Why Industry 5.0 Needs XAI 2.0?*

Szymon Bobek[1],*, Sławomir Nowaczyk[2],
Joao Gama[3], Sepideh Pashami[2], Rita P. Ribeiro[3], Zahra Taghiyarrenani[2],
Bruno Veloso[3], Lala Rajaoarisoa[5], Maciej Szelążek[4] and Grzegorz J. Nalepa[1]

[1] *Jagiellonian Human-Centered Artificial Intelligence Laboratory (JAHCAI), Mark Kac Center for Complex Systems Research, and Institute of Applied Computer Science, Jagiellonian University, 31-007 Kraków, Poland*

[2] *Center for Applied Intelligent Systems Research, Halmstad University, Sweden*

[3] *University of Porto, Porto, Portugal and INESC TEC, Porto, Portugal*

[4] *AGH UST, Kraków, Poland*

[5] *IMT Nord Europe, Univ. of Lille, Center for digital systems, F-59000 Lille, France*

Abstract

Advances in artificial intelligence trigger transformations that make more and more companies enter Industry 4.0 and 5.0 eras. In many cases, these transformations are gradual and performed in a bottom-up manner. This means that in the first step, the industrial hardware is upgraded to collect as much data as possible without actual planning of the utilization of the information. Furthermore, the data storage and processing infrastructure is prepared to keep large volumes of historical data accessible for further analysis. Only in the last step are methods for processing the data developed to improve or gain more insight into the industrial and business processes. Such a pipeline makes many companies face a problem with huge amounts of data, an incomplete understanding of how the existing knowledge is represented in the data, under which conditions the knowledge no longer holds, or what new phenomena are hidden inside the data. We argue that this gap needs to be addressed by the next generation of XAI methods which should be expert-oriented and focused on knowledge generation tasks rather than model debugging. The paper is based on the findings of the EU CHIST-ERA project on Explainable Predictive Maintenance (XPM).

Keywords

Explainability, Predictive Maintenance, Industry 5.0

1. Industrial data is a black box itself

Explainable artificial intelligence (XAI) aims to bring transparency, trust, and intelligibility to automated decision-making systems based on AI algorithms. In recent years, the primary focus of XAI was to build methods that introduce transparency into black-box models, such as deep

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

*Corresponding author.

szymon.bobek@uj.edu.pl (S. Bobek); slawomir.nowaczyk@hh.se (S. Nowaczyk); jgama@fep.up.pt (J. Gama); sepideh.pashami@hh.se (S. Pashami); rpribeiro@fc.up.pt (R. P. Ribeiro); zahra.taghiyarrenani@hh.se (Z. Taghiyarrenani); bveloso@fep.up.pt (B. Veloso); lala.rajaoarisoa@imt-nord-europe.fr (L. Rajaoarisoa); szelazek.m@gmail.com (M. Szelążek); grzegorz.j.nalepa@uj.edu.pl (G. J. Nalepa)

0000-0002-6350-8405 (S. Bobek); 0000-0002-7796-5201 (S. Nowaczyk); 0000-0003-3357-1195 (J. Gama); 0000-0003-3272-4145 (S. Pashami); 0000-0002-6852-8077 (R. P. Ribeiro); 0000-0002-1759-8593 (Z. Taghiyarrenani); 0000-0001-7980-0972 (B. Veloso); 0000-0001-9624-5843 (L. Rajaoarisoa); 0000-0002-1857-602X (M. Szelążek); 0000-0002-8182-4225 (G. J. Nalepa)

CEUR Workshop Proceedings (CEUR-WS.org)

neural networks. This includes a variety of model-agnostic methods such as LIME [1], SHAP [2], Anchor [3], LORE [4], and LUX [5], but also model-specific approaches that take advantage of gradient-based deviations from the reconstruction of the input features [6] or the internal features of particular machine learning algorithms, such as GradCam [7], dedicated to deep neural networks. In parallel to the aforementioned research, another trend is also emerging that forces the construction of efficient glass-box [8], inherently interpretable models such as explainable boosting machines [9], and, most recently, prototype deep neural networks.

Regardless of the underlying techniques used in the XAI methods, their goal is always to explain the model's decisions by delivering a description of a relationship between the model input and the model output. Such a description can be presented in various forms [10]: feature importance, feature impact, decision rules, prototypes, examples, etc. This builds end-user trust in the model and allows for better model understanding, which can be used to model debugging, data debiasing, etc., contributing to boosting the model performance.

However, at the same time, one should assume that high-performing models encapsulate important knowledge, which can be used to better understand not only the model itself but also the data and the process that generates it. This is especially important in areas such as industry, where data is a black box itself – as in most cases it comes as unlabeled, noisy, incomplete, and, most importantly, without any explicitly formulated background knowledge about the processes that generated them. It is even more essential for Industry 5.0, where the collaboration between AI and humans is one of the core assumptions, and the pitfalls and false hopes for current XAI have already been reported [11, 12]. Therefore, shedding more light on the data and focusing on knowledge generation regarding the underlying mechanism that is the source of the data can help improve not only the model but also the whole process (i.e., business process, manufacturing process, maintenance process) on which the next generation of models can be trained, as presented in Fig. 1. Surprisingly, this motivation was never a foundation for the state-of-the-art XAI algorithms. Therefore, in our opinion, the new XAI methods for Industry 5.0 require several major changes in the design assumptions, including the following:

A1 – Presenting knowledge that has already been available in the domain is closer to *confirmation*, rather than explanation. The new XAI methods should focus more on extracting knowledge from predictive models than be limited only to the explanations that mainly serve the purpose of a model or dataset debugging.

A2 – Explanation is an act of *knowledge transfer*. This imposes a need for research on methods that will allow bidirectional transfer between humans and AI systems. The new generation of XAI algorithms should inherently be designed to allow humans to formulate expectations and needs to be addressed by the XAI algorithm and receive explanations on the desired level of abstraction.

A3 – New knowledge discovered in the explanation process may not be consistent with domain knowledge. The new generation of XAI algorithms should generate explanations that can be validated against domain knowledge and contested by a human [13]. Therefore, they should be tightly coupled with argumentative frameworks [14].

A4 – Explanations are useful if they are actionable [15]. New XAI algorithms should be designed so that the knowledge they extract can be directly utilized by machine learning models, the business process, or instant decision-making, closing the explanation loop.

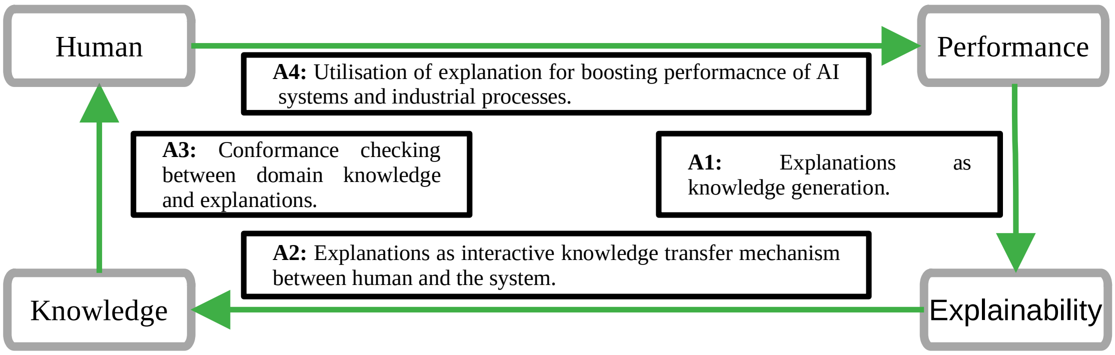
Human
Performance
Explainability
Knowledge
A4: Utilisation of explanation for boosting performacnce of AI systems and industrial processes.
A3: Conformance checking between domain knowledge and explanations.
A1: Explanations as knowledge generation.
A2: Explanations as interactive knowledge transfer mechanism between human and the system.

Figure 1: XAI2.0 should explicitly be focused on extracting knowledge from the model, and generating explanations in a more holistic manner.

2. Showcases and limitations of current XAI in industry

The assumptions brought by us in the previous section aroused based on our experience from eXpainable Predictive Maintenance (XPM)[1] project where we work on integrating explanations into Artificial Intelligence (AI) solutions within the area of Predictive Maintenance (PM) [16]. In the following section, we present several showcases of major challenges for practical XAI applications in four selected cases: electric vehicles, metro trains, steel plants and wind farms.

In the steel industry, one of the questions is how to translate the decision–making process performed by an ML algorithm in a way that can be understood by a domain expert who is not necessarily an ML specialist. This is an essential aspect, neglected in many of the current XAI solutions, as successful implementation of new solutions requires compliance with the technologies and procedures used in the company **(A4)**. Therefore, it is crucial to develop XAI methods that will allow incorporating real-life knowledge into quality decision support tasks and creating a semantic connection between the data and human specialists [17] **(A2)**.

Another issue related to the proper utilization of domain knowledge in explanations is related to so-called physics-guided or physics-informed systems, popular in industries that relay on processes that have strong theoretical background in mathematics or physics. In such a case, the AI system can be enhanced with that knowledge, as shown in [18]. However, current XAI methods do not allow us to take full advantage of the fact that domain knowledge has been used in the ML model, although it could be used as part of the quality control of explanations in terms of compliance with known theoretical foundations **(A3)**.

Addressing **(A1)**, **(A2)**, and **(A3)** is also important in the automotive industry. Transfer Learning and, more specifically, Domain Adaptation, have been shown to effectively address dynamic and diverse environments [19]. However, explanations of how domains differ and how AI/ML models capture and align these variations are still lacking. An industry-specific Explainable Domain Adaptation (XDA) approach is required to extract information on domain changes that may not be immediately apparent to domain experts [20] **(A1)**, **(A3)**. Only by leveraging the explanations and knowledge acquired through XDA, can experts make informed decisions and develop business strategies (for example, maintenance plans or usage guidelines) that align with the requirements and expectations of diverse customers and fleet operators **(A1)**.

[1] See https://www.chistera.eu/projects/xpm.

Analogous problems were observed by us in explaining failures and anomalies. In estimating time until a system failure, the explanation should correspond to describing survival or hazard functions. Survival analysis is a widely used method in various industries for that purpose, however, it frequently falls short in providing explanations for its estimations. Therefore, it is crucial to develop explanations that express the influential and distinctive factors that have shaped the observed survival patterns [21] **(A1)**, **(A3)**.

In the case of online anomaly detection algorithms [22, 23], some rule-based methods [24, 25] allow domain experts to identify an abnormal behavior of a specific sensor or module **(A1)**. Nonetheless, despite the capability to report anomalies in real-time, these rule-based models do not identify the root cause of the failure **(A3)** or allow the transfer of domain knowledge to the model **(A2)**. In the specific case of the mobility industry, imprecise explanations can lead to a wrong decision of maintenance teams, consequently causing the faulty component to not be repaired. The explanation layer must be designed to have a human in the loop to improve knowledge transfer between the domain expert and the model.

In the wind farm industries maintenance driven by an anomaly explanation approach should assist maintenance planners in making accurate diagnostic decisions [6]. The question is how to generate explanations and provide specific inferences about the problem (its severity and complexity), that can be validated against the domain's knowledge **(A3)**. In such a setting, the explanation provided by the most popular XAI algorithms, such as SHAP, LIME, Anchor, etc. is finally not sufficient. The difficulty is, on the one hand, as the Shapley value for each characteristic is obtained by the contribution of the features to all possible subsets of other features so that for N characteristics, the calculation of SHAP values is exponential in the number of features [26]. Thus, the application is complex for the decision-support purposes. On the other hand, LIME method is very sensitive to small perturbations in the input data that can cause a large change in the values of the features [27]. That implies that in some situations, it is possible to generate different explanations for very similar observations, which may confuse the target user (operator or maintenance manager). Accordingly, to make efficient the decision-making process, we should provide detailed and persistent explanations based on knowledge of the domain expertise, as well as the exploitation of data and inspection reports [28]. Indeed, the integration of this information should improve significantly the monitoring and control of industrial processes **(A4)**.

3. Summary

To make Explainable Artificial Intelligence (XAI) more useful, several changes can be implemented, such as moving toward context-aware, human-centric, understandable and trustworthy explanations. We argue that this cannot be delivered with current XAI methods unless the new generation of XAI systems includes the assumptions A1-A4 as design requirements. For the development of AI systems that means a need for redesigning the ML/DM pipeline to make an XAI an integral part of it and allocating more resources to research efficient ways of human-AI interaction.

Acknowledgment

The paper is funded from the XPM project funded by the National Science Centre, Poland under CHIST-ERA programme (NCN UMO-2020/02/Y/ST6/00070) and Swedish Research Council under grant CHIST-ERA-19-XAI-012 and by the EC project HumanE-AI: Toward AI Systems That Augment and Empower Humans(grant #820437)

References

[1] M. T. Ribeiro, S. Singh, C. Guestrin, "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, 2016, p. 1135–1144. URL: https://doi.org/10.1145/2939672.2939778. doi:10.1145/2939672.2939778.

[2] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable ai for trees, Nature Machine Intelligence 2 (2020) 56–67. URL: https://doi.org/10.1038/s42256-019-0138-9. doi:10.1038/s42256-019-0138-9.

[3] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: high-precision model-agnostic explanations, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18, AAAI Press, New Orleans, Louisiana, USA, 2018, pp. 1527–1535.

[4] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, F. Giannotti, Local Rule-Based Explanations of Black Box Decision Systems, 2018. URL: http://arxiv.org/abs/1805.10820. doi:10.48550/arXiv.1805.10820, arXiv:1805.10820 [cs].

[5] S. Bobek, G. J. Nalepa, Introducing uncertainty into explainable ai methods, in: M. Paszynski, D. Kranzlmüller, V. V. Krzhizhanovskaya, J. J. Dongarra, P. M. A. Sloot (Eds.), Computational Science – ICCS 2021, Springer International Publishing, Cham, 2021, pp. 444–457.

[6] J. Randriarison, L. Rajaoarisoa, M. Sayed-Mouchaweh, Faults explanation based on a machine learning model for predictive maintenance purposes, in: Proceedings of the 7th edition in the series of the International Conference on Control, Automation and Diagnosis,, 2023, p. 1.

[7] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization, CoRR abs/1610.02391 (2016). URL: http://arxiv.org/abs/1610.02391. arXiv:1610.02391.

[8] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215. URL: https://doi.org/10.1038/s42256-019-0048-x. doi:10.1038/s42256-019-0048-x.

[9] Y. Lou, R. Caruana, J. Gehrke, G. Hooker, Accurate intelligible models with pairwise interactions, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 623–631. URL: https://doi.org/10.1145/2487575.2487579. doi:10.1145/2487575.2487579.

[10] M. Kuk, S. Bobek, B. Veloso, L. Rajaoarisoa, G. J. Nalepa, Feature importances as a tool for root cause analysis in time-series events, in: Proceedings of the International Conference on Computational Science (ICCS), 2023, p. 1.

[11] A. Adadi, M. Berrada, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), IEEE Access 6 (2018) 52138–52160. doi:10.1109/ACCESS.2018.2870052, conference Name: IEEE Access.

[12] S. Verma, A. Lahiri, J. P. Dickerson, S.-I. Lee, Pitfalls of Explainable ML: An Industry Perspective, 2021. URL: http://arxiv.org/abs/2106.07758. doi:10.48550/arXiv.2106.07758, arXiv:2106.07758 [cs].

[13] G. Vilone, L. Longo, A novel human-centred evaluation approach and an argument-based method for explainable artificial intelligence, in: I. Maglogiannis, L. Iliadis, J. Macintyre, P. Cortez (Eds.), Artificial Intelligence Applications and Innovations, Springer International Publishing, Cham, 2022, pp. 447–460.

[14] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and explainable artificial intelligence: a survey, The Knowledge Engineering Review 36 (2021) e5. doi:10.1017/S0269888921000011.

[15] I. Linkov, S. Galaitsi, B. D. Trump, J. M. Keisler, A. Kott, Cybertrust: From explainable to actionable and interpretable artificial intelligence, Computer 53 (2020) 91–96. doi:10.1109/MC.2020.2993623.

[16] S. Pashami, S. Nowaczyk, Y. Fan, J. Jakubowski, N. Paiva, N. Davari, S. Bobek, S. Jamshidi, H. Sarmadi, A. Alabdallah, R. P. Ribeiro, B. Veloso, M. Sayed-Mouchaweh, L. Rajaoarisoa, G. J. Nalepa, J. Gama, Explainable predictive maintenance, 2023. arXiv:2306.05120.

[17] M. Szelażek, S. Bobek, G. J. Nalepa, Semantic data mining-based decision support for quality assessment in steel industry, Expert Systems n/a (2023) e13319. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.13319. doi:https://doi.org/10.1111/exsy.13319. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/exsy.13319.

[18] J. Jakubowski, P. Stanisz, S. Bobek, G. J. Nalepa, Roll wear prediction in strip cold rolling with physics-informed autoencoder and counterfactual explanations, in: 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), 2022, pp. 1–10. doi:10.1109/DSAA54385.2022.10032357.

[19] Z. Taghiyarrenani, S. Nowaczyk, S. Pashami, M.-R. Bouguelia, Multi-domain adaptation for regression under conditional distribution shift, Expert Systems with Applications 224 (2023) 119907.

[20] A. Berenji, S. Nowaczyk, Z. Taghiyarrenani, Data-centric perspective on explainability versus performance trade-off, in: Advances in Intelligent Data Analysis XXI, Springer Nature Switzerland, Cham, 2023, pp. 42–54.

[21] A. Alabdallah, S. Pashami, T. Rögnvaldsson, M. Ohlsson, Survshap: A proxy-based algorithm for explaining survival models with shap, in: 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), 2022, pp. 1–10. doi:10.1109/DSAA54385.2022.10032392.

[22] N. Davari, S. Pashami, B. Veloso, S. Nowaczyk, Y. Fan, P. M. Pereira, R. P. Ribeiro, J. Gama, A fault detection framework based on lstm autoencoder: A case study for volvo bus data set, in: Advances in Intelligent Data Analysis XX: 20th International Symposium on Intelligent Data Analysis, IDA 2022, Rennes, France, April 20–22, 2022, Proceedings, Springer, 2022, pp. 39–52.

[23] N. Davari, B. Veloso, R. P. Ribeiro, J. Gama, Fault forecasting using data-driven modeling: A case study for metro do porto data set, in: Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part II, Springer, 2023, pp. 400–409.

[24] R. P. Ribeiro, S. M. Mastelini, N. Davari, E. Aminian, B. Veloso, J. Gama, Online anomaly explanation: A case study on predictive maintenance, in: Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part II, Springer, 2023, pp. 383–399.

[25] G. Vilone, L. Longo, A quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods, Frontiers in Artificial Intelligence 4 (2021).

[26] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4765–4774.

[27] D. Alvarez-Melis, J. T. S., On the robustness of interpretability methods, CoRR abs/1806.08049 (2018). URL: http://arxiv.org/abs/1806.08049. arXiv:1806.08049.

[28] M. Sayed-Mouchaweh, L. Rajaoarisoa, Explainable decision support tool for iot predictive maintenance within the context of industry 4.0, in: 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), 2022, pp. 1492–1497. doi:10.1109/ICMLA55696.2022.00234.