Explainable Artificial Intelligence: What Do You Need to Know?

Sam Hepenstal¹(✉) and David McNeish²

¹ Defence Science and Technology Laboratory, Porton Down, Salisbury SP4 0JQ, UK
shepenstal@dstl.gov.uk

² Defence Science and Technology Laboratory, Portsmouth West, Fareham PO17 6AD, UK
dgmneish@dstl.gov.uk

Abstract. In domains which require high risk and high consequence decision making, such as defence and security, there is a clear requirement for artificial intelligence (AI) systems to be able to explain their reasoning. In this paper we examine what it means to provide explainable AI. We report on research findings to propose that explanations should be tailored, depending upon the role of the human interacting with the system and the individual system components, to reflect different needs. We demonstrate that a ‘one-size-fits-all’ explanation is insufficient to capture the complexity of needs. Thus, designing explainable AI systems involves careful consideration of context, and within that the nature of both the human and AI components.

Keywords: Explainable AI · Algorithmic transparency · Interpretability

1 Introduction

The use of artificial intelligence (AI) to support high risk and high consequence decision making faces a significant challenge; can the system outcomes and the processes by which they are reached be explained to the human decision maker? This is particularly relevant within the defence and security domains where decisions have profound consequences and may, for example, result in detention, injury or even death. Human accountability for these types of decisions is of paramount importance, but is dependent on the decision making process being intelligible to human agents (Floridi and Cows 2019). Based on recent research by the authors, this paper seeks to address the following questions: What requires an explanation in order to achieve explainable AI? And how might explanation needs vary across different user roles or system components? This research focusses on the defence and security domain but the findings are likely to be generalizable to the use of AI to support high risk and high consequence decision making more broadly.

2 There Are Various Interpretations of Explainable AI

Explainable AI (XAI) is a slippery notion which can refer to several distinct concepts (Lipton 2016) and there is as yet no widely agreed definition. As pointed out by Preece

(2018) the desire to understand and explain the outputs from AI-based systems is not new, although recent developments in the field of Machine Learning (ML) have led to a resurgence in interest. Despite the abundance of overlapping terminology within the field of XAI, this section aims to describe some of the key concepts which have emerged from the literature.

Hoffman et al. (2018) describe the aim of XAI as helping human agents develop a strong mental model of the system in order to provide one or more of the following benefits; justification for individual outputs, an understanding of how the system works in general including its capabilities and limitations, and the ability to accurately predict what it would do given certain conditions. According to Hoffman and Klein (2017), from a psychological and philosophical perspective, explanation is closely related to causal reasoning and includes abduction, retrospection (including counterfactual reasoning) and prospection (projection into the future). Causal reasoning is central to sense-making, the development of mental models, decision making, re-planning, coordination and anticipatory thinking. In the broadest sense, XAI seeks to answer questions such as: What is the system doing? How is it doing it? Why is it doing it? And what would it do given a certain set of conditions? (Hoffman et al. 2018; Weller 2017).

An important distinction to draw is between local and global explanation. Local explanation focusses on justifying a single decision or output, whilst global explanation focusses on overall system behaviour (Hoffman et al. 2018; Weller 2017; Doshi-Velez and Kim 2017; Ribeiro et al. 2016). XAI is often described in terms of transparency and post-hoc explanation (Preece et al. 2018; Mittelstadt et al. 2019; Lipton 2016; Preece 2018). Transparency seeks to reveal information about the internal structure of a model and its training data in order to communicate how the system reaches an output (Tomsett et al. 2018, Preece 2018). Post-hoc explanations aim to provide information about the cause or reason for an output (Tomsett et al. 2018) generally without referring to the internal structure of the model (Preece 2018). Or put another way, transparency focusses on how the system works internally whilst post-hoc explanations address how the model behaves (Mittelstadt et al. 2019). These ideas are echoed within the framework (Fig. 1)

Fig. 1. Algorithmic transparency framework (Hepenstal et al. 2019).

proposed by Hepenstal et al. (2019) which distinguishes between the ease with which a user can explain the results provided by a system, and the ability to inspect and verify the goals and constraints of the system within context.

3 Roles Influence Explanation Needs

Ultimately explanation needs to be context specific; the goals, motivations, needs, capabilities and limitations of the target audience must be considered (Tomsett et al. 2018; Weller 2017; Ribeiro et al. 2016). Drawing on the division between expert and lay users proposed by Ras et al. (2018), as well as the simple distinctions made in early AI research (Preece 2018), the stakeholders who need XAI can be grouped into two broad categories; developers and users. As presented in Fig. 2, the terms developers and users are used by the authors to mean those who influence system design and those affected by the system respectively.

Fig. 2. Two broad categories of stakeholders requiring XAI.

Our research sought to investigate the explanation needs of these two stakeholder groups through a participatory design workshop. This involved twelve participants with either a military or AI background, from either the developer or user stakeholder communities. The twelve participants were split into three groups of four with a mixture of backgrounds in each to encourage interdisciplinary working. The workshop focussed on two hypothetical scenarios which were used to stimulate and focus discussion by illustrating the possible application of AI decision support in a military context. These scenarios were not intended to represent any in-service or future systems used by the UK MOD. The first scenario focussed on a system which analysed data from sensors worn by infantry soldiers in order to support task allocation decisions made by their commander. The second focussed on a system which analysed ship sensor information in order to detect and classify likely incoming threats and support the commander in choosing the best course of action. The workshop approach was based on the design thinking process and built upon the methods described by McNeish and Maguire (2019).

The participants felt that explainability was highly important for both developers (those who influence system design) and users (those affected by the system) of AI-based systems. Two reasons for its importance, common to both stakeholder groups, were regulatory and legal compliance and the ability to investigate when something goes wrong. Whilst developers need explainability for testing and iterative improvement of

system performance, users need to build their trust and confidence in the system when optimising and being able to justify their decision making.

Five key stakeholder needs emerged and are summarised in Fig. 3. Each of these attributes is considered from the perspective of the two stakeholder groups. We draw out the differences, if any, and provide a rationale as to why that may be the case. Overall the most common theme was the need to understand the input data and sensor performance, with almost twice as many comments attributed to it than any other individual theme. The remaining four received almost the same number of comments as one another. These five areas of explanation needs are not unique to military applications of AI systems. However the diverse and dynamic nature of military operations, including the potential for lethal force, may drive their particular relevance within the military domain.

Fig. 3. Proportion of comments attributed to each stakeholder group for each theme.

Although the need to understand the input data and sensor performance was considered relevant to both stakeholder groups, it was attributed more often to developers. Both users and developers may need to know the number and type of sensors providing the data used as inputs to the system, whether or not these sensors are working, the impact of the operational environment on sensor reliability, and their operational range. Ultimately they are likely to want to know how all of these factors could impact the accuracy and reliability of the outputs from a decision support system. They also need to know, particularly from a developer’s perspective, whether the right parameters are being measured; in other words how reliable the input metrics are as indicators for the task performance measures being predicted. Users in particular need to understand the ‘completeness’ of the input data being used as a basis for predictions including whether any particularly important sensors were not being used. From a developer perspective understanding the minimum number and type of information sources required to produce accurate and reliable predictions is important, particularly when used to inform

safety critical decisions. When using multiple data sources as inputs, understanding the relative priority of each and its influence on prediction accuracy would be important for both developers and users, particularly if different data sources seem to contradict. In summary, both developers and users will need an understanding of the provenance of the data being used to train the system and the inputs used to inform real-time predictions.

Building an appropriate level of confidence and trust in the system was a challenge most frequently associated with the user stakeholder group. This included the level of confidence or uncertainty in the data sources and the system outputs and across different types of scenario. Both justification for individual outputs as well as an understanding of the system reliability across multiple uses were seen as necessary for informing this appropriate level of user confidence and trust.

Both developers and users need to understand the capabilities and limitations of the system when used in different operational contexts, and therefore how generalizable the system might be. Contextual factors could include different physical environments, missions and tasks, types of threat, and interoperability with other systems (for example during coalition operations). These factors may have implications for what training data should be used and the resultant prediction accuracy. Also related to this theme was the need to understand how the system is likely to respond to an unexpected or unknown threat.

The need to inform the verification and validation activities was a common theme across both scenarios. In particular this included understanding the sufficiency of test cases used during the verification of the system as well as the characteristics of the training and input data. Explanation may be required when validating the system against the user requirements; in particular, the ability to assess its impact on overall mission performance and the face validity of information being presented to the user. Another specific concern was understanding a system which adapts and learns over time; an online learning system for example. This may require some form of through-life assurance which would require the system to enable both developers and users to understand how it has changed and whether this might impact the system safety argument.

Finally, authority and accountability and was the most common theme within the second scenario which focussed on threat detection in a maritime environment. Both developers and users need to understand who is accountable for safety-critical and mission-critical decisions. Who is responsible for decisions and how much authority the human and the system should have over individual actions must be clear in order to inform design decisions relating to human-machine interaction, assessment by regulators, and to ensure the users understand what is expected of them. This is particularly important from a legal perspective and for the management and application of rules of engagement to ensure that there is sufficient human control over the ultimate outcome. Although the legal and ethical drivers for XAI are mentioned in previous research (Lipton 2016; Samek et al. 2017), the nature of military decision making which by definition could lead to the use of lethal force, adds significant weight to this challenge and has legal implications relating to international humanitarian law and the application of rules of engagement. Put simply the severe consequences of inappropriate levels of trust in a ML system drive the requirement for clarity over decision making roles.

4 System Components Influence Explanation Needs

Real world systems are complex, often incorporating many algorithms and components. Our research to date has considered a conversational agent (CA) system for information retrieval, to support criminal investigations.

When a user interacts with the CA they trigger a series of processes. Firstly, entities are extracted from their input text which can be later used in a database query. Next, the users input text is matched to a CA intention, which defines how the query will be structured, managed, and responded to. An intention is a collection of functional attributes and each of these has its own algorithmic processes and models. Additionally, the CA is interacting with a database, which has a defined structure and a degree of completeness and currency.

To date, much research into the area of XAI (Gunning 2019) has focussed upon ML systems for image classification, where there is a single model that is being explained with a single explanation. In these cases a typical approach is to identify the features in the image that are important in defining the classification, for example as a heat map over the image. Tools such as LIME (Ribeiro et al. 2016) allow users to inspect significant features visually, by highlighting them. These feature based approaches present what Gilpin et al. (2018) describe as a combination of interpretability and completeness, where interpretability is linked to explaining the internals of a system and completeness is to describe them as accurately as possible. The internals are captured by identifying and conveying the features within the input or output data itself. For example, if a user wishes to classify images of cats, a heat map layer that highlights the image pixels in the input images that are crucial to return the classification of ‘cat’, provides some explanation of the internals of the system.

Providing explanations becomes more complicated when applied to a real world system, such as our CA. As described previously, an interaction with a CA triggers various processes, including models for how to retrieve information. An explanation of the model used can be provided through the input or output data, to meet the needs described by Gilpin et al. (2018).

Table 1. Example Conversational Agent transcript (synthetic data).

Stage	Agent	Text entry
1	Human	Can you find people who are similar to Paul Peterson?
2	Machine	Yes, Sam Smythe is the most similar person to Paul Peterson. They are both employees of Dean Lane Services and have been offenders in criminal activities. Both have been involved in activities related to drugs. They both own vehicles

For example, in Table 1, the CA can respond and explain why ‘Sam Smythe’ has been chosen as similar to ‘Paul Peterson’. It does so using features from data within the output, which are the entities and their relationships with one another. Previous research

to develop an algorithmic transparency framework (Hepenstal et al. 2019) has identified the need to provide visibility of system goals and constraints, in addition to this explanation of the response data, so that a user can have a better understanding of when system outputs can be used. A further issue is that the explanation in Table 1 is related only to the information retrieval model used to return the response. If a user wanted to understand how to trigger an alternative intention, they could not grasp this from the explanation given because the intention classification involves an entirely different and separate system component. We would need to provide an additional feature based explanation for the ML classification of a user's intention, by highlighting the words that are important in the classification. With this simple example, we demonstrate that for real world systems explanations are not universally applicable, nor appropriate, across distinct system components and should therefore be tailored carefully.

To capture user needs for explanations of CA systems we conducted interviews with four intelligence analysts, each with more than 10 years of experience in an operational role. Our focus was upon using a CA for information retrieval tasks and interviews therefore covered a scripted investigation scenario, where an analyst is asking questions of a CA. Each interview lasted approximately an hour. We presented interviewees with a series of questions and corresponding CA responses with two explanation conditions. For one condition, responses described the data alone (Table 2) and in the other condition, responses described the data and the system process (Table 3). All data included in the text is fictional. For both conditions, we provided an accuracy measure for concept match confidence in the form of a numerical value between zero and one. We switched the order in which we presented conditions to analysts, so that two analysts saw condition 1 then 2, and two saw condition 2 then 1. We did this to help mitigate any ordering effects, for example, when we asked analysts to identify additional understanding needs or to compare the two conditions. We were not attempting to test the differences between the two conditions; rather we used the conditions as a starting point from which we could explore additional needs with the analysts.

Table 2. CA condition 1: Data only explanation

Q: Are any suspects connected to James White?
A: Yes, James White is connected to Frank Howes and Paul Keen

Table 3. CA condition 2: Data and system explanation

Q: Are any suspects connected to James White?
A: Yes, James White is connected to Frank Howes and Paul Keen James White was victim in an assault where Frank Howes and Paul Keen were suspects Concept match confidence = 0.92 I am looking for a connecting path between James White and suspects (Descriptor) I am looking for single shortest paths and therefore, will not consider more complex connections

We first reduced the interview transcript data to identify a total of 114 distinct statements made by the analysts, with counts for each analyst ranging from 24 statements to 34. To analyse the statements we used an approach called Emergent Themes Analysis (ETA), as described by Wong and Blandford (2004, 2002) where broad themes, which are similar ideas and concepts, are identified, indexed and collated.

Through rigorous coding of the data, we identified that analyst statements could be broadly associated with the core functional components of a CA that need explaining, for example entity extraction from the user’s query or the system processes applied. This is an interesting finding, indicating that analysts have specific considerations for each function of a CA. We should not treat a user’s understanding of the data in the same way as we provide an understanding of the intention classification, or the extracted entities, the response, or system processes. From broad CA component themes, we have drawn out the detailed understanding needed by an analyst in an explanation. These are the sub-themes and, in Table 4, we show that explanations can be tailored for each system component by placing greater emphasis on the relevant sub-themes. For example, system explanations require some understanding of the data, including clarification of source, currency and structure.

Table 4. CA component core understanding needs.

CA component theme	Sub-theme (common for multiple analysts)	Summary of sub-theme(s)
CA intention interaction	Clarification (3), Continuation (2)	Clear language to understand classification (i.e. no confusing response metric) and information to support continuation of investigation
Data	Clarification (3)	Clarification of data updates and source, and data structure to aid forming questions
Extracted entities	Clarification + Verification (3)	More information of entities extracted for clarification and verification
Response	Clarification (4), Justification (4), Exploration (2)	Justification of response with underlying data, clarification of language (not trying to be human) and terminology, ability to explore results in more detail
System processes	Continuation (4), Verification (4), Clarification (3), Exploration (2), Justification (2)	User wants system understanding to support continuation of investigation, to allow them to verify processes are correct and explore them in more or less detail and justify their use/approach and constraints

5 Conclusion

When the diversity of stakeholders and the complexity of AI-based systems with multiple components are taken into account, it is clear that a ‘one-size-fits-all’ approach to XAI is insufficient. Instead the particular needs of the target audience, be they stakeholders who influence the design of the system (developers) or those affected by the system (users), must be actively considered during the design process. Also, rather than focussing on single explanations to explain single models, specific system components may require different forms of explanation. These explanations should focus at both the global and local level, providing explanation of the individual outputs as well as the process by which they have been reached. Importantly, the system goals and constraints should be explained to both groups in order to develop an appropriate level of trust in the outputs. This will also aid recipients when deciding whether or not the system is appropriate to use within a given situation.

In summary, explanation needs in the real world are complex and context dependent. Ongoing work being conducted by the authors is focussed on understanding additional needs for explanations, for example when applied to different tasks, exploring the strengths and weaknesses of different approaches to explaining AI systems, methods for evaluating explanations, and the demonstration of a human-centred design approach to the development of XAI (specifically a conversational agent).

© Crown copyright (2020), Dstl. This material is licensed under the terms of the Open Government Licence except where otherwise stated. To view this licence, visit <http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3> or write to the Information Policy Team, The National Archives, Kew, London TW9 4DU, or email: psi@nationalarchives.gsi.gov.uk.

References

- Blandford, A., Wong, W.: Situation awareness in emergency medical dispatch. *Int. J. Hum Comput Stud.* **61**, 421–452 (2004)
- Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017)
- Floridi, L., Cowls, J.: A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1 (2019)
- Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: an overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, Turin (2018)
- Gunning, D.: DARPA’s explainable artificial intelligence (XAI) program. In: Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI 2019) 2019 New York, NY, USA. Association for Computing Machinery, II
- Hepenstal, S., Kodagoda, N., Zhang, L., Paudyal, P., Wong, W.: Algorithmic transparency of conversational agents. In: Trattner, C., Parra, D., Riche, N. (eds.) *IUI 2019 Workshop on Intelligent User Interfaces for Algorithmic Transparency in Emerging Technologies*, pp. 17–20. Los Angeles, CA, USA (2019)
- Hoffman, R., Klein, G.: Explaining explanation, part 1: theoretical foundations. *IEEE Intell. Syst.* **32**, 68–73 (2017)
- Hoffman, R., Miller, T., Mueller, S.T., Klein, G., Clancey, W.J.: Explaining explanation, part 4: a deep dive on deep nets. *IEEE Intell. Syst.* **33**, 87–95 (2018)

- Lipton, Z.C.: The Mythos of model interpretability. In: 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016). New York, NY, USA (2016)
- Mcneish, D., Maguire, M.: A participatory approach to helicopter user interface design. *Ergonomics & Human Factors* 2019. Stratford-upon-Avon (2019)
- Mittelstadt, B., Russell, C., Wachter, S.: Explaining explanations in AI. In: FAT* 2019: Conference on Fairness, Accountability, and Transparency (FAT* 2019). Atlanta, GA, USA (2019)
- Preece, A.: Asking “Why” in AI: explainability of intelligent systems-perspectives and challenges. *Intell. Syst. Account. Finance Manage.* **25**, 63–72 (2018)
- Preece, A., Harborne, D., Braines, D., Tomsett, R., Chakraborty, S.: Stakeholders in explainable AI. In: AAAI FSS-18: Artificial Intelligence in Government and Public Sector. Arlington, Virginia, USA (2018)
- Ras, G., van Gerven, M., Haselager, P.: Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. In: Escalante, H.J., et al. (eds.) *Explainable and Interpretable Models in Computer Vision and Machine Learning*. TSSCML, pp. 19–36. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98131-4_2
- Ribeiro, M.T., Singh, S., Guestrin, C.: “Why should i trust you?”: explaining the predictions of any classifier. In: KDD 2016 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York (2016)
- Samek, W., Wiegand, T., Muller, K.R.: *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models* (2017)
- Tomsett, R., Braines, D., Harborne, D., Preece, A., Chakraborty, S.: Interpretable to whom? a role-based model for analyzing interpretable machine learning systems. In: 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018). Stockholm, Sweden (2018)
- Weller, A.: Challenges for transparency. In: 2017 ICML Workshop on Human Interpretability in Machine Learning (WHI 2017). Sydney (2017)
- Wong, W., Blandford, A.: Analysing ambulance dispatcher decision making: Trialing Emergent Themes Analysis. HF2002 human factors conference design for the whole person - integrating physical, cognitive and social aspects: a joint conference of the Ergonomics society of Australia (ESA) and the Computer human interaction special interest group, 2002 Canberra, Australia. Ergonomics Society of Australia

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.