A Systematic Method to Understand Requirements for Explainable AI (XAI) Systems

Mark Hall, *Airbus Central R&T, UK*
Daniel Harborne, *Cardiff University, UK*
Richard Tomsett, *IBM Research, UK*
Vedran Galetic, *Airbus Central R&T, UK*
Santiago Quintana-Amate, *Airbus Central R&T, UK*
Alistair Nottle, *Airbus Central R&T, UK*
Alun Preece, *Cardiff University, UK*

Abstract*—This paper presents a five-step systematic method in the development of an explainable AI (XAI) system, to (i) understand specific explanation requirements, (ii) assess existing explanation capabilities and (iii) steer future research and development in this area. A case study is discussed whereby the method was developed and applied within an industrial context. *This paper is a summary of research originally published at the XAI workshop at IJCAI, 2019 [1].

I. Introduction

Requirements for explainable artificial intelligence (XAI) systems are dependent upon the application and to whom the explanations are intended for [2], [3]. There are significant research efforts toward developing new techniques to make AI systems more explainable. Simultaneously, there is a growing body of research into metrics and ways in which such explainable methods and tools may be formally evaluated [4], [5]. In practice it is challenging to directly compare the effectiveness of these explainable techniques, without a formal set of requirements with respect to a given application [6]. Furthermore, a key challenge exists in understanding unique requirements for explanations within AI systems [7]. Thus, if research is to be undertaken that applies to real-word industry problems for developer and user communities, then efforts need to be aligned to understand requirements for explainable AI systems.

The contribution of this research is to provide an industry-based engineering foundation to steer future AI research. It addresses a key gap in the literature towards understanding the requirements of stakeholders for XAI systems. The paper presents a systematic method that can be used in the development of an explainable AI system, to (i) understand specific explanation requirements, (ii) assess existing explanation capabilities and (iii) steer future research and development in this area.

II. Developing the Method: Case Study

A. Method

There is still debate surrounding the meaning of key terms such as explainability and interpretability within the literature, with terms often used interchangeably. For the purposes of this research, a simplistic and pragmatic conceptual model was developed from the literature to provide clarity to apply this to an industrial case. This is shown in Figure 1 and described below. In this context, an *outcome* refers to the result of an action or process. For example, an outcome could be an explanation artifact or the resultant change in the mental state of an explainee. In addition, *evaluation* here refers to the assessment of each of these outcomes.

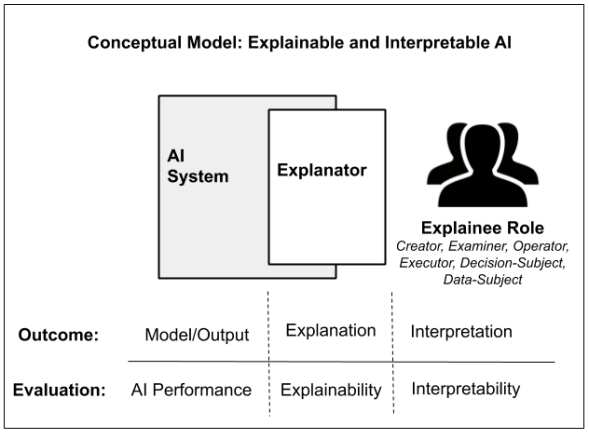

Fig. 1. Explainable & Interpretable AI Conceptual Model

- The outcome of an *AI system* is a *model* or some decision *output*, evaluated in terms of *AI system performance*.
- An *explanator* forms part of the AI system which generates *explanation* artifacts, which in turn are evaluated in terms of *explainability* effectiveness.
- An *explainee agent* will consume *explanation* artifacts, and their ability to understand this is evaluated in terms of explanation *interpretability*.

Once this conceptual model was adopted, the following five-step method was then developed to be applied within the industrial context, in order to understand requirements for an explainable AI system.

Five-Step Method:

1) Determine the relevant explainee roles within the ecosystem [2].

2) Determine the relevant explanation characteristics.
3) Capture explanation requirements from individuals related to the specific roles. This can be undertaken by a more traditional requirements engineering process, or applying agile methods for requirements elicitation.
4) Assess the ability of appropriate explainable methods to meet these requirements.
5) Map existing explainable techniques to XAI system requirements. This identifies which existing techniques address specific requirements, and highlight any gaps in capabilities to steer further research and development for the given application.

This systematic method was developed and followed within an industrial research project, to understand the requirements for designing an explainable AI system. The AI system was developed as a proof-of-concept by data science researchers alongside domain experts within the engineering discipline.

B. Application

End users for an AI system within the company were identified and their role labels within the ecosystem were established. Semi-structured interviews then took place with these subject matter experts, which enabled initial insights to be gained into the differing types of requirements. A total of 12 individuals were formally interviewed, supported by various informal discussions that took place. The individuals were all experienced aerospace engineers with a variety of engineering responsibilities including system design, testing, verification and validation. The formal interviews ranged between 45-60 minutes, and the discussion was facilitated by the researcher. It became clearer through these discussions what the differing needs would be. Through the course of these discussions, this activity helped to provide some clarity to the situation. However, it became clear that there needed to be a common lexicon to develop a shared understanding if formal requirements were to be developed for each role type. This observation was addressed by deriving a set of explanation characteristics, both from the literature and characteristics that emerged from these interviews.

C. Discussion

Method Strengths:

- Identifying key stakeholders and their roles within the AI ecosystem, as outlined by [2] proved to be beneficial in practice to recognise differing end-user needs early in the process.
- The explanation characteristics proved to be effective for framing discussion between data science researchers and subject matter experts. The initial semi-structured interviews that were conducted without these characteristics did not result in clear requirements, demonstrating the need for a set of established explanation characteristics that may be systematically addressed.
- The explanation characteristics also provided the means by which to assess the capability of existing explainable techniques in XAI.

Method Limitations:

- Throughout the discussions with individuals representing different role types, it became evident that requirements differed depending on the following, which would benefit from further consideration:
  - AI Application The requirements would differ if the system were to be a decision-support machine learning system, or embedded as an autonomous component within wider engineering systems.
  - Criticality - The level of risk associated with the AI system influences the explainability requirements.
  - System life cycle stage - The discussions highlighted that a specific role may need to undertake tasks at different points in the life cycle.
- Explanation characteristics - These would benefit from being developed further to provide greater granularity and support a more comprehensive assessment of explainable techniques.

III. Conclusion

This paper presented a five-step systematic method for understanding XAI requirements, that was developed and applied for an industrial AI system. It established an approach to understanding explainable requirements in a given AI system. Future work should include application of this method to a variety of scenarios, and further refinement in an iterative manner.

Acknowledgment

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] M. Hall, D. Harborne, R. Tomsett, V. Galetic, S. Quintana-Amate, A. Nottle, and A. Preece, "A systematic method to understand requirements for explainable ai (XAI) systems," in *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI)*, 2019.

[2] R. Tomsett, D. Braines, D. Harborne, A. Preece, and S. Chakraborty, "Interpretable to whom? A role-based model for analyzing interpretable machine learning systems," *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*, pp. 1–8, 2018.

[3] D. Bohlender and M. Köhl, "Towards a characterization of explainable systems," *arXiv:1902.03096*, 2019.

[4] S. Mohseni, N. Zarei, and E. D. Ragan, "A survey of evaluation methods and measures for interpretable machine learning," *arXiv:1811.11839*, 2018.

[5] D. Gunning, "DARPA's explainable artificial intelligence (XAI) program," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, ser. IUI '19, 2019.

[6] F. Doshi-Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," *arXiv:1702.08608*, 2017.

[7] C. T. Wolf, "Explainability scenarios: Towards scenario-based XAI design," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, ser. IUI '19, 2019, pp. 252–257.