Towards a model- and data-focused taxonomy of XAI systems

Jan-Peter Kucklick
Paderborn University (UPB), Germany, jan.kucklick@uni-paderborn.de

Towards a model- and data-focused taxonomy of XAI systems

Jan-Peter Kucklick[1]

[1] Paderborn University (UPB), Business Information Systems, Paderborn, Germany
jan.kucklick@upb.de

Abstract. Explainable Artificial Intelligence (XAI) is currently an important topic for the application of Machine Learning (ML) in high-stakes decision scenarios. Related research focuses on evaluating ML algorithms in terms of interpretability. However, providing a human understandable explanation of an intelligent system does not only relate to the used ML algorithm. The data and features used also have a considerable impact on interpretability. In this paper, we develop a taxonomy for describing XAI systems based on aspects about the algorithm and data. The proposed taxonomy gives researchers and practitioners opportunities to describe and evaluate current XAI systems with respect to interpretability and guides the future development of this class of systems.

Keywords: Explainable Artificial Intelligence, XAI, Interpretability, Decision Support Systems, Taxonomy

1 Introduction

Explainable Artificial Intelligence (XAI), the ability to explain an algorithm's decision-making in a human-understandable way, is one important topic in the information systems and computer science disciplines [1–4]. Currently, many modern applications are based on Machine Learning (ML) or Artificial Intelligence (AI), where an algorithm is trained for making predictions. With the development of more and more powerful algorithms, the performance of these systems rose. However, their complexity also increased over time. Consequently, many modern ML models are so-called black boxes, which leave their mechanics of reasoning for decisions disguised. According to Adadi and Berrada [2], there are four main motivations to use XAI: Justification of decisions, supervision and governance of the algorithm's decisions, debugging capabilities to improve the system, and knowledge generation. The first three aspects are especially of interest in high-stakes decisions, where false and unjustified decisions have severe consequences such as in medicine, finance, and legal practice [1, 4].

Currently, research on XAI is focusing on technical aspects like finding new interpretable mathematical algorithms, (post-hoc) methods that help to explain the black-box systems, and evaluation methods of the quality of XAI systems [1, 5–7] or the user acceptance and effectiveness of proposed XAI methods [8–12].

17th International Conference on Wirtschaftsinformatik,
February 2022, Nürnberg, Germany

Hence, most research is mainly focusing on mathematical aspects of ML models [8, 9, 13, 14], an aspect mostly related to the algorithmic interpretability [15]. Nevertheless, a decision support system can be deconstructed into the combination of an algorithm and data [16].

Based on this separation, we argue that the interpretability of the whole system is also dependent on the used features, their representation, and characteristics - a perspective that has not been in focus in related work. One domain that relies on sophisticated methods and interpretability is real estate appraisal which implemented deep learning methods to improve the predictive performance of algorithms. However, real estate agents rely on interpretability of the features to justify their decisions [17]. In this paper, we propose a taxonomy to characterize intelligent systems in a technical way, including algorithms and features in relation to interpretability. Data Scientists, IT managers, and researchers are enabled by this taxonomy to set up their XAI strategy, to rate their current intelligent systems and to guide the development process according to their interpretability needs. As also the data perspective is included in the taxonomy, it provides an enhanced view on how to improve the XAI system.

2 Research Methodology

As the development of a taxonomy is a complex task, we follow the conceptual-to-empirical approach, including a conceptualization and an evaluation phase [18]. We start with a literature review on XAI based on the following search terms: ("explainability" or "interpretability" or "transparency") and ("artificial intelligence" or "machine learning" or "deep learning" or "decision support systems") in the AIS eLibrary, Google Scholar and ACM Digital Library and EBSCO Host [19]. The related work is analyzed according to quality, as only peer reviewed papers are considered, and their fit in terms of content. The taxonomy is conceptualized on the consolidated theoretical foundation gathered by the literature review. In an evaluation step, the taxonomy is refined by incorporating knowledge from applying the taxonomy. This short paper proposes a taxonomy, which will be evaluated and extended to a holistic framework in a full version paper.

3 Conceptualization of XAI and Discussion

As a starting point, related work about model interpretability is used. *Inherently interpretable algorithms* are for example a Linear Regression or a Decision Tree [3, 13, 14]. *Classical machine learning models* like a Random Forest or Extreme Gradient Boosting build upon the concept of Decision Trees, but their accuracy might be higher by reducing variance. However, the complexity of hundreds of trees reduces their interpretability level [14]. In addition to classical ML algorithms, *deep learning* might add further accuracy, especially when using text or image data [8, 14, 20]. Nevertheless, deep learning also provides the least amount in interpretability from an algorithmic view due to their chained high-dimensional non-linear functions [13]. Consequently, this results in reduced algorithmic transparency, and it often remains unclear why and

under which conditions these highly complex algorithms result in an optimal solution [15]. Therefore, we identified three different levels of interpretability, *high, medium, low,* according to the interpretability of the different algorithm classes.

Nonetheless, there is more than just a model-centric approach to XAI, as data also plays an important role in an intelligent system [16].

Therefore, it is also necessary to evaluate the data with respect to its interpretability. Hard and soft information, a concept established in finance, relates to the contextual interpretation of data [21]. Information that can be interpreted context independently because of the clear representation in numerical or categorical variables is called *hard information*, like the price of a share. In contrast, *soft information* captures latent aspects that cannot be easily reduced to a number, like the mood and buying motivation of a potential customer. This information is highly context dependent and complex [21]. Following this argumentation, soft information is more difficult to interpret, as contextual aspects need to be considered for understanding. Additionally, while soft information might have a high value for the application, as they often capture previously omitted factors, the information is concealed in unstructured data coming from text or images [21]. The complex feature representation often reduces interpretability.

While the concept of hard and soft information mainly has a perspective related to content (information), it is worthwhile to decompose the data from a technical view as well. One technical aspect is the number of features in a model relating to its complexity. Even when an inherently transparent algorithm like a Linear Regression is used, thousands of variables significantly reduce the interpretability [22]. This relates to the concept of simulatability suggested by Lipton [15]. Simulatability describes the interpretability of a system when a human can understand and simulate the behavior of the algorithms at once. Systems that are highly complex due to many features are hard or impossible to understand for a human even if inherently transparent algorithms are used [3, 15]. Therefore, we separate models by the number of features. Models having *single to ten* features are cognitively processable by humans [23]. These systems are more interpretable than having *ten to hundreds, or hundreds to millions* of variables.

However, having only a few variables does not make a system interpretable per se. The expressiveness of the features also plays a significant role. When features are not interpretable by themselves, meaning that the unit of the feature is not understandable, it violates the interpretability concept of decomposability [15]. Decomposability refers to the property that each part of a model, input, parameters, and calculation have an intuitive interpretation [3]. For example, the used principal components of variables like in [24] is violating the assumption of decomposability due to anonymous features.

Consequently, the expressiveness of features can be separated into human understandable variables and feature creation methods [22]. When features are selected by hand [22, 23] and no transformations are applied (*observed*), interpretability tends to be high, as these variables represent the information in a human understandable way. Many sophisticated data preprocessing techniques like standardization or log-transformation of variables create less interpretable features [25]. Methods like polynomial features, dimensionality reduction like Principal Component Analysis (PCA) [24] or Topic Modeling for text data [26] create complex feature representations

(*designed*). These features are based on mathematical combinations and the original representation of the variable is changed. Therefore, the transformed features might not be inherently interpretable. Nevertheless, data exploration might help to explain these features. One example is to name a topic in a topic model based on the words in the topic [26]. The third group of features is called *generated*. Often, deep learning is used to extract features based on hidden layers of image data like in real estate appraisal to fuse tabular and image data [17, 27–29]. While deep learning generates these variables, they are impossible to interpret, because they have no human interpretable label. Hence the groups *observed*, *designed*, and *generated* for feature creation are ordered by descending interpretability. The expressiveness of the feature ranges from *inherent definition* (like square feet of a house) via *analysis per data exploration* (e.g., in Topic Models) to *anonymous* (e.g. PCA representation of features).

Table 1. Technical XAI taxonomy, ML=Machine Learning, DL = Deep Learning; Table rows can be assigned to the interpretability level independently of each other.

			Interpretability		
			high	**medium**	**low**
Model	**Algorithmic**	**Transparency**	Inherently interpretable models	Classical ML models	DL models
Feature	**Content**	**Information Type**	Hard Information		Soft Information
		Expressiveness	Inherent definition	Analysis per exploration	Anonymous
	Technical	**Number of Features**	Single to Ten	Ten to Hundred	Hundred to Millions
		Input Feature Creation	Observed	Designed	Generated

Finally, our proposed taxonomy (**Table 1**) supplements the model-centric perspective from the related work and divides the intelligent system into algorithms and features. Features are evaluated in terms of content according to their information type and expressiveness, and technically according to their number and method of creation, considering the complexity of the representation. The taxonomy should be read row-wise, evaluating each aspect of the XAI system independently, indicating potential areas for improving the interpretability of the system.

For discussion, we use the seven-gap framework proposed by Martens and Provost [12]. The authors use a socio-technical perspective on an intelligent system and state that the system's overall performance can be improved when the seven gaps are closed. While gap one (between model and truth) is closed by building more accurate models, three other gaps of the system (between the user-groups and the model) can be closed

when the user's mental decision model and the algorithmic model are aligned. The essential idea is that different user groups have different interpretability requirements, ranging from outcome justification to a deep understanding of the system. By providing a model that offers these aspects of interpretability, the gaps are closed. The proposed taxonomy (**Table 1**) has a technical focus on XAI systems, however several aspects relate to a human user by judging the interpretability on concepts like simulatability and decomposability [15], taking the human mental limitations and thus their decision model into account. Consequently, it helps to build models that are better understood or aligned with the user's mental model, helping to reduce the three gaps between the users and the model in the seven-gap model. This could increase user acceptance [11]. Following the argumentation of Martens and Provost [12], the effectiveness of the decision support system rises. From an application perspective, justification of decision and governance of the ML system can be established [2].

Although it may seem that an interpretable system is not well-performing as simpler algorithms, fewer variables, and no complex preprocessing technique are used, this might not be necessarily true. First, some authors [23] argue that these models can outperform complex black box ML techniques when the features are created intelligently. Second, we stress that performance in terms of ML and XAI might include different aspects. While in ML, accuracy relates to the algorithmic accuracy measured in F1-Score or root-mean-squared-error [14, 16], the algorithmic performance only relates to the first of the seven gaps in the model of Martens and Provost [12]. Performance in the light of the seven-gaps model is related to the overall effectiveness of an intelligent systems, taking the user-groups and the usage of the system into account [12].

4 Outlook

Our work is not without limitations. This paper is still research in progress and the evaluation and revision phase in the taxonomy development model is the next research step. Therefore, we aim to review data analytics studies in the domain of real estate appraisal about XAI and evaluate their suggested model based on our taxonomy. Furthermore, a post-hoc interpretability method dimension based on the effectiveness of the methods could be integrated in the taxonomy. While this short paper's contribution is the proposed taxonomy, future research should extend this research to a framework including a user dimension. Moreover, typologies of XAI systems should be deducted from the framework, taking their application frequencies into account. Implications for practice are that this taxonomy helps IT Managers to evaluate the current XAI capabilities of their intelligent system. In a development and implementation phase, this taxonomy can guide data scientists in their decision-making of what algorithms and features to use in their project. For research, this taxonomy provides the ability to standardize the comparison of XAI systems, potentially creating new research aspects. In conclusion, the proposed taxonomy helps to tap into the new class of XAI system with the potential to create higher user-acceptance of ML models and thus more effective decision support systems.

References

1. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A Survey of Methods for Explaining Black Box Models. ACM Comput. Surv. 51, (2018). https://doi.org/10.1145/3236009.
2. Adadi, A., Berrada, M.: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access. 6, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052.
3. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion. 58, 82–115 (2020). https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.012.
4. Meske, C., Bunde, E., Schneider, J., Gersch, M.: Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. Inf. Syst. Manag. 0, 1–11 (2020). https://doi.org/10.1080/10580530.2020.1849465.
5. Ribeiro, M.T., Singh, S., Guestrin, C.: "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. pp. 1135–1144. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778.
6. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017).
7. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity Checks for Saliency Maps. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.) Advances in Neural Information Processing Systems. pp. 9525–9536. Curran Associates, Inc. (2018).
8. Herm, L.-V., Wanner, J., Seubert, F., Janiesch, C.: I don't get it, but it seems valid! The connection between explainability and comprehensibility in (X) AI research. In: ECIS 2021 Proceedings (2021).
9. Wanner, J., Herm, L.-V., Heinrich, K., Janiesch, C., Zschech, P.: White, Grey, Black: Effects of XAI Augmentation on the Confidence in AI-based Decision Support Systems. In: ICIS 2020 Proceedings (2020).
10. Lukyanenko, R., Castellanos, A., Samuel, B., Tremblay, M., Maass, W.: Research Agenda for Basic Explainable AI. In: AMCIS 2021 Proceedings (2021).
11. Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Wortman Vaughan, J.W., Wallach, H.: Manipulating and Measuring Model Interpretability. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–52. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3411764.3445315.
12. Martens, D., Provost, F.: Explaining data-driven document classifications. Mis Q. 38, 73–100 (2014).
13. Du, M., Liu, N., Hu, X.: Techniques for Interpretable Machine Learning. Commun.

ACM. 63, 68–77 (2019). https://doi.org/10.1145/3359786.
14. Asatiani, A., Malo, P., Nagbøl, P.R., Penttinen, E., Rinta-Kahila, T., Salovaara, A.: Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems. J. Assoc. Inf. Syst. 22, 325–352 (2021).
15. Lipton, Z.C.: The Mythos of Model Interpretability. Queue. 16, (2018). https://doi.org/10.1145/3236386.3241340.
16. Sarker, I.H.: Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput. Sci. 2, (2021). https://doi.org/10.1007/s42979-021-00592-x.
17. Kucklick, J.-P., Müller, O.: A Comparison of Multi-View Learning Strategies for Satellite Image-Based Real Estate Appraisal. In: The AAAI-21 Workshop on Knowledge Discovery from Unstructured Data in Financial Services (2021).
18. Nickerson, R.C., Varshney, U., Muntermann, J.: A method for taxonomy development and its application in information systems. Eur. J. Inf. Syst. 22, 336–359 (2013). https://doi.org/10.1057/ejis.2012.26.
19. Webster, J., Watson, R.T.: Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Q. 26, xiii–xxiii (2002).
20. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature. 521, 436–444 (2015). https://doi.org/10.1038/nature14539.
21. Liberti, J.M., Petersen, M.A.: Information: Hard and Soft. Rev. Corp. Financ. Stud. 8, 1–41 (2019). https://doi.org/10.1093/rcfs/cfy009.
22. Gosiewska, A., Kozak, A., Biecek, P.: Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering. Decis. Support Syst. (2021). https://doi.org/https://doi.org/10.1016/j.dss.2021.113556.
23. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1, 206–215 (2019). https://doi.org/https://doi.org/10.1038/s42256-019-0048-x.
24. Kostic, Z., Jevremovic, A.: What Image Features Boost Housing Market Predictions? IEEE Trans. Multimed. 22, 1904–1916 (2020).
25. Molnar, C.: Interpretable Machine Learning - A Guide for Making Black Box Models Explainable. Leanpub, Victoria, BC, Canada (2020).
26. Debortoli, S., Müller, O., Junglas, I., vom Brocke, J.: Text mining for information systems researchers: An annotated topic modeling tutorial. Commun. Assoc. Inf. Syst. 39, 110–135 (2016).
27. Xu, C., Tao, D., Xu, C.: A survey on multi-view learning. arXiv Prepr. arXiv1304.5634. (2013).
28. Li, Y., Yang, M., Zhang, Z.: A survey of multi-view representation learning. IEEE Trans. Knowl. Data Eng. 31, 1863–1883 (2018).
29. Bency, A.J., Rallapalli, S., Ganti, R.K., Srivatsa, M., Manjunath, B.S.: Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, 24-31 March. pp. 320–329 (2017). https://doi.org/10.1109/WACV.2017.42.