg, A Goldenberg, J Kelly, X Liu, HE Naguib, G Nejat, AP Schoellig, et al. 2020. Making Sense of the Robotized Pandemic Response: A Comparison of Global and Canadian Robot Deployments and Success Factors. (2020). arXiv:2009.08577
- [2] Cindy L Bethel. 2009. Robots without faces: non-verbal social human-robot interaction. (2009).
- [3] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, Vol. 8. 8–13.
- [4] Simon Burton, Ibrahim Habli, Tom Lawton, John McDermid, Phillip Morgan, and Zoe Porter. 2020. Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. *Artificial Intelligence* 279 (2020), 103201.
- [5] Katherine Rose Driggs Campbell. 2017. *Tools for Trustworthy Autonomy: Robust Predictions, Intuitive Control, and Optimized Interaction*. University of California, Berkeley.
- [6] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. 2018. Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 307–315.
- [7] D de Martini, M Marchegiani, P Newman, M Gadd, and L Kunze. [n.d.]. Sense-Assess-eXplain (SAX): building trust in autonomous vehicles in challenging real-world driving scenarios. In *2020 IEEE Intelligent Vehicles Symposium (IV)*. Institute of Electrical and Electronics Engineers.
- [8] Fotios Dimeas and Nikos Aspragathos. 2016. Online stability in human-robot cooperation with admittance control. *IEEE transactions on haptics* 9, 2 (2016), 267–278.
- [9] Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot. 2018. Making machine learning robust against adversarial inputs. *Commun. ACM* 61, 7 (2018), 56–66.
- [10] Renate Häuslschmid, Max von Buelow, Bastian Pfleging, and Andreas Butz. 2017. Supporting trust in autonomous driving. In *Proceedings of the 22nd international conference on intelligent user interfaces*. 319–329.
- [11] Thomas Hellström and Suna Bensch. 2018. Understandable robots-what, why, and how. *Paladyn, Journal of Behavioral Robotics* 9, 1 (2018), 110–123.
- [12] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. 2017. Explainable Agency for Intelligent Autonomous Systems. In *AAAI*, Vol. 17. 4762–4763.
- [13] Stefanos Nikolaidis, Yu Xiang Zhu, David Hsu, and Siddhartha Srinivasa. 2017. Human-robot mutual adaptation in shared autonomy. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 294–302.
- [14] Tatsuya Nomura and Kayoko Kawakami. 2011. Relationships between Robot’s Self-Disclosures and Human’s Anxiety toward Robots. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Vol. 3. IEEE, 66–69.
- [15] Mohan Sridharan and Ben Meadows. 2019. Towards a Theory of Explanations for Human–Robot Collaboration. *KI-Künstliche Intelligenz* 33, 4 (2019), 331–342.