Artificial Intelligence and Social Computing, Vol. 122, 2024, 145–155
https://doi.org/10.54941/ahfe1004651

Assessing Explainability in Time Series Models: A User-Centered Approach in Industrial Settings

Emmanuel Brorsson[1], Yanqing Zhang[1], Elmira Zohrevandi[2], Nilavra Bhattachaya[3], Andreas Theodorou[4], Andreas Darnell[5], Rasmus Tammia[6], and Willem D. van Driel[7,8]

[1]ABB AB Corporate Research Center, Västerås, 72226, Sweden
[2]Linköping University, Campus Norrköping, SE-601 74 Norrköping, Sweden
[3]ABB AG Corporate Research Center, Ladenburg, 68526, Germany
[4]Universitat Politècnica de Catalunya (UPC), Spain
[5]Södra Skogsägarna, Växjö, 35251, Sweden
[6]New Boliden, Boliden, 93632, Sweden
[7]Signify, Eindhoven, 5656 AE, The Netherlands
[8]Delft University of Technology, Delft, 2600 AA, The Netherlands

ABSTRACT

This paper investigates methods for evaluating the explainability of transformer models analyzing time series data, a largely unexplored area in the field of explainable AI (XAI). The study focuses on **application-grounded methods involving human subject experiments with domain experts.** On-site evaluations were conducted in two industrial settings involving 14 control room operators. The evaluation protocols consisted of methods to **measure the metrics** of subjective comparison, forward simulatability, and subjective satisfaction. The results indicate that the chosen combination of evaluation metrics provide a multi-faceted assessment on quality and relevance of explanations from an operator's perspective in industrial settings, in turn contributing to the field of user-centered XAI evaluation, particularly in the context of time series data and offers insights for future work in this area.

Keywords: Explainable AI, Transformer models, Time series data, Explainability evaluation

INTRODUCTION

The integration of Artificial Intelligence (AI) in industrial processes is becoming increasingly common due to its ability to improve product consistency and reduce operational costs (Javaaid *et al.*, 2021). Integration of AI-based models into process industries has a key potential in supporting operators by predictive analytics (Peres, 2020) and transform operations by improving speed, flexibility and scalability (Woo, 2020).

While AI can complement human operators in industrial applications, it is most effective when designed to enhance human capabilities rather than replace them (Longo, 2017; Williamson, 2021). However, their lack of understanding can negatively affect operators' trust in these models and, hence,

impact the gain from automation (Liu *et al.*, 2023; Gade, 2019). A solution in calibrating trust can be considered the use of *transparent* and *eXplainable AI* (XAI)*;* i.e. systems whose decision-making mechanism is communicated in one way or another (Miller 2019; Theodorou *et al.*, 2017).

It is not just that the technical implementation of XAI techniques is an ongoing challenge, so is the evaluation of their effectiveness. The inclusion of the context-specific stakeholders is crucial; their domain knowledge strengthens their confidence and understanding of the AI agent (Das *et al.*, 2021; Vantrepotte *et al.*, 2021; Zhang, 2020) and facilitates balancing the benefits of explanations with associated costs (Li *et al.*, 2020a; Li *et al.*, 2020b). However, current research either completely ignores human understanding by focusing exclusively on technical metrics (Colin *et al.,* 2022; Miller *et al.,* 2017) or focus on end users, ignoring that non-expert users perceive different styles of explanations differently (Ehsan *et al.,* 2019). Other work shows that the majority of XAI research neglects to consider domain experts in their evaluations (Lopes *et al.,* 2022*;* Nauta *et al.,* 2023). Furthermore, the lack of consideration of a wide range of models, tasks, and data types—including time-series data—in existing literature regarding XAI evaluation has resulted in evaluation methods being biased towards particular models and tasks, most prominently; neural networks performing classification tasks on image data (Nauta *et al.,* 2023). The complexity of time-series data, which includes a temporal aspect that image data lacks, increases the challenge for domain experts to understand explainability (Schlegel *et al.,* 2019), resulting in some researchers advising against user evaluations of XAI for models using time-series data (Rojat *et al.,* 2021).

To fill these gaps in research, this paper aims to present a selection of user-centered XAI evaluation metrics and examples of results for time series deep learning models in process industries. Our work focuses on time-constant processes, where there is an inherent increase in complexity between the causal relationships of operators' tasks, predictions, and process data (Gade, 2019; Zhou, 2021). Our contributions include: 1) a protocol that combines metrics and supporting methods for evaluating XAI for time series models; 2) examples of results generated by the protocol in an evaluation setting with process industry operators 3) an evaluation and discussion of the use of the protocol in time-constant processes.

TIME-CONSTANT PROCESSES

Time-constant processes are characterized as how quickly the process responds to changes in the input. The time constant is a parameter that describes how a first-order linear time-invariant system reacts to a step input. It quantifies how quickly the process variable responds to changes in the input where a greater time-constant denotes a slower system response. Delignification in paper pulp production and flotation in mining are two examples of process industries where control room operators face challenges in maintaining a stable operation as the underlying process is inherently dynamic. Small adjustments to control parameters can significantly affect the system efficiency and the financial outcome. In time-constant processes, the

interconnectivity between physical parameters further adds to the complexity of the process. Visual analytics interfaces that display relevant historical data points and their relationships, as well as consequences of operator adjustments can support in finding suitable strategies (Zohrevandi *et al.*, 2023). To assess the quality of explanations designed for operators of such complex processes, evaluation activities need to be tailored to capture a coherent view of how a model and its explainers aligns with activities and mental models of users.

EXPERIMENTAL DESIGN METHODOLOGY

We have conducted an application-grounded (Doshi-Veles and Kim, 2017) evaluation study to assess the explainability of a time-series based transformer model (Lim *et al.*, 2021). To achieve this, a web-based dashboard was developed (see Figure 1) based on initial explainability requirements that were identified during previous field studies to the industrial plants currently investigated. A contextual inquiry approach was applied during evaluations by using observations and semi-structured interviews (Duda *et al.*, 2020) to support additional methods to measure the chosen metrics (presented in section Dependent Measures).

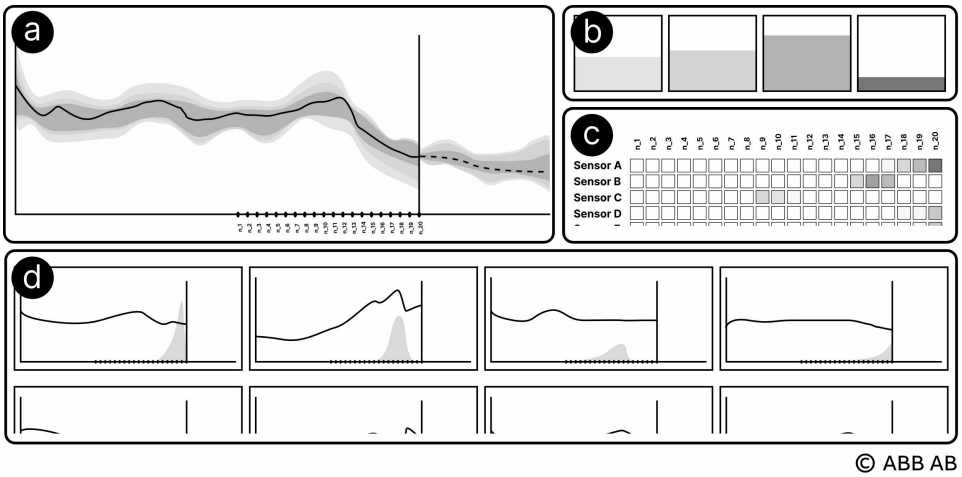

Figure 1: A schematic representation of the designed dashboard that was used in the study. The dashboard contained four views. An overview window (a) which showed the predicted time-series graph considering the sensor data and the recovery profile for the key process parameter. A precision-intervals view (b) which showed how often the key performance parameter would lie within each interval. A feature-importance view (c) which represented the sensor weight values and an on-demand representation view (d) which visualized the time-series data of all sensors for the time-period of interest selected by the operator.

Dependent Measures

To measure the quality of explanations, three metrics were chosen based on their coverage of evaluation properties proposed by Nauta *et al.*, (2023). Table 1 lists the dependent measures and specifies the categories each measure

evaluates. Three user-centric dependent measures were adopted; 1) *forward simulatability,* 2) *subjective comparison*, and, 3) *subjective satisfaction*.

1) *Forward simulatability* specifies the extent to which the user relies on explanations to predict a hidden model output (Doshi-Velez and Kim, 2017). The measure evaluates the extent to which the user has understood why the model has generated a certain output (Hase and Bansal, 2020) i.e. output-completeness. High completeness is in general desired, and being one of few methods that involves users to evaluate output-completeness (Nauta *et al.,* 2023), forward simulation becomes an important method to consider during user evaluations. This measure has been used to evaluate text and tabular data (Hase and Bansal, 2020), image data (Kim *et al.,* 2016) and time-series-based interactive simulation environments (Hoernle *et al.,* 2019). To our knowledge, our work is the first to apply forward simulatability during evaluation of time-series-based ML models in time-constant processes.

2) *Subjective comparison* tasks the user with comparing different explainers, which is a powerful metric for comparing different alternatives at early stages of design. Previous researchers have used this metric in evaluation activities in automated fact checking (Atanasova *et al.,* 2020), recommendation systems (Chen *et al.,* 2018), image recognition (Ghorbani *et al.,* 2019), and text classification (Liu *et al.,* 2018).

3) *Subjective satisfaction* is used to gather insights about satisfaction, usefulness, fluency, relevance, trust, and, sufficiency for explainers with a user in a specific context (Nauta *et al.,* 2023).

Procedure

To comply with safety regulations, the interview sessions were conducted in a room other than the control room where the operators usually work. The participant was first introduced to the project goals, the experimental procedure and the consent form. The prototype was then displayed on a computer screen. Each session included one participant and two facilitators, one leading the session while the other asking follow-up questions. The participants could interact with the interface using a computer mouse only. Audio data was recorded and participants' interactions with the interface on the screen were captured through video.

To measure subjective comparison, participants were introduced one explainer at a time, and using a think-aloud protocol (Ericsson and Simon, 1993), the participant was instructed to explain what they saw in front of them while they used the mouse to interact with parts of the interface at a time. When the whole interface was exposed, the participant was asked to compare the different explainers and their combined contribution to understanding the model forecast.

To measure forward simulatability, a previously unseen model forecast was hidden using a piece of paper while providing full access to the explainers in an attempt to predict the model output. To communicate their prediction, the participants were asked to point out their prediction on the blocking piece of paper, as well as verbally announce the predicted increase or decrease in units.

To measure subjective satisfaction, semi-structured interview questions were used together with supporting observations of user interactions, as well as a 5-point Likert scale questionnaire presented at the end of each session.

RESULTS

In this section we report the types of results generated by each metric and their supporting methods.

Participants

A total of fourteen process control operators participated in our study: eight process control operators from the mining industry and six process control operators of paper pulp production. All operators reported Swedish as their first language. Breaking down the demographics, it is worth noting that thirteen operators identified as male and just a single operator from the mining industry was female. The operators had varying levels of general knowledge of ML models, but none of them had seen the designed dashboard previously.

Table 1. The adopted dependent measures used in our evaluation study. Each measure evaluates a set of conceptual properties as initially presented by Nauta *et al.* (2023). For each metric, Nauta *et al.*'s proposed couplings between metrics and properties are compared with the couplings identified by the evaluations carried out in this work. *X* marks full coverage while *O* marks partial coverage by support from another metric.

	Correctness	Output-completeness	Consistency	Continuity	Contrastivity	Covariate Complexity	Compactness	Compositionality	Confidence	Context	Coherence	Controllability
Subjective comparison (Nauta *et al.*, 2023)							X	X		X	X	
Subjective comparison (time-constant processes)							X	O		X	X	
Forward simulatability (Nauta *et al.*, 2023)		X		X			X	X		X		
Forward simulatability (time-constant processes)		X					O	O		X	O	
Subjective satisfaction (Nauta *et al.*, 2023)						X	X	X		X	X	
Subjective satisfaction (time-constant processes)							X	X		X	X	

Subjective Comparison

The metric of subjective comparison was measured by allowing participants to interact with the prototype, comparing multiple explainers with each other while speaking out regarding what they saw in front of them. This metric was

found to provide insights mainly regarding the properties of *context*, *coherence*, and *compactness* (size of the explanation in terms of human capabilities to process it; see Table 1).

Firstly, the think aloud protocol used during subjective comparison provided insights into how the explainer aligned with previous experience, expectations and background knowledge of participants. As an example from this study, depending on their understanding of AI models, participants varied in how they described the feature importance explainer, those with less knowledge of AI expressed significant gaps in their understanding of what this explainer showed. These are examples of findings that could have been overlooked if the facilitator would initially explain the interface to the participant rather than the other way around.

Secondly, comparing multiple explainers at the same time provided insights regarding the interplay between them (*compactness*), e.g. if they complemented each other or if there was any redundancy. For example, it was found that participants used the feature importance matrix as guidance of what sensors to look at into more detail, indicating that it provided an overview similar to Shneiderman's information seeking mantra *"Overview first, zoom and filter, then details-on-demand"* (1996, p. 2). It was also found that model attention for specific sensors and time steps in the lookback period were mostly redundant as this same information was already communicated through the feature importance matrix.

Forward Simulatability

To assess the metric of forward simulatability, model output was hidden from the participant, requiring them to predict the model forecast using only explainers. This metric was found to provide insights mainly regarding the properties of *output-completeness* (the extent to which the explainer(s) cover the model output), *context* (how well the explanation aligns with user needs and experience levels) and *coherence* (alignment with mental models of operators; see Table 1).

It was found that for time-constant processes using time-series data models, the ability to assess *output-completeness* through forward simulatability might not produce as reliable results as for other data or model types because of either lack of AI knowledge or task similarity to commonly applied operator strategies. Only 2 out of 7 participants from delignification, and 2 out of 6 from flotation, all of which expressed greater understanding of the explainers during *subjective comparison*, relied on explainers such as feature importance and attention to predict the hidden model forecast, as is intended during forward simulation. The remaining operators in each process applied commonly used prediction strategies of analysing values from sensors they normally prioritize during their daily work. When the forecast is hidden, participants face a similar situation they normally encounter in their daily work where they have to predict where the process is heading to assess feasible actions. A reduced understanding of what the explainers show might have defaulted these participants to use familiar strategies.

Interestingly, these shortcomings of forward simulatability to evaluate *output-completeness* did highlight the metric's strength of producing insights for the properties of *context* and *coherence* in time-constant processes. In these situations, operators have specific strategies and information needs tied to the situation they find themselves in and explanations need to provide relevant information for the specific situations its viewers find themselves in (Miller, 2019). As such, forward simulatability can become a tool for exploring alignment with task requirements and mental models, or if explainers are well understood – *output-completeness*.

Subjective Satisfaction

In our tests, subjective satisfaction was evaluated by semi-structured interview questions throughout the evaluation sessions. The metric was found to provide great support to forward simulation and subjective comparison for gathering detailed insights.

Compared to Nauta *et al.*'s pairing of metrics and components (see Table 1), the properties of *compactness*, *context*, *coherence* and *compositionality* (how something is explained, the format & structure and organization of the explanation) and were found to not be as strongly associated in the metrics of forward simulatability and subjective comparison. However, with the added support of subjective satisfaction using semi-structured interview questions, detailed findings regarding the properties could be extracted.

For example, observing critical incidents, i.e. something notable about usability or user experience, such as miss-clicks or prolonged silence while focusing on a single object (del Galdo *et al.,* 1986) during forward simulation and subjective comparison was found to provide great support for assessing *compactness* or *context* in that participants would spend extraneous time on observing explanations that were less understandable. By asking a follow-up question regarding the relevance of the contents, participants are clearly invited to fill in the gaps. One such example from this work was in the flotation use case. Here, all 8 participant struggled with finding the right information during forward simulation. But it was not after a series of follow-up questions that it became clear that sensors that participants normally took into account were missing in the interface, in turn disrupting their ability to assess the forecast according to their mental model of the process, indicating that the *coherence* of the explainers was low. Findings like these indicates that forward simulatability and subjective comparison can create the frame of mind of where properties like *compactness*, *context*, *compositionality* and *coherence* can be evaluated, but additional metrics such as subjective satisfaction are needed to extract a fuller picture. Similar patterns can be found in other research frameworks, such as contextual inquiry which exploits the value of framing by targeted questions while the user is performing tasks in their natural environment (Duda, 2020).

Limitations

Application grounded evaluations depends on involving the intended end users carrying out their tasks with the support of the ML model (Doshi-Veles

and Kim, 2018), which can pose challenges in contexts such as process industries because of safety reasons or limited access to operators. The evaluations of this paper were carried out with the intended end users, however not in their natural environment because of safety risks of interfering with the industrial process. Instead, a separate system in a separate room was used, using methods such as forward simulation to trigger normal task behavior.

Participants were exposed to all explainers at once during subjective comparison. While allowing access to all explainers at once provides the possibility to generate insights regarding the holistic use, some fidelity might be lost regarding which single explainer might be the most useful for the participant.

CONCLUSION

This work has presented evaluation metrics and supporting methods for user-centred XAI evaluation of time series data models in time-constant processes, which is a largely unexplored space in existing research. The selected metrics of forward simulation, subjective comparison, and subjective satisfaction, were found to evaluate XAI properties in somewhat different ways in time-constant processes using time-series data compared to previous research for other ML tasks and data types. Forward simulation was found to indicate if explainers are output complete for time series models for operators with a higher degree of ML understanding. The metric was also found to support assessment of how well explainers support existing prediction strategies. If forward simulation is to be used specifically for evaluating *output-completeness* for time series models, we recommend balancing the hidden forecast according to correctness, as suggested by Hase and Bansal (2020), in turn reducing the risk that the participant would coincidentally predict the model output when predicting the process or vice versa. Subjective comparison was found to mainly produce insights regarding interplay and redundancies between multiple explainers as well as indicate misunderstandings of the contents of individual explainers. Subjective satisfaction was found to provide crucial support in generating additional details for several evaluation properties, highlighting the benefits of using multiple metrics and methods to gather insights from several angles such as observed behaviour, spoken thoughts, and answered questions. Overall, this work contributes to the understanding of user-centered XAI evaluation, particularly in the context of time series data in time-constant processes and offers insights for future work in this area.

ACKNOWLEDGMENT

The present study is funded by VINNOVA Sweden (2021-04336), Bundesministerium für Bildung und Forschung (BMBF; 01IS22030), and Rijksdienst voor Ondernemend Nederland (AI2212001) under the project *Explanatory Artificial Interactive Intelligence for Industry*[1] (EXPLAIN).

[1] https://explain-project.eu/

REFERENCES

Atanasova, P., Simonsen, J. G., Lioma, C. and Augenstein, I. (2020). Generating fact checking explanations. arXiv preprint arXiv:2004.05773.

Chen, C., Zhang, M., Liu, Y. and Ma, S. (2018). "Neural attentional rating regression with review-level explanations". Proceedings of the 2018 world wide web conference (pp. 1583–1592).

Colin, J., Fel, T., Cadène, R., Serre, T. (2022). What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods. Advances in Neural Information Processing Systems 35, 2832–2845.

Das, D., Banerjee, S., Chernova, S. (2021). "Explainable AI for Robot Failures: Generating Explanations that Improve User Assistance in Fault Recovery". Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. Presented at the HRI '21: ACM/IEEE International Conference on Human-Robot Interaction, ACM, Boulder CO USA, pp. 351–360. https://doi.org/10.1145/3434073.3444657

del Galdo, E. M., Williges, R. C., Williges, B. H. and Wixon, D. R. (1986). "An evaluation of critical incidents for software documentation design". Proceedings of the Human Factors Society Annual Meeting (Vol. 30, No. 1, pp. 19–23). Sage CA: Los Angeles, CA: SAGE Publications.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Duda, S., Warburton, C. and Black, N. (2020). Contextual research: Why we need to research in context to deliver great products. In Human-Computer Interaction. Design and User Experience: Thematic Area, HCI 2020, Held as Part of the 22nd International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22 (pp. 33–49). Springer International Publishing.

Ehsan, U., Riedl, M. O. (2019). On design and evaluation of human-centered explainable AI systems. Glasgow'19.

Ehsan, U., Wintersberger, P., Liao, Q. V., Mara, M., Streit, M., Wachter, S., Riener, A. and Riedl, M. O. (2021). Operationalizing human-centered perspectives in explainable AI. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1–6).

Ericsson, K. A., & Simon, H. A. (1993). Protocol analysis: Verbal report as data. Cambridge, Massachusetts, London, England: The MIT Press.

Gade, K., Geyik, S. C., Kenthapadi, K., Mithal, V., Taly, A. (2019). "Explainable AI in Industry". Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Presented at the KDD '19: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM, Anchorage AK USA, pp. 3203–3204. https://doi.org/10.1145/3292500.3332281

Ghorbani, A., Wexler, J., Zou, J. Y. and Kim, B. (2019). Towards automatic concept-based explanations. Advances in neural information processing systems, 32.

Hase, P. and Bansal, M. (2020). Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? arXiv preprint arXiv:2005.01831.

Hoernle, N., Gal, K., Grosz, B., Lyons, L., Ren, A. and Rubin, A. (2019). Interpretable models for understanding immersive simulations. arXiv preprint arXiv:1909ee.11025.

Kim, B., Khanna, R. and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. Advances in neural information processing systems, 29.

Li, N., Adepu, S., Kang, E., Garlan, D. (2020a). "Explanations for human-on-the-loop: a probabilistic model checking approach". Proceedings of the IEEE/ACM

15th International Symposium on Software Engineering for Adaptive and Self-Managing Systems. Presented at the SEAMS '20: IEEE/ACM 15th International Symposium on Software Engineering for Adaptive and Self-Managing Systems, ACM, Seoul Republic of Korea, pp. 181–187. https://doi.org/10.1145/3387939.3391592

Li, N., Camara, J., Garlan, D., Schmerl, B. (2020b). Reasoning about When to Provide Explanation for Human-involved Self-Adaptive Systems, in: 2020 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS). Presented at the 2020 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS), IEEE, Washington, DC, USA, pp. 195–204. https://doi.org/10.1109/ACSOS49614.2020.00042

Liu, D., Wang, Y., Liu, C., Yuan, X., Yang, C., Gui, W. (2023). Data Mode Related Interpretable Transformer Network for Predictive Modeling and Key Sample Analysis in Industrial Processes. IEEE Trans. Ind. Inf. 19, 9325–9336. https://doi.org/10.1109/TII.2022.3227731

Liu, H., Yin, Q. and Wang, W. Y. (2018). Towards explainable NLP: A generative explanation framework for text classification. arXiv preprint arXiv:1811.00196.

Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748–1764.

Lopes, P., Silva, E., Braga, C., Oliveira, T., Rosado, L. (2022). XAI Systems Evaluation: A Review of Human and Computer-Centred Methods. Applied Sciences 12, 9423. https://doi.org/10.3390/app12199423

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence, Volume 267, 1–38.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M. and Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. ACM Computing Surveys, 55, pp. 1–42.

Peres, R. S., Jia, X., Lee, J., Sun, K., Colombo, A. W., Barata, J. (2020). Industrial Artificial Intelligence in Industry 4.0 - Systematic Review, Challenges and Outlook. IEEE Access 8, 220121–220139. https://doi.org/10.1109/ACCESS.2020.3042874

Rojat, T., Puget, R., Filliat, D., Del Ser, J., Gelin, R. and Díaz-Rodríguez, N. (2021). Explainable artificial intelligence (xai) on timeseries data: A survey. arXiv preprint arXiv:2104.00950.

Shneiderman, B., (1996). "The eyes have it: A task by data type taxonomy for information visualizations". In Proceedings 1996 IEEE symposium on visual languages (pp. 336–343). IEEE.

Schlegel, U., Arnout, H., El-Assady, M., Oelke, D. and Keim, D. A. (2019), October. Towards a rigorous evaluation of XAI methods on time series. In 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) (pp. 4197–4201). IEEE.

Theodorou, A., Wortham, R. H. and Bryson, J. J. (2017). Designing and implementing transparency for real time inspection of autonomous robots. Connection Science, 29(3), pp. 230–241.

Vantrepotte, Q., Berberian, B., Pagliari, M., Chambon, V., (2021). Leveraging human agency to improve confidence and acceptability in human-machine interactions (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/6pvnh

Williamson, S., Vijayakumar, K., (2021). Artificial intelligence techniques for industrial automation and smart systems. Concurrent Engineering 29, 291–292. https://doi.org/10.1177/1063293X211026275

Woo, W. L., (2020). Future Trends in I&M: Human-machine Co-creation in the Rise of AI. IEEE Instrum. Meas. Mag. 23, 71–73. https://doi.org/10.1109/MIM.2020.9062691

Zhang, Y., Liao, Q. V., Bellamy, R. K. E., (2020). "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making". Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Presented at the FAT*'20: Conference on Fairness, Accountability, and Transparency, ACM, Barcelona Spain, pp. 295–305. https://doi.org/10.1145/3351095.3372852

Zhou, J., Gandomi, A. H., Chen, F., Holzinger, A., (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. Electronics 10, 593. https://doi.org/10.3390/electronics10050593

Zohrevandi, E., Brorsson, E., Darnell, A., Bång, M., Lundberg, J., Ynnerman, A., (2023). "Design of an Ecological Visual Analytics Interface for Operators of Time-Constant Processes", in: 2023 IEEE Visualization and Visual Analytics (VIS). Presented at the 2023 IEEE Visualization and Visual Analytics (VIS), IEEE, Melbourne, Australia, pp. 131–135. https://doi.org/10.1109/VIS54172.2023.00035