Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 207 (2022) 3884–3893

Procedia Computer Science

www.elsevier.com/locate/procedia

26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)

Towards an Integrated Evaluation Framework for XAI: An Experimental Study

Qiyuan Zhang[a,b,c], Mark Hall[d], Mark Johansen[a,b,c], Vedran Galetic[d], Jacques Grange[a,b,c], Santiago Quintana-Amate[d], Alistair Nottle[d], Dylan M Jones[a,b,c], Phillip L Morgan[a,b,c*]

[a]Cardiff University Centre for AI, Robotics and Human-Machine Systems (IROHMS); [b]Cardiff University Human Factors Excellence Research Group (HuFEx); [c]School of Psychology, Cardiff University, Cardiff, CF10 3AT, UK; [d]Airbus Central R&T, Bristol, UK, BS16 1EJ

Abstract

Increasing prevalence of opaque black-box AI has highlighted the need for explanations of their behaviours, for example, via explanation artefacts/proxy models. The current paper presents a paradigm for human-grounded experiments to evaluate the relationship between explanation fidelity, human learning performance, understanding and trust in a black-box AI by manipulating the complexity of an explanatory artefact. Decision trees were used in the current experiment as exemplar interpretable surrogate models, providing explanations approximating black-box behaviour, by means of explanation by simplification. Consistent with our hypotheses: 1) explanatory artefacts brought about better learning, while greater decision tree depths led to greater interpretability of the AI's performance and greater trust in the AI; and 2) explanatory artefacts facilitated learning and task performance even after they were withdrawn. Findings are discussed in terms of the interplay between human understanding, trust and AI system performance, highlighting the simplifying assumption of a monotonic relationship between explanation fidelity and interpretability.

Keywords: : Explainable AI; understandability; interpretability; trust; category learning; decision aids; surrogate models; decision trees; explanation fidelity; explanatory artefacts

[*] Corresponding author. Tel.: (00) 44 (0)29 2251 0784
E-mail address: morganphil@cardiff.ac.uk

10.1016/j.procs.2022.09.450

1. Introduction

AI systems can be distinguished in terms of those that are transparent, in that they provide intrinsically comprehensible explanations, and those that are opaque and do not. As machine learning has become more ubiquitous and powerful, so has an appreciation of the difficulties that arise from the fact that most of the systems do not allow scrutiny of the way they reason. This opaque, 'black-box' character of some AI has the possibility that problems may be solved, but the means by which they are solved may not be clearly interpretable by humans. That such a system cannot (or for some proprietary reason, will not) yield an account of its actions, beyond the simple generic characteristics of the elements comprising the AI black-box, can become a stumbling block to its adoption by human users. Work in the emerging field of Explainable Artificial Intelligence (XAI) [1]–[4] aims to provide explicit explanations for the behaviour of the AI where there were few.

While there is a plethora of methods now available for producing explanations of AI black-box models, studies on the utility and acceptability (e.g. understanding, trust etc.) of this information for human users is relatively rare. It has commonly assumed that any explanation is helpful, but it is not clear what impact it has on users' understanding and the trust with which systems are held. One key goal of this paper is to develop a method to measure explanation quality, and to evaluate the interplay between explanation fidelity, explanation interpretability, understanding and trust.

1.1. Methods of Explaining AI Black-boxes and Decision Trees

As machine learning systems become more complex, it becomes less realistic to expect an exhaustive test of conditions to provide assurance and understanding about its safe and reliable operation. XAI is based on the idea that if the system can provide an explanation of its reasoning, its operational validity can be assessed for a wide range of criteria, ones that go beyond simple indices of performance. For example, as the character of unreliability has been fleshed out, criteria such as avoiding discrimination [5] and avoiding technical debt [6] have grown in importance to the point where the explanation is seen as a right that should be enshrined in law [7].

XAI techniques can generally be categorised as (i) transparent (as opposed to opaque/black-box) and (ii) post-hoc [8], [9]. Transparency refers to the level to which a system provides information about its internal workings or structure and the data on which it has been trained [10]. Post-hoc approaches are distinct since as they involve extracting information from trained AI models. While post-hoc interpretations often do not elucidate precisely how a model works, they may nonetheless confer useful information for practitioners and end-users of AI [9]. One approach to generate post-hoc explanations is to restrict the amount of explanatory information by highlighting only those features that have a big impact on the model's predictions. Two notable examples of feature-importance techniques are LIME [11] and SHAP [12].

Another approach is to provide a simplified 'surrogate' model to approximate black-box AI behaviour in an inherently explainable form. They usually comprise the training of a supplementary machine learning model whose primary aim is to explain the black-box AI's behaviour and rationale. The utilities of surrogate models are partly dependent on their fidelity, defined as: "the proportion of predictions made by the surrogate that match the predictions made by the black-box, while accuracy measures the proportion of predictions made by the surrogate that match the actual value of the target" [13]. One promising approach to producing a proxy model is decision trees, which maps explanations directly onto the architecture of human cognition. Human concepts tend to be structured hierarchically. The harmony of the representational structure with the structure of the world is illustrated by the fact that a decision tree promotes comprehensibility. Its graphical structure allows ease of navigation through the domain. Moreover, the tree focuses attention on the most relevant attributes by presenting only a subset while the relative importance of attributes is indicated by their proximity to the root of the tree. On the basis that hierarchical structures are somehow more inherently compatible with the architectonic structure of the cognitive system, decision trees will be more easily processed and their content and structure more easily remembered than other arrangements. While we do not make an empirical comparison between different surrogate methods, the adoption of decision trees was made in an attempt to provide an easily-digested cognitive presentational format.

Current AI explainability techniques may be ingenious but are of questionable utility. There is little agreement as to what constitutes an explanation, and more importantly, little effort has been invested into whether the explanation can be understood by a human user, whether a specialist or non-specialist. Sometimes the explanations are summaries of predictions made by the model rather than explanations *per se*. Some forms of explanation assume the user has a high level of understanding of the context in which the explanation is provided, that is, it relies on user knowledge to some degree, but assumptions about user understanding may be unwarranted. In particular, the complexity of an explanation also carries with it the possibility that different observers will draw different inferences.

1.2. The DARPA Evaluation Framework to Explainability

An important step towards developing "good" explanations is to establish a universal, context-agnostic evaluation framework which can be applied to all types of explanation methods. The current study draws on the DARPA approach [1], [14], and while arguably incomplete, it serves as a productive starting point for developing a framework in which to situate an assessment of AI black-box understandability. The approach involves taking measurements of three types, arising from corresponding stages in what they refer to as the XAI Process, broadly as follows: The user receives an explanation, from which an assessment can be made of the user's satisfaction as well as the degree to which it meets criteria of "goodness". The explanation should establish or revise the user's mental model whose character can be probed in a variety of ways, but essentially comprises a test of comprehension. As a final stage in the cycle, a test of performance is undertaken; here, the focus is on the improvement in efficiency brought about by progressive refinement of the user's mental model.

A core feature of the DARPA approach is the assessment of the 'goodness' of an explanation. The 'goodness' of an explanation is assumed to engender trust, and the refinement of the mental model engenders appropriate use. Relatively little empirical work has been devoted to understanding what constitutes explanation goodness. One exception is the development of a The Explanation Goodness Checklist [2] that purports to be a means for establishing what makes explanations good and comprises seven statements that require a Yes/No response. This is promising first step in establishing the psychological response to explanations and we adopt it in the present study, with some modifications. We modify the checklist to include a means for the person to make a graded response, instead of saying 'yes' or 'no', as suggested by Hoffman et al. [2]. So, for instance, the question 'The explanation of how the [software, algorithm, tool] work is sufficiently complete' is better expressed by selecting a point on a scale (say, from one to nine, with one end-anchor being 'not at all' and the other 'extremely').

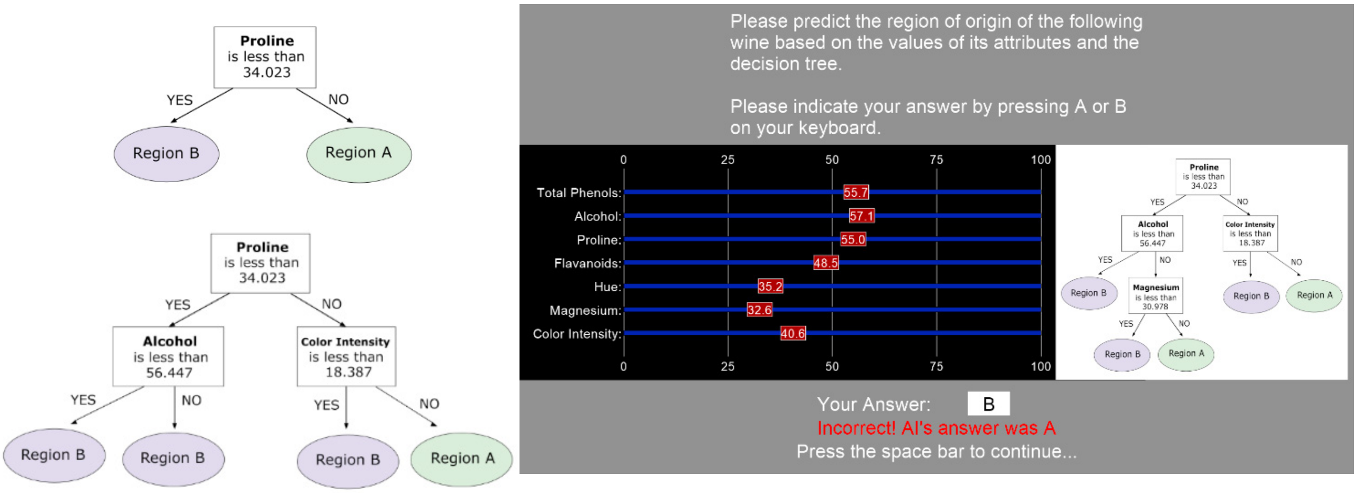
Within the DARPA program, 23 XAI approaches were explored in 11 user studies with approximately 12 700 participants (see [14] for a summary). The key conclusions were that most users liked explanations and found them useful; however, there was often consequential information overload. Many tasks (being too easy) were subject to a ceiling effect—especially true of images that had to be identified—so that explanations did not have a beneficial effect on performance (and only sometimes on positive ratings of use). Explanations sometimes had unforeseen effects suggesting that they do not always cohere logically with the domain of activity. Despite the large scale and wide scope of the program, some types of AI were not considered; for example, there was a bias toward object recognition. The program did not comprise systematic study of different types of explanation (though several types were distinguished in early program documentation). There was a tendency to focus on 'one-shot' explanations, not explanations developing over time in a dynamic sense (that is, the issue of whether explanation can be improved upon by allowing the user extensively interaction with the system was not addressed).

The current paper addresses the above limitations and makes an attempt to develop a unified means of measuring the quality of explanations, we delineated several criteria for experimental studies: 1) The effectiveness of explanations needs to be assessed using both performance measures (prediction accuracy, decision times, and so forth) and self-report judgments including understandability and trust; 2) The studies need to manipulate more than two levels of the degree of explanation complexity/fidelity; 3)The studies need to use larger samples of users to facilitate appropriate statistical conclusions; 4) The samples of users need to be 'naive' inasmuch as they are not the people who developed the system.

1.3. The Present Study

The aim of the present study was to develop and test an empirical paradigm to examine the relationship between explanation fidelity, understanding, learning and trust in AI black-boxes using decision trees as an example. Participants with no experience in AI or computer science were asked to participate in a scenario in which they learnt how the AI used the attributes of instances to make categorizations with or without the assistance of decision trees as explanation artefacts. Then they tried to predict the output of the AI in a classification task. Explanation fidelity was directly manipulated in terms of the depth of the decision tree which acted as interpretable surrogate models to approximate the behaviour of AI black-boxes.

The key research questions were as follows: How does the user's understanding of AI black-boxes relate to the complexity/fidelity of the explanation artefacts for the AI's behaviour? How does the user's task performance relate to the complexity/fidelity of the explanation? How does the user's trust in the AI vary with the complexity of the explanation? To what extent does a decision aid (e.g., a decision tree), facilitate learning and performance after the aid is withdrawn? We had two hypotheses: 1) More detailed explanatory artefacts, operationalized as greater decision tree depths, will correspond to greater interpretability of the AI's performance, better learning and greater trust; 2) More detailed explanatory artefacts will facilitate learning and task performance even after the decision aid is withdrawn, namely cognitive resilience.



Please predict the region of origin of the following wine based on the values of its attributes and the decision tree.

Please indicate your answer by pressing A or B on your keyboard.

Your Answer: B

Incorrect! AI's answer was A

Press the space bar to continue...

Figure 1 Decision trees used by participants in categorization tasks with a depth of one and two (left) and Computer interface of categorisation tasks in learning phase with decision trees of depth of three (right)

2. Methods

2.1. Participants

A total of 57 Psychology undergraduates at Cardiff University completed the experiment as part of a paid participant panel. The experiment was approved by Cardiff University's School of Psychology ethical review process with a full risk assessment.

2.2. Materials

The dataset used in the study is derived from publicly available, real data (Load_Wine Dataset) and initially consisted of 178 samples where each of the samples contains 13 numerical attributes. The output or target column describing the type of wine initially had three categories (A, B and C). Based on pilot data, the difficulty of this version of the learning task was reduced by eliminating category C and cutting the number of attributes from thirteen to the seven most diagnostic, as determined by SHAP feature importance (total phenols, alcohol, proline, flavonoids, hue, magnesium, and color intensity.). The resulting categories contained 130 samples (59 in category A and 71 in category B), each with seven attributes. While the attributes were all on different scales, we normalized them onto scales from 0 to 100 to facilitate their intelligibility. Participants undertook one of four conditions, three decision tree conditions with various tree depths and a control condition (with no decision tree). Figure 1 shows the decision trees used in the experiment proper, with a depth of one, two or three levels, corresponding to the three study conditions.

2.3. Design

The experiment used a mixed design involving one independent variable manipulated between-participants and one within-participant variable. The between-participant variable was the decision tree depth, which had four levels: Tree depths of one, two and three contrasted with a control condition with no tree. The within-participant variable was the degree of decision support, which had three levels: Feedback plus decision tree, decision tree only with no feedback, and no decision tree nor feedback. The key dependent variables were response accuracy (as proportion correct) and decision time, judged goodness of the artefact's explanation of the AI's behaviour and rated trust in the AI black-box.

2.4. Procedure

2.4.1. Learning tasks

Participants were instructed that an AI had been trained to predict the geographic region of a wine based on the values of its seven most important attributes and that the AI's performance was now nearly perfect. They were asked to undertake the "same prediction task as the AI" and attempt "to perform the task as well as the AI". They were instructed that they would be presented with a series of wine instances and they should try to learn to correctly categorise their region based on the displayed numerical attributes using the feedback given. Participants in the three decision tree conditions were introduced to decision trees with instructions on how to use them (details in the materials section). Participants in the control condition were not shown any decision tree. Instead, they were warned that they would initially have to guess categories but that they could learn the task by paying attention to the feedback on whether or not their answer was correct on each trial.

.

All conditions in the experiment, had an initial categorisation training phase, where participants received feedback after each response, followed by a testing phase, without feedback. This testing phase will allow an evaluation of how much learning has taken place, uncontaminated by the presence of feedback. In the training phase, each response was followed immediately by written feedback indicating whether it was 'correct' or 'incorrect' and displaying the correct answer. Each phase had sixty instances presented in random order. Each trial displayed a wine's seven attributes in an integrated numerical and analogue display (Figure 1): *total phenols, alcohol, proline, flavonoids, hue, magnesium, color intensity*. The analogue scales were from 0 to 100 with graduations every 25 values. A particular attribute value was displayed both as a point on the scale and as an actual number (to one decimal place) at that point.

Participants in the control condition saw only the attributes of the wines. However, participants in the non-control conditions had the explanatory artefact, the decision tree, on the screen during the initial training and testing phases, for an example see Figure 1. Those undertaking experimental conditions also did an additional testing phase at the end in which neither decision tree nor feedback was provided.

2.4.2. Judgement tasks

Participants were asked to make judgments about their view of the AI at three points in the procedure – immediately after the exposure to the decision tree (experimental conditions only), after the training phase and after the test phase.

They were asked to make the following judgments on a 9-point judgment scale ranging from 1-not at all, to 9-extremely, which were derived from the Hoffman et al. [2] Explanation Goodness Checklist. Participants in the control and experimental conditions were asked: How well do you understand the AI's categorisation? How well can you now predict the AI's categorisation? How much do you trust the AI? Participants in the experimental conditions were also asked: How helpful is the decision tree to your categorisation? How satisfactory is the decision tree's explanation of the AI's categorisation? How much do you trust the decision tree's explanation of the AI? How accurate is the decision tree's explanation of the AI's categorisation? How complete is the decision tree's explanation of the AI's categorisation? How detailed is the decision tree's explanation of the AI's categorisation? These were presented to participants one question at a time.

3. Results and Discussion

3.1. Performance Measures

Figure 2 (left panel) shows learning accuracy as a proportion of correct categorisations. There was a significant main effect of phase reflecting an improvement in accuracy from the training phase (with feedback) to test phase (with no feedback) ($F(1, 53) = 37.884, p < .001, \eta^2 = .417$). There was also a significant main effect of tree depth ($F(3, 53) = 25.724, p < .001, \eta^2 = .593$), produced by a marked difference between the less accurate control condition and the three, very similar, decision tree conditions. Strikingly these values peak in test phase (with no feedback) and fall in test phase 2 (with no tree, with no feedback). The rise and fall are both significant ($F(1, 38) = 120.401, p < .001, \eta^2 = .760$; $F(1, 38) = 20.331, p < .001, \eta^2 = .349$, respectively).

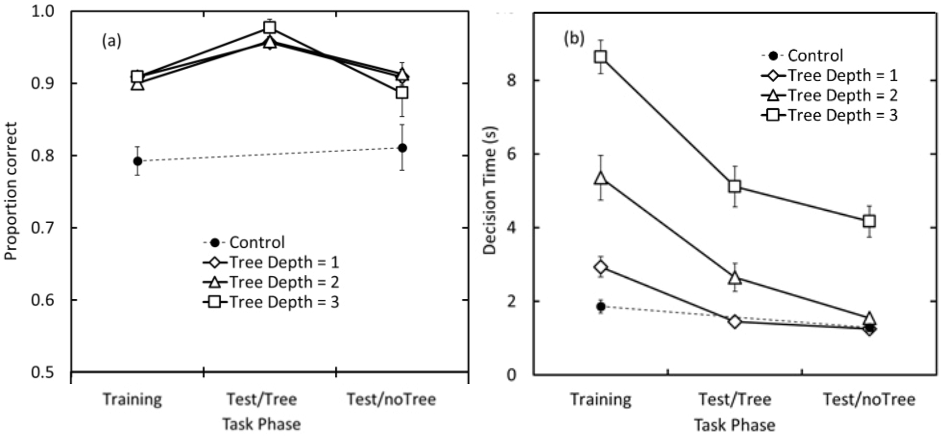


Figure 2 Mean percentage of correct responses, left panel, and mean decision time, right panel, across all conditions in Training (with feedback), Test with Tree (without feedback) and Test with no Tree (without feedback).

Figure 2 (right panel) shows the average decision time over the three phases and its interaction with decision tree depth. There was a significant main effect of phase ($F(2, 76) = 131.640, p < .001, \eta^2 = .776$), reflecting a general speeding of decision time across phases, most pronounced between Phase 1 and 2 ($F(1, 38) = 171.697, p < .001, \eta^2 = .819$). There was also a significant main effect of depth of the tree ($F(2, 38) = 37.049, p < .001, \eta^2 = .661$), with the most pronounced difference between decision tree depth three and depth two, showing decision time to be roughly three times slower when decision tree depth was three than when it was one. More interestingly, there was a significant interaction between phase and depth corresponding to slower decision times at greater levels of tree depth ($F(4, 76) = 7.274, p < .001, \eta^2 = .277$).

3.2. Judgment Measures

Figure 3 (left panel) shows judged trust in the AI in relation to tree depth and phase. Most values of the trust were above the mid-point of the scale indicating generally high trust. There was a significant main effect of tree-depth ($F(2, 41) = 4.292$, $p = .020$, $\eta^2 = .173$), characterized by a tendency for trust to be progressively higher as the tree depth increases. There was no systematic change in trust across phases ($F(2, 26) = .829$, $p = .448$, $\eta^2 = .060$). Data from the control condition show barely any change from the after-training phase to the after-test phase. What is perhaps surprising is that this level of trust is generally higher than values recorded for tree depth 1 and tree depth 2 which suggests that decision trees with low complexities can potentially reduce human trust in AI.

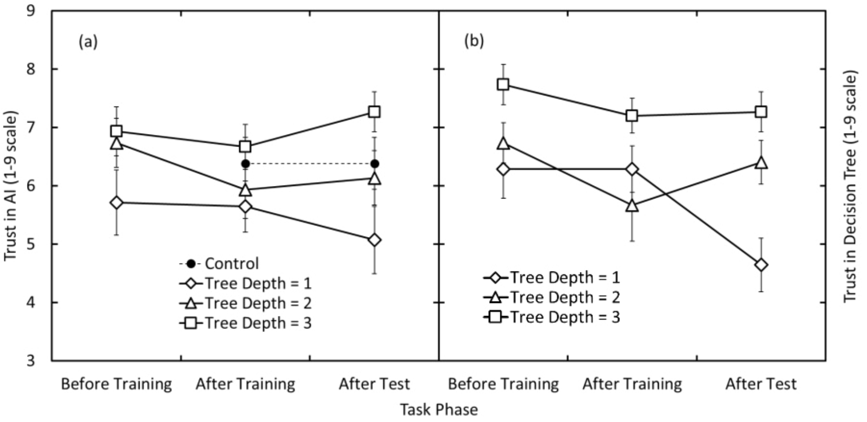


Figure 3. Trust ratings in both AI (left panel) and decision tree (right panel) across three experimental conditions at various temporal positions.

Judged trust in the Decision Tree was shown in Figure 3 (right panel). Over all conditions, most of the values are above the middle of the judgment scale, with values in the tree depth 3 condition being particularly close to the 'extremely' boundary. There is a significant main effect of tree depth ($F(2, 41) = 7.985$, $p = .001$, $\eta^2 = .280$), with tree depths 1 and 2 being broadly similar and tree depth 3 being markedly superior in engendering trust. The fact that the ratings on trust in AI and trust in the decision tree yielded similar patterns of results is indicative of the strong association of the two constructs. A correlational analysis revealed medium to large correlations between the two across all three phases ($r = .458$, $p < .01$ for before-training phase; $r = .596$, $p < .01$ for after-training phase; and $r = .744$, $p < .01$ for after-test phase), indicating that trust in the AI was related to trust in the decision tree.

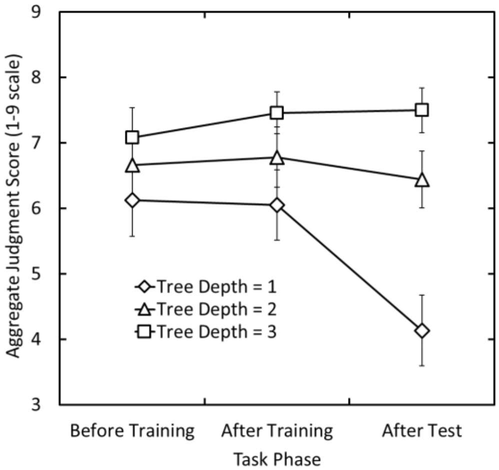


Figure 4. Mean aggregated scores on subjective measurement scales across three experimental conditions at various temporal positions

3.3. Judgment Scale Assessments

Given that the judgments for the seven questions of understandability were highly correlated with each other (Cronbach's Alpha = .937 for before training; .912 for after training; and .953 for after test phases), we aggregated the data by averaging the judgments for the different scales together, see Figure 4. The main effect of tree depth was significant with greater depth associated with higher ratings of understandability ($F(2, 41) = 12.208$, $p < .001$, $\eta^2 = .373$). There was also a significant main effect of phase (F(2, 82) = 5.078, p = .008, η2 = .110), largely brought about by an overall decline from after-training phase to after-test phase in the tree-depth 1 condition, corresponding to a significant interaction ($F(4, 82) = 4.809$, $p = .002$, $\eta^2 = .190$). Overall, this indicates that greater tree depth corresponds to better understanding and greater trust.

4. General Discussion

The impenetrability of increasingly sophisticated black-box AI has brought into relief the need for explanations of their behaviours, an example of which is in terms of explanation artefacts/proxy models. The results of the present experiment show how human learning performance, understanding and trust in a black-box AI relate to each other in the context of the complexity of an explanatory artefact, a decision tree. Specifically, the influence of the complexity of the explanatory artefact was predominantly on decision time because of a ceiling effect that arose from an emphasis on accuracy rather than speed. With the constraint that the study only included a limited range of tree depths, we can be confident with the conclusion that the greater the decision tree depth, the slower the decision time, the greater the judged understanding of the AI and the greater the trust in the AI, although the possible non-linear relation between tree depth and decision time needs further explication. Consistent with our hypotheses: 1) the results show that explanatory artefacts bring about better learning and that greater decision tree depths lead to greater interpretability of the AI's performance and greater trust in the AI; and 2) explanatory artefacts facilitate learning and task performance even after the decision aid is withdrawn.

On the face of it, increasing fidelity brings with it a number of benefits. However, it would be appropriate to be circumspect about whether these benefits continue indefinitely with increasing fidelity. It is likely that as the depth increases, the penalty associated with processing and integrating such a large amount of information may increase to the point where human performance begins to degrade [3]. Opportunities for human error increase with the number of decision points associated with the explanation. Additionally, as complexity increases, the cost of user engagement increases. This may manifest itself in terms of fatigue, resulting in a disinclination to engage with the fuller explanation and a reduced cognitive capacity to deal with the explanation. At the same time, constraints of the environment in which both the AI and humans co-operate will determine the optimum trade-off between fidelity and decision time. Clearly, environments that demand speedy response carry a risk of lower accuracy. This underscores the need for a systems'-based contextualisation of the AI black-box's operation to optimising the utility of the explanation.

One important objective of the current study was to explore how explanations of AI black-boxes promoted trust by facilitating human understanding. Many regard trust as the most important enabler to human acceptance, adoption and reliance on AI and automation [17], [18]. Engendering positive attitudes to AI—including trust—is crucial to its acceptance. As with other forms of automation, a failure to inculcate trust can lead to misuse, disuse or abuse of the system [19]. In the interpersonal domain, trust can be defined as an expectancy that others can be relied upon [20] and involves acceptance of vulnerability to another whose behaviour is not under one's control [21]–[23]. As such, trust is a social construct that involves multiple actors or different parties (a 'trustor' and a 'trustee'). The definition also points to an emotional component of trust, as the acceptance of vulnerability involves feelings that the trustor has about the trustee. In the present research context, the trustor is the user and the trustees are the AI black-boxes as well as the explanatory surrogate model for the AI. Like interactions between people, social and emotional dynamics play a role when humans interact with computers, in particular with AI black-boxes. But humans tend to behave less cooperatively when interacting with artificial agents compared to when they interact with humans [24]. However, this 'cooperation gap' can be reduced when computers incorporate human-like characteristics, i.e., their behaviour is made understandable to humans, for example, by explanatory surrogate models. [25])

Our results suggest a strong link between trust and understanding of the AI black-box and explanation artefacts. That is, the more comprehensive the explanation (i.e., the greater the decision tree depth in this case), the greater the level of reported trust and understanding, with the average correlation between trust and understanding being .628. It should also be noted that trust ratings in the AI were lower for tree-depth 1 and 2 conditions compared to the control conditions where there was no decision tree, suggesting that if the explanation is too simple, it might reduce trust, relative to no explanation at all.

5. Limitations and Future Research

Taken together, our results suggest the possibility of a non-monotonic relationship between understandability and explanation fidelity. However, our study was not able to fully explore this possibility due to the limited range of the decision-tree depths as well as the monotonic nature of our participants (i.e., non-experts). Further research is needed to explore this possibility by extending the range of evaluated fidelity values, the diversity of participants in terms of prior-knowledge of the AI, as well as the level of task difficulties. This need is further emphasised by emerging evidence that explanations are perceived as most useful when they apply to incorrect outcomes of AI activity [16]. In addition, trust seems likely to vary with the ambiguity or noisiness of the explanation, as implied by task difficulty.

There are two possible bases for the present results on trust: On the one hand, they could have been based on a cold calculus using statistical regularities and reliability of the AI system. On the other hand, they could have been the product of motivational and emotional factors. For instance, the investment of so much time in the processing of deep decision trees may have been rated, i.e., cognitive dissonance mechanisms may have been at work: ‘Why would I work so hard at deciding if this system is untrustworthy?’ Future research should explore users’ metacognitive rationale.

More generally, [25] note the aspect of ethical governance and stress the importance of being able to systematically and transparently measure and compare system capabilities, with standardised tests or benchmarks. They argue that technology can only be trusted if it is beneficial, safe, well-regulated and if investigations are robust if and when there are issues or indeed incidents. Thus, the aim of providing explanations should not only be to promote trust but perhaps more importantly, to indicate when trust might not be appropriate. Future research should more broadly assess the relationship between explanation fidelity, trust and system performance, particularly when fidelity is relatively poor.

To conclude, explanations help facilitate understanding, trust and learning. In particular, explanations consolidate learning and promote performance once provided, even after they are taken away. Explanation fidelity increases understanding, at least up to a point. In AI we should trust, but only with sufficient explanations that aren’t too complex.

Acknowledgement

This paper is dedicated to our dear friend and colleague **Professor Dylan Marc Jones OBE DSc (30th March 1948 – 8th April 2022)**, without whom this work would not have been possible. Dylan had been at Cardiff University for almost 50-years and was the father of Human Factors at the School of Psychology. Most recently, Dylan was Co-Director of IROHMS and HuFEx. This work is also a tribute to **Prof Bill Macken (16th May 1965–17th Feb 2020)**, who played an instrumental role in the conceptualisation of the project. We also thank Dr Job van der Schalk for his comments in early project meetings. The work was funded by Airbus [Airbus Operations Ltd (grant number GB762702827)]] awarded to Professor Phil Morgan and Professor Dylan M Jones.

References

[1] D. Gunning, “Explainable Artificial Intelligence (XAI): Program Update Novmeber 2017,” *Def. Adv. Res. Proj. Agency*, 2017.

[2] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, “Metrics for Explainable AI: Challenges and Prospects,” *arXiv Prepr. arXiv1812.04608*, 2018, Accessed: Apr. 02, 2020. [Online]. Available:

http://arxiv.org/abs/1812.04608.

[3] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, Oct. 2018, pp. 80–89, doi: 10.1109/DSAA.2018.00018.

[4] A. Páez, "The Pragmatic Turn in Explainable Artificial Intelligence (XAI)," *Minds Mach.*, 2019, doi: 10.1007/s11023-019-09502-w.

[5] N. Bostrom and E. Yudkowsky, "The ethics of artificial intelligence," in *The Cambridge Handbook of Artificial Intelligence*, 2014.

[6] D. Sculley *et al.*, "Hidden technical debt in machine learning systems," in *Advances in Neural Information Processing Systems*, 2015, vol. 2015-Janua, pp. 2503–2511.

[7] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision making and a 'right to explanation,'" *AI Mag.*, 2017, doi: 10.1609/aimag.v38i3.2741.

[8] M. Hall *et al.*, "A Systematic Method to Understand Requirements for Explainable AI (XAI) Systems," in *Proceedings of the IJCAI Workshop on eXplainable Artificial Intelligence (XAI 2019)*, 2019.

[9] Z. C. Lipton, "The mythos of model interpretability," *Commun. ACM*, vol. 61, no. 10, pp. 35–43, Jun. 2018, doi: 10.1145/3233231.

[10] 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 Prepr. arXiv1806.07552*, Jun. 2018, Accessed: Mar. 31, 2020. [Online]. Available: http://arxiv.org/abs/1806.07552.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, doi: 10.1145/2939672.2939778.

[12] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017.

[13] X. Renard, N. Woloszko, J. Aigrain, and M. Detyniecki, "Concept Tree: High-Level Representation of Variables for More Interpretable Surrogate Decision Trees," *arXiv Prepr. arXiv1906.01297*, 2019, Accessed: Apr. 01, 2020. [Online]. Available: http://arxiv.org/abs/1906.01297.

[14] D. Gunning, E. Vorm, J. Y. Wang, and M. Turek, "DARPA's explainable AI (XAI) program: A retrospective," *Appl. AI Lett.*, vol. 2, no. 4, p. e61, Dec. 2021, doi: https://doi.org/10.1002/ail2.61.

[15] "Load_Wine Dataset." [Online]. Available: https://scikitlearn.org/stable/modules/generated/sklearn.datasets.load_wine.html.

[16] S. C. H. Yang, W. K. Vong, Y. Yu, and P. Shafto, "A Unifying Computational Framework for Teaching and Active Learning," *Top. Cogn. Sci.*, 2019, doi: 10.1111/tops.12405.

[17] V. Riley, "Operator reliance on automation: Theory and data," in *Automation and Human Performance: Theory and Applications*, 2018.

[18] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*. 2004, doi: 10.1518/hfes.46.1.50_30392.

[19] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Hum. Factors*, 1997, doi: 10.1518/001872097778543886.

[20] J. B. Rotter, "A new scale for the measurement of interpersonal trust," *J. Pers.*, 1967, doi: 10.1111/j.1467-6494.1967.tb01454.x.

[21] D. E. Zand, "Trust and administrative problem solving," *Adm. Sci. Q.*, 1972.

[22] F. D. Schoorman, R. C. Mayer, and J. H. Davis, "Including versus excluding ability from the definition of trust," *Acad. Manag. Rev.*, 1996.

[23] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer, "Not so different after all: A cross-discipline view of trust," *Academy of Management Review*. 1998, doi: 10.5465/AMR.1998.926617.

[24] E. B. Sandoval, J. Brandstetter, M. Obaid, and C. Bartneck, "Reciprocity in Human-Robot Interaction: A Quantitative Approach Through the Prisoner's Dilemma and the Ultimatum Game," *Int. J. Soc. Robot.*, 2016, doi: 10.1007/s12369-015-0323-x.

[25] A. F. T. Winfield and M. Jirotka, "Ethical governance is essential to building trust in robotics and artificial intelligence systems," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, 2018, doi: 10.1098/rsta.2018.0085.