Interaction Design for Explainable AI

Workshop Proposal

Prashan Madumal
The University of Melbourne
Melbourne, Australia
pmathugamaba@unimelb.edu.au

Ronal Singh
The University of Melbourne
Melbourne, Australia
rr.singh@unimelb.edu.au

Joshua Newn
The University of Melbourne
Melbourne, Australia
joshua.newn@unimelb.edu.au

Frank Vetere
The University of Melbourne
Melbourne, Australia
f.vetere@unimelb.edu.au

ABSTRACT

As artificial intelligence (AI) systems become increasingly complex and ubiquitous, these systems will be responsible for making decisions that directly affect individuals and society as a whole. Such decisions will need to be justified due to ethical concerns as well as trust, but achieving this has become difficult due to the 'black-box' nature many AI models have adopted. Explainable AI (XAI) can potentially address this problem by explaining its actions, decisions and behaviours of the system to users. However, much research in XAI is done in a vacuum using only the researchers' intuition of what constitutes a 'good' explanation while ignoring the interaction and the human aspect. This workshop invites researchers in the HCI community and related fields to have a discourse about human-centred approaches to XAI rooted in interaction and to shed light and spark discussion on interaction design challenges in XAI.

KEYWORDS

Explainable AI, Explainable Interfaces

ACM Reference Format:
Prashan Madumal, Ronal Singh, Joshua Newn, and Frank Vetere. 2018. Interaction Design for Explainable AI: Workshop Proposal. In *Proceedings of the 30th Australian Computer-Human Interaction Conference (OzCHI '18), December 4–7, 2018, Melbourne, VIC, Australia.* ACM, Melbourne, VIC, Australia, 2 pages. https://doi.org/10.1145/3292147.3293450

1 INTRODUCTION

With AI systems becoming an integral part of everyday life, Explainable AI (XAI) will play a pivotal role in promoting *trust*, *transparency*, *confidence* and *ethics* in these systems. The XAI initiation[1] by DARPA led by David Gunning further strengthens this notion and highlights the growing need for XAI systems.

As Gunning [5, 7] proposes, XAI systems should contain two key components: (1) *explainable model* and (2) *explanation interface*. The *explainable model* component is concerned with building more interpretable models of the underlying AI model and generating explanations whereas the *explanation interface* focuses on transferring the generated explanations to the end-user. While explainable models have been widely explored in the AI community with respect to machine learning (e.g. [4]) and AI planning (e.g. [1, 2, 6, 11]), explanation interfaces have been underexplored. This sets the premise of our proposed workshop, where we aim to have a discourse surrounding the explanation interface and the harmony that it should maintain with explainable models.

2 MOTIVATION

AI community at large has embraced XAI in recent years, evidenced from the ongoing interest shown in high-profile AI and HCI venues since the first workshop on XAI[2] in IJCAI 2017. The XAI[3] and Interpretability[4] workshops ran for three days in IJCAI/ICML/AMAAS 2018. At IUI 2018, included a paper session on Explainable IUIs, and CHI 2018, saw two paper sessions on Explaining and Explainable Systems[5] and Explaining Players[6] with papers focusing on explanation interaction of agents [3, 10].

As Miller [9, pg 10] states, the process of explanation is inherently interactive and social with two processes. First, a *cognitive process*, namely the process of determining an explanation for a given event, called the *explanandum*, in which the causes for the event are identified, and a subset of these causes is selected as the explanation (or *explanans*). Second, the *social process* of transferring knowledge between explainer and explainee, usually an interaction between a human, in which the goal is that the explainee has enough information to understand the causes of the event. More recently Madumal et al [8] also explored the dialogical aspect of explanations noting the importance of dialogue structures and interactions in XAI systems.

In this workshop, we will address the topics related to the design of human-computer interfaces for XAI. Effective knowledge transfer through an explanation depends on a combination of explanation dialogues, psychological and philosophical theories on explanations and interfaces that can accommodate explanations.

[1] https://www.darpa.mil/program/explainable-artificial-intelligence

OzCHI '18, December 4–7, 2018, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-6188-0/18/12.
https://doi.org/10.1145/3292147.3293450

[2] http://home.earthlink.net/ dwaha/research/meetings/ijcai17-xai/
[3] http://home.earthlink.net/ dwaha/research/meetings/faim18-xai/
[4] https://sites.google.com/view/whi2018/home
[5] https://chi2018.acm.org/technical-program/?sessionId=-L6UzqrouPr-ofgv6jrI
[6] https://chi2018.acm.org/technical-program/?sessionId=-L6UzvPWF03Xk78XD-f8

The aim is to explore explanation interfaces that enable users to understand and interact with intelligent systems, which ultimately promotes trust among the users and the intelligent system.

This workshop invites researchers in the HCI community and related fields to have a discourse about human-centred approaches to XAI rooted in interaction. We welcome multidisciplinary contributions that inform or intersect with XAI. These include but are not limited to human-human interaction, human-computer interfaces, human factors, computer sciences, cognitive science, interactive design, theoretical approaches of explainability and transparent AI, fairness, accountability, and trust.

We also seek submissions that contribute to answering some of the fundamental questions raised by other researchers in the field such as:

- What is an explanation? What should they look like?
- What, when and how to explain?
- How to evaluate explanations or how the explanation is provided?

3 PARTICIPANTS

In this workshop, we want to bring together researchers and practitioners from the HCI community, and researchers and practitioners who are interested in explainable AI and in particular, human-computer interfaces for explainable AI.

Participants will benefit from community engagement and discussion of concepts regarding explainable systems within the field of HCI. The outcome of these discussions will benefit not only the OzCHI community but also the broad AI and HCI community that is actively pursuing research in the field of explainable AI.

4 WORKSHOP ORGANIZERS

Prashan Madumal is a PhD Candidate at the A.I. and Autonomy Lab and the Microsoft Research Centre for Social NUI at The University of Melbourne. He is under the supervision of A/Prof. Tim Miller, Prof. Frank Vetere, Prof. Liz Sonenberg. He has worked as a software engineer previously in Sysco Labs in Sri Lanka. His current research interests are focused on Explainable AI and Human-Agent Interaction.

Ronal Singh is a Research Fellow in Human Agent Collaboration at the Microsoft Research Centre for Social NUI at the University of Melbourne. Ronal recently completed his PhD in the School of Computing and Information Systems at the university. As part of the Human Agent Collaboration project, he is investigating the use of AI planning and eye gaze to improve the interactions between an intelligent agent and a human user. His research interests are in human-agent teamwork and multi-modal human-agent interactions.

Joshua Newn is a PhD Candidate at the Microsoft Research Centre for Social NUI at The University of Melbourne under the supervision of Prof. Frank Vetere and Dr. Eduardo Velloso. His current research interests aim to advance the use of eye tracking in social contexts to enrich the interaction between humans and intelligent user interfaces; allowing systems to make social inferences through our gaze behaviours.

Frank Vetere is a Professor at the School of Computing and Information Systems at the University of Melbourne. He directs the Microsoft Research Centre for Social NUI and leads the Interaction Design Laboratory. His expertise is in Human-Computer Interactions (HCI) and Social Computing, with interests in design-thinking and in technologies for ageing well. His research aims to generate knowledge about the use and design of information and communication technologies for human wellbeing and social benefit. He applies human-oriented design techniques, interpretations of ethnographies, and evaluation of technologies (field-based studies and lab-based experiments) to create knowledge about the design and use of ICTs.

REFERENCES

[1] Joost Broekens, Maaike Harbers, Koen Hindriks, Karel Van Den Bosch, Catholijn Jonker, and John-Jules Meyer. 2010. Do you get it? User-evaluated explainable BDI agents. In *German Conference on Multiagent System Technologies*. Springer, 28–39.

[2] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. 2017. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. *IJCAI International Joint Conference on Artificial Intelligence* (2017), 156–163. https://doi.org/10.24963/ijcai.2017/23 arXiv:1802.01013

[3] Jonathan Dodge, Sean Penney, Claudia Hilderbrand, Andrew Anderson, and Margaret Burnett. 2018. How the Experts Do It: Assessing and Explaining Agent Behaviors in Real-Time Strategy Games. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 562, 12 pages. https://doi.org/10.1145/3173574.3174136

[4] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. Ml (2017), 1–13. arXiv:1702.08608 http://arxiv.org/abs/1702.08608

[5] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces*. ACM, 211–223.

[6] Maria Fox, Derek Long, and Daniele Magazzeni. 2017. Explainable Planning. *IJCAI - Workshop on Explainable AI* (2017).

[7] David Gunning and Darpa Io. [n. d.]. Explainable Artificial Intelligence (XAI) Explainable AI - What Are We Trying To Do ? ([n. d.]), 1–18.

[8] P. Madumal, T. Miller, F. Vetere, and L. Sonenberg. 2018. Towards a Grounded Dialog Model for Explainable Artificial Intelligence. *ArXiv e-prints* (June 2018). arXiv:cs.AI/1806.08055

[9] Tim Miller. 2017. Explanation in Artificial Intelligence: Insights from the Social Sciences. (2017). https://doi.org/arXiv:1706.07269v1 arXiv:1706.07269

[10] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 649, 13 pages. https://doi.org/10.1145/3173574.3174223

[11] Michael Winikoff. 2017. Debugging Agent Programs with Why?: Questions. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS '17)*. IFAAMAS, 251–259.