

# Using Ontologies to Enhance Human Understandability of Global Post-hoc Explanations of Black-box Models (Extended Abstract)

Roberto Confalonieri<sup>1</sup>, Tillman Weyde<sup>2</sup>, Tarek R.Besold<sup>3</sup> and Fermín Moscoso del Prado Martín<sup>4</sup>

<sup>1</sup>Faculty of Computer Science, Free University of Bozen-Bolzano, Domenikanerplatz 3, Bolzano-Bozen, Italy

<sup>2</sup>Dept. of Computer Science, City, University of London, GB-EC1V 0HB London

<sup>3</sup>Philosophy & Ethics, Faculty of IE/IS, Eindhoven University of Technology, 5600 MB Eindhoven

<sup>4</sup>Lingvist Technologies OÜ, Tallinn, Estonia

## 1. Extended Abstract

Explainable AI (XAI) has been identified as a key factor for developing trustworthy AI systems [1, 2]. The reasons for equipping intelligent systems with explanation capabilities are not limited to user rights and acceptance. Explainability is also needed for designers and developers to enhance system robustness and enable diagnostics to prevent bias, unfairness, and discrimination, as well as to increase trust by all users in why and how decisions are made [3].

The interpretability of AI systems has been described long time ago since mid 1980s [4], but until recently it becomes an active research focus in computer science community due to the advances of big data and various regulations of data protection in developing AI systems, such as the GDPR. For example, according to the GDPR, citizens have the legal right to an explanation of decisions made by algorithms that may affect them (e.g., see Article 22). This policy highlights the pressing importance of transparency and interpretability in algorithm design.

XAI focuses on developing new approaches for explanations of black-box models by achieving good explainability without sacrificing system performance. In the ML literature, techniques for explaining black-box models are typically classified as local and global methods [5]. Whilst local methods take into account specific examples and provide local explanations, global methods aim to provide an overall approximation of the behavior of the black-box model. Global explanations are usually preferable over local explanations, because they provide a more general view about the decision making process of a black-box.

---

*3rd International Workshop on Data meets Applied Ontologies in Explainable Artificial Intelligence. DAO-XAI @ BAKS 2021*


✉ roberto.confalonieri@unibz.it (R. Confalonieri); t.e.veyde@city.ac.uk (T. Weyde); tarek.besold@gmail.com (T. R.Besold); fermosc@gmail.com (F.M. d. P. Martín)

🌐 <http://www.inf.unibz.it/~rconfalonieri/> (R. Confalonieri)

🆔 0000-0003-0936-2123 (R. Confalonieri)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Other approaches are based on hybrid or neuro-symbolic systems, advocating a tight integration between symbolic and non-symbolic knowledge, e.g., by combining symbolic and statistical methods of reasoning [6]. The construction of hybrid systems is widely seen as one of the grand challenges facing AI today. However, there is no consensus regarding how to achieve this, with proposed techniques in the literature ranging from knowledge extraction and tensor logic [7] to inductive logic programming and other approaches.

Knowledge representation—in its many incarnations as ontologies, knowledge graphs, etc.—is a key asset to enact hybrid systems, and it can pave the way towards the creation of transparent and human-centric explainable knowledge-enabled systems [8]. Linking explanations to structured knowledge, for instance in the form of ontologies, brings multiple advantages. First, they can be used to provide a sound and machine interpretable conceptualisation of the system, facilitating knowledge sharing and reuse. In this way it is easier to capture user requirements and promote the reuse of components [9]. Second, they can be used not only to enrich explanations (or the elements therein) with semantic information—thus facilitating evaluation and effective knowledge transmission to users—but they also create a potential for supporting the customisation of the levels of specificity and generality of explanations to specific user profiles or audiences [10].

This extended abstract builds on the work presented in [11] which describes an extension of TREPAN, a seminal global explanation approach that extracts surrogate decision trees from black-box models. TREPAN was extended to take into account explicit knowledge, modeled by means of ontologies, to extract human-understandable explanations.

TREPAN is a tree induction algorithm that recursively extracts decision trees from oracles, in particular from feed-forward neural networks [12]. The algorithm is model-agnostic, and it can be applied to explain any black-box classifier (e.g., Multi-Layer Perceptron, Random Forest). TREPAN combines the learning of the decision tree with a trained machine learning classifier (the oracle).

The proposed extension of the TREPAN algorithm, called TREPAN RELOADED, uses a modified information gain that, in the creation of split nodes, gives priority to features associated with more general concepts defined in a domain ontology. This was achieved by means of an information content measures defined using the idea of refinement operators [13].

The understandability of the extracted explanations was tested with humans in a user study with four different tasks. Results were evaluated in terms of response times and correctness, subjective ease of understanding and confidence, and similarity of free text responses. The results showed that decision trees generated with TREPAN Reloaded, taking into account domain knowledge, were significantly more understandable throughout than those generated by standard TREPAN. The enhanced understandability of post-hoc explanations was achieved with little compromise on the accuracy with which the surrogate decision trees replicate the behaviour of the original neural network models.

## References

- [1] A. B. Arrieta, N. D. Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelli-

- gence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion* 58 (2020) 82–115. doi:10.1016/j.inffus.2019.12.012.
- [2] R. Confalonieri, L. Coba, B. Wagner, T. R. Besold, A historical perspective of explainable artificial intelligence, *WIREs Data Mining and Knowledge Discovery* 11 (2021). doi:<https://doi.org/10.1002/widm.1391>. arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1391>.
- [3] E. Mariotti, J. M. Alonso, R. Confalonieri, A framework for analyzing fairness, accountability, transparency and ethics: A use-case in banking services, in: *IEEE International Conference on Fuzzy Systems 2021 (Fuzz-IEEE 2021)*, 2021. To appear.
- [4] M. R. Wick, W. B. Thompson, Reconstructive expert system explanation, *Artificial Intelligence* 54 (1992) 33–70.
- [5] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, *ACM Comp. Surv.* 51 (2018) 1–42.
- [6] A. D. Garcez, M. Gori, L. C. Lamb, L. Serafini, M. Spranger, S. N. Tran, Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning, *IfCoLoG Journal of Logics and their Applications* 6 (2019) 611–631. arXiv:1905.06088.
- [7] L. Serafini, A. S. d’Avila Garcez, Learning and reasoning with logic tensor networks, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10037 LNAI (2016) 334–348. doi:10.1007/978-3-319-49130-1\_25.
- [8] S. Chari, D. M. Gruen, O. Seneviratne, D. L. McGuinness, Foundations of Explainable Knowledge-Enabled Systems, in: *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*, volume 47 of *Studies on the Semantic Web*, 2020.
- [9] S. Chari, O. Seneviratne, D. M. Gruen, M. A. Foreman, A. K. Das, D. L. McGuinness, Explanation Ontology: A Model of Explanations for User-Centered AI, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12507 LNCS (2020) 228–243. doi:10.1007/978-3-030-62466-8\_15. arXiv:2010.01479.
- [10] M. Hind, Explaining Explainable AI, *XRDS* 25 (2019) 16–19. doi:10.1145/3313096.
- [11] R. Confalonieri, T. Weyde, T. R. Besold, F. Moscoso del Prado Martín, Using ontologies to enhance human understandability of global post-hoc explanations of black-box models, *Artificial Intelligence* 296 (2021). doi:<https://doi.org/10.1016/j.artint.2021.103471>.
- [12] M. W. Craven, J. W. Shavlik, Extracting tree-structured representations of trained networks, in: *NIPS 1995*, MIT Press, 1995, pp. 24–30.
- [13] N. Troquard, R. Confalonieri, P. Galliani, R. Peñaloza, D. Porello, O. Kutz, Repairing Ontologies via Axiom Weakening, in: *AAAI 2018*, 2018, pp. 1981–1988.