Conversational Explanations: Explainable AI through human-machine conversation

Dave Braines*,†
*Emerging Technology, IBM Research UK,
†Crime and Security Research Institute, Cardiff University, UK

***Abstract*—Explainable AI has significant focus within both the research community and the popular press. The tantalizing potential of artificial intelligence solutions may be undermined if the machine processes which produce these results are black boxes that are unable to offer any insight or explanation into the results, the processing, or the training data on which they are based. The ability to provide explanations can help to build user confidence, rapidly indicate the need for correction or retraining, as well provide initial steps towards the mitigation of issues such as adversarial attacks, or allegations of bias. In this tutorial we will explore the space of Explainable AI, but with a particular focus on the role of the human users within the human-machine hybrid team, and whether a conversational interaction style is useful for obtaining such explanations quickly and easily.**

I. Tutorial

The tutorial is broken down into three broad areas which are dealt with sequentially:

A. Explainable AI

What is it? Why do we need it? Where is the state of the art? Starting with the philosophical definition of explanations [1] and the role they serve in human relationships, this will cover the core topic of explainable AI [2], looking into different techniques for different kinds of AI systems, different fundamental classifications of explanations (such as transparent, post-hoc and explanation by example) and the different roles that these may play with human users in a human-machine hybrid system. Examples of adversarial attacks and the role of explanations in mitigating against these will be given, along with the need to defend against bias (either algorithmic or through training data issues).

B. Human roles in explanations

Building on the work reported in ”Interpretable to whom?” [3] this section examines the different roles that a human (or machine) user within the system may be fulfilling, and why the role has an important part to play in determining what kind of explanation may be required. In almost all current AI explanation-related research the role of the user is not a primary consideration, but we assert that the ability to create a meaningful explanation must take this into account. The goals of the users will vary depending on their role, and the explanations that will serve them in achieving these goals will also vary.

C. Conversational explanations

The role of conversational machine agents (such as Alexa, Siri and Google) are becoming increasingly commonplace, but the typical interactions that these agents fulfil are fairly simple. Conversational interactions can be especially useful in complex or evolving situations where the ability to design a rich and complete user interface in advance may not be possible. In our ongoing research we are investigating the role of a conversational interaction with AI explanations and will report the findings so far in this section. There will also be a live interactive demo for optional use by the audience during this session.

II. Intended Audience

The intended audience for this tutorial are researchers in any field where complex algorithms or processes can be used to inform human decision-making. The participants will be taken through a general overview of explanation in both human and machine contexts, and how the role of the agent will have a significant impact on what kind of explanation might be useful. The workshop will then move into some ongoing research into the role of conversation as a tool to enable explanations in human-machine hybrid systems, along with an interactive demonstration of an early version of this capability.

III. Instructor's biography

Dave Braines is the Chief Technology Officer for Emerging Technology, IBM Research UK, and is a Fellow of the British Computer Society. As a member of the IBM Research division he is an active researcher in the field of Artificial Intelligence and is currently focused on Machine Learning, Deep Learning and Network Motif analysis. He has published over 100 conference and journal papers and is currently the industry technical leader for a 10 year research consortium comprised of 17 academic, industry and government organizations from the UK and US. Dave is passionate about human-machine cognitive interfaces and has developed a number of techniques to support deep interactions between human users and machine agents.

Since 2017 Dave has been pursuing a part-time PhD in Artificial Intelligence at Cardiff University, and in his spare time he likes to get outdoors for camping, walking, kayaking, cycling or anything else that gets him away from desks and screens!

Acknowledgements

This research was sponsored by the U.S. Army Research Laboratory and the UK Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the UK Ministry of Defence or the UK Government. The U.S. and UK Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copy-right notation hereon.

References

[1] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, no. 1, pp. 1–38, 2019.

[2] C. Molnar, *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/, 2019, https://christophm.github.io/interpretable-ml-book/.

[3] R. Tomsett, D. Braines, D. Harborne, A. Preece, and S. Chakraborty, "Interpretable to whom? a role-based model for analyzing interpretable machine learning systems," *arXiv preprint arXiv:1806.07552*, 2018.