Exploring How Expertise Impacts Acceptability of AI Explanations: A Case Study from Manufacturing

Zibin Zhao
University of Warwick
Coventry, CV4 7AL, UK
Zibin.Zhao@warwick.ac.uk

Cagatay Turkay
University of Warwick
Coventry, CV4 7AL, UK
Cagatay.Turkay@warwick.ac.uk

CHI'23 Workshop on Human-Centered Explainable AI (HCXAI),, April 28–29, 2023
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract

There has been extensive recent research on the explainability of AI to support and improve the accountability of decision-making in various application domains. There is limited research, however, to understand in what ways explainability is required and how likely the end-users will accept these techniques to support their everyday work. These questions are particularly interesting when one considers how experienced the practitioners are in their area. In this qualitative case study, we collaborated with a manufacturing industry coordinator to design XAI-based bearing fault diagnosis simulation tasks and collect data on perceptions of explanations from practitioners with varying levels of knowledge and experience. Our preliminary findings suggest that practitioners' positive attitude towards explainable AI does not necessarily translate to their acceptance or utilisation in their work, potentially due to variations in industry experience and domain knowledge. Additionally, we reflect on the methodology of this interdisciplinary case study.

Author Keywords

XAI application; User characteristics; Manufacturing

CCS Concepts

•**Human-centered computing** → **User studies;**

Experimental Setup

Data set: We used a public bearing-fault dataset of accelerometer-measured vibration signals from Case Western Reserve University (CWRU) [22]. 1039 vibration signals with fault types located on the inner race and ball (see Figure 1) were manually sorted into two classes. Figure 2 depicts these two vibration signals.

Underlying model: We then trained a K-nearest neighbours (KNN) model using 727 instances and tested its classification accuracy using 312 instances, achieving 82.5%.

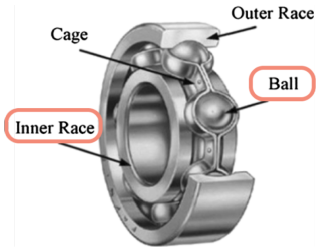

Figure 1: The bearing fault locations.

Introduction

There is a growing body of explainability methods to enhance the accountability of AI and machine learning in several domains, including high-risk ones [1, 20]. There is now consensus that a single explanation cannot suit all circumstances due to variations in goals of explainability, demands of stakeholders and requirements particular to the application context [2]. It is therefore essential to take into account the diverse explanation needs and the priorities of the stakeholders when designing and trying to deploy explainability solutions [17, 13, 4].

It has been argued that users' prior knowledge and expertise can affect their interactions with explainable AI systems [5] and their trust in the explanations provided [7]. However, prior research has mainly focused on the user's machine-learning knowledge and experience, with limited attention given to how users' domain-specific knowledge impacts their use of explainable AI (XAI) explanations [12, 21]. In particular, there is a scarcity of studies on how users' domain knowledge impacts their perceptions of XAI in real-world decision-making scenarios [3]. While recent studies in healthcare [6, 3] and manufacturing [11, 9, 8, 18] have shown promise for the application of XAI, there are still many unknowns in understanding practitioners' attitudes towards explainability, in particular when the impact of experience is considered.

To this end, this paper aims to explore how users' domain knowledge and years of industry experience affect their perceptions of XAI, using a case study from the manufacturing industry. As the study is ongoing, this paper primarily focuses on the methodology, as well as interesting preliminary results and reflections on the research process.

Research design

In this case study, we simulated the use of XAI methods in bearing fault diagnostics, a task in which an engineer tries to understand what is wrong with a *bearing*, a connector that transfers rotational movement between machine elements. In this diagnostic task, the goal of the engineer is to identify what kind of fault does the bearing have so that they can devise a solution. Typically, engineers will use the vibration signals recorded from these bearings and try to observe particular patterns to identify the kind of fault. Following this, we observed user perception based on their agreement with the XAI explanations and the model's results, as perceived usefulness is crucial in determining user acceptance of new technology [10].

Several studies have found that users without AI backgrounds prefer explanations that they can easily comprehend [14, 3]. In view of this, we utilized a post-hoc XAI method called Local Interpretable Model-Agnostic Explanations (LIME) [19] to explain the KNN model's prediction. Specifically, we used a version of LIME for time series [15], which divides an input sequence into a pre-determined number of segments and perturbs them similarly to standard LIME [16]. In the interview, participants explored explanations similar to Figure 3. For instance, if the KNN model predicted a single signal indicating a bearing fault on the ball, LIME would highlight the top 10 significant features in that signal that contributed to the prediction. As seen in the second graph of Figure 3, the red parts of the signal indicate sections that caused the model to predict an inner race fault, while the green parts represent key features that the model used to predict a ball fault.

Research implementations

This study used contextual exploration and task-focus interviews from Benk et al. [4]. The data and explanations were

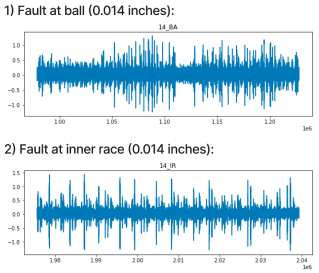

Figure 2: 1) Signal for all ball faults with a fault size of 0.014; 2) Signal for all inner race faults with a fault size of 0.014.

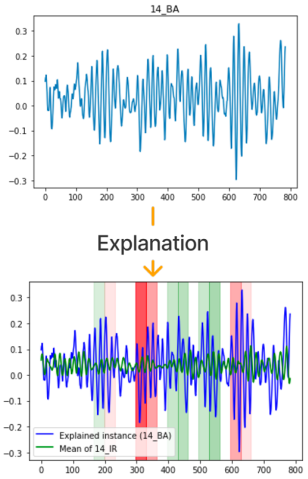

Figure 3: An explanation of a prediction from the fault classifier. Green areas indicate features positively contributing to the classification with red sections contributing negatively.

presented in print format, and participants were encouraged to work with sketching and note-taking. In the pre-research interview, we gathered specific contextual information by consulting with four bearing fault diagnosis engineers about their daily use and perceptions of a machine learning-based diagnosis system. According to their feedback, this diagnosis system utilized by engineers can only warn about bearing failures, but it cannot locate them.

The task-focused interviews consisted of four simulated tasks based on the collected contextual information. Each task included three phases. In the first phase, participants were asked to identify four bearing vibration signal plots *without* AI input. In the second phase, participants were shown the model's predictions and asked if they would alter their first-phase decisions. In the final phase, we showed participants explanations of the model's results and asked which portions they agreed with. Meanwhile, we asked a senior engineer to participate in the experiment as a facilitator to assist in clarifying any domain-specific questions.

Since this study is still ongoing, we have so far obtained data from only six bearing fault diagnosis practitioners from different Chinese manufacturing companies. While we are not yet in a position to make clear conclusions due to the small sample size, a number of interesting preliminary findings and reflections have emerged.

Preliminary findings

Users' domain knowledge and experience seem to both influence their perceptions of XAI explanations. Previous studies have shown that domain expertise affects users' trust in explanations, with users who have high domain-specific knowledge being more willing to trust the system with an explanation [10, 3]. In our study (see Figure 4), we found that users with high-level domain knowledge were generally more accepting of XAI explanations, but some users with low domain-specific knowledge changed their prior diagnosis decisions after seeing the model's results and explanations. Participants with high domain knowledge but low domain-specific experience (P3&P5) were also willing to use the ML model with explanations, but their behaviour (e.g., their diagnosis decisions and how they highlighted key features) was not affected by the explanations and model results during the interview process (see Figure 5). On the other hand, participants with low domain knowledge but rich industry experience (P2&P4) had a negative attitude towards adopting explanations, but they changed their diagnosis decisions because of the model's predictions and presented explanations in one of the tasks.

The participants' use of explanations was contrary to their attitude towards them. In other words, the participant's positive attitude towards the explainable method does not mean they will accept or implement the explanations in their workflows. The participants' various degrees of industry experience and personal attitudes towards new technologies have likely contributed to this result.

Methodological reflections

Real-world interpretations are not (always) binary. Domain experts, thanks to their background in their respective area would have distinct practices in how they perform the tasks and what they think the "right" answer would be – which may not necessarily be the same as the "ground truth" in the data. For instance, in our experiment we made use of the CWRU dataset where a signal (Figure 5) in one of the tasks led to disagreements. According to the data, the true label of this instance was an inner race fault, but the model predicted a ball fault. Four of the six participants also classfied this as a ball fault – in agreement with the "false" KNN model. Participants also reported that the sig-

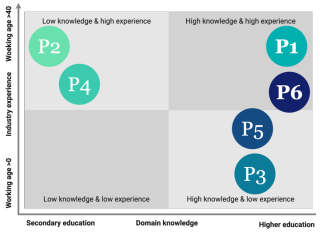

Figure 4: The distribution of our participants' domain knowledge levels and years of experience.

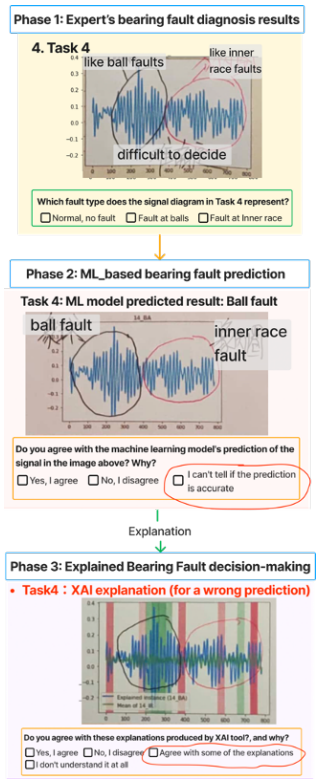

Figure 5: A record of one participant's response to one of the simulated tasks.

nal might present a crossover of the two fault types at that point in time. Since we were interested in observing how participants responded to "correct" and "wrong" predictions, such disagreements made setting the tasks challenging. As we have seen in our bearing fault classification task, multiple interpretations and classifications of data is possible and all of these interpretations could be equally correct.

Site-specific data and tasks. Relatedly, unlike "lab" settings where there are clearer, generalizable truths, real-world applications are often messy and noisy. Experts might disagree with each other and the data. This makes publicly available data harder to use in experiments with experts who might have disagreements with the experts who curated the original dataset. This adds an additional layer of bias that can not be easily accounted for prior to the study even with data sets experimentally tested. Such unknowns make the interpretation of the observations harder. A potential way to mitigate this could be to always aim to use data sets directly generated in the setting where the experiment will be run but this is harder with multi-site studies.

Mimicking the real-world is tough. Although we made efforts to create an experimental setting that mimics actual working conditions by collaborating with industry experts and gathering industry-related information, it is important to note that the data sets used in our design are not derived from real-world working environments, and the machine learning models are yet to be tested in such environments. Existing use of technologies such as AI are often limited in manufacturing and experts usually have minimal prior exposure to them. It is therefore highly important to design tasks that come from everyday practice, to gradually ease-in the participants to the study through those familiar tasks, and then introduce how the "technology" (in our case XAI) would fit in. There is a fine balance between keeping the simulated situations familiar while introducing interventions that transform them. Finding this is a challenge and requires multiple rounds of piloting and a co-design process to get the protocol as balanced as possible.

Find a mediator. We observed that involving a practitioner as a consultant and coordinator in such interdisciplinary studies is of utmost importance. During the pre-study interviews, where such a coordinator was not present, participants often mentioned that they did not know much about AI, causing unease when sharing their views and feelings about using intelligent diagnostic software. However, in the task-oriented interviews, the coordinator was the main interviewer facing the participants. In that setting, participants talked comfortably and were more willing to share their opinions and feelings. This experience made it clear that participants identified more with the industry representative, feeling more comfortable to share their views where they see the experiment as an exchange with a peer rather than being observed and judged by a researcher. Identifying an *industrial mediator* to join the study could significantly improve the effectiveness of the study and provide more in-depth insights, but this requires a careful preparation phase where the mediator is properly trained on the tasks.

Conclusion

This paper presents an ongoing case study that investigates how users' domain knowledge and experience impact the applicability of XAI in real-world applications. Our research has preliminarily shown that it is important to consider not only users' domain knowledge but also their years of experience in the field when discussing how their profession affects their needs and acceptance of AI. We also recommend the use of an intermediary person to facilitate real-world application studies and suggest being mindful of potential biases when using public data sets in field studies.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. DOI: http://dx.doi.org/10.1109/ACCESS.2018.2870052

[2] V. Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Qingzi Vera Liao, Ronny Luss, Aleksandra Mojsilovic, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John T. Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. *ArXiv* abs/1909.03012 (2019).

[3] Sarah Bayer, Henner Gimpel, and Moritz Markgraf. 2021. The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems* 0, 0 (2021), 1–29. DOI: http://dx.doi.org/10.1080/12460125.2021.1958505

[4] Michaela Benk, Raphael P. Weibel, and Andrea Ferrario. 2022. Creative Uses of AI Systems and their Explanations: A Case Study from Insurance. *ArXiv* abs/2205.00931 (2022).

[5] Angelos Chatzimparmpas, Rafael M. Martins, Ilir Jusufi, Kostiantyn Kucher, Fabrice Rossi, and Andreas Kerren. 2020. The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. *Computer Graphics Forum* (2020). DOI:http://dx.doi.org/10.1111/cgf.14034

[6] Yuhan Du, Anna Markella Antoniadi, Catherine McNestry, Fionnuala M. McAuliffe, and Catherine Mooney. 2022. The Role of XAI in Advice-Taking from a Clinical Decision Support System: A Comparative User Study of Feature Contribution-Based and Example-Based Explanations. *Applied Sciences* 12, 20 (2022). DOI: http://dx.doi.org/10.3390/app122010323

[7] Shi Feng and Jordan Boyd-Graber. 2019. What Can AI Do for Me? Evaluating Machine Learning Interpretations in Cooperative Play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 229–239. DOI: http://dx.doi.org/10.1145/3301275.3302265

[8] Anna-Christina Glock. 2021. Explaining a Random Forest With the Difference of Two ARIMA Models in an Industrial Fault Detection Scenario. *Procedia Computer Science* 180 (2021), 476–481. DOI: http://dx.doi.org/https://doi.org/10.1016/j.procs.2021.01.360 Proceedings of the 2nd International Conference on Industry 4.0 and Smart Manufacturing (ISM 2020).

[9] Claudia V. Goldman, Michael Baltaxe, Debejyo Chakraborty, and Jorge Arinez. 2021. Explaining Learning Models in Manufacturing Processes. *Procedia Computer Science* 180 (2021), 259–268. DOI:http://dx.doi.org/https://doi.org/10.1016/j.procs.2021.01.163 Proceedings of the 2nd International Conference on Industry 4.0 and Smart Manufacturing (ISM 2020).

[10] Shirley Gregor and Izak Benbasat. 1999. Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly* 23, 4 (1999), 497–530. http://www.jstor.org/stable/249487

[11] Md Junayed Hasan, Muhammad Sohaib, and Jong-Myon Kim. 2021. An Explainable AI-Based Fault Diagnosis Model for Bearings. *Sensors* 21, 12 (2021). DOI:http://dx.doi.org/10.3390/s21124070

[12] Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2019. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics* 25, 8 (2019), 2674–2693. DOI: http://dx.doi.org/10.1109/TVCG.2018.2843369

[13] Markus Langer, Daniel Oster, Timo Speith, Lena Kästner, Kevin Baum, Holger Hermanns, Eva Schmidt, and Andreas Sesing. 2021. What Do We Want From Explainable Artificial Intelligence (Xai)? ? a Stakeholder Perspective on Xai and a Conceptual Model Guiding Interdisciplinary Xai Research. *Artificial Intelligence* 296, C (2021), 103473. DOI: http://dx.doi.org/10.1016/j.artint.2021.103473

[14] Neda Mesbah, Christoph Tauchert, Christian Michael Olt, and Peter Buxmann. 2019. Promoting Trust in AI-based Expert Systems. In *Americas Conference on Information Systems*.

[15] Emanuel Metzenthin. 2021. LIME For Time. (2021). https: //github.com/emanuel-metzenthin/Lime-For-Time Accessed: 2023-02-10.

[16] Oliver Mey and Deniz Neufeld. 2022. Explainable AI Algorithms for Vibration Data-Based Fault Detection: Use Case-Adadpted Methods and Critical Evaluation. *Sensors* 22, 23 (2022). DOI: http://dx.doi.org/10.3390/s22239037

[17] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Trans. Interact. Intell. Syst.* 11, 3–4, Article 24 (sep 2021), 45 pages. DOI: http://dx.doi.org/10.1145/3387166

[18] Simon Neugebauer, Lukas Rippitsch, Florian Sobieczky, and Manuela Gei. 2021. Explainability of AI-predictions based on psychological profiling. *Procedia Computer Science* 180 (2021), 1003–1012. DOI:http://dx.doi.org/https: //doi.org/10.1016/j.procs.2021.01.361 Proceedings of the 2nd International Conference on Industry 4.0 and Smart Manufacturing (ISM 2020).

[19] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. DOI: http://dx.doi.org/10.1145/2939672.2939778

[20] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 180–186. DOI:http://dx.doi.org/10.1145/3375627.3375830

[21] Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and Their Needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 74, 16 pages. DOI: http://dx.doi.org/10.1145/3411764.3445088

[22] Case Western Reserve University. 2004. Bearing Data Center: Download a Data File. https://engineering.case.edu/bearingdatacenter/download-data-file. (2004). Accessed: 2022-11-01.