Knowledge-driven and Human-understandable Explanations of Black Box Models and Process Mining

Roberto Confalonieri, Pietro Galliani, Oliver Kutz, Marco Montali,
Guendalina Righetti, Nicolas Troquard, and Markus Zanker
Faculty of Computer Science, Free University of Bozen-Bolzano, Bolzano, Italy
firstname.surname@unibz.it

1 Research at UniBZ

Approaches to *Responsible and Trustworthy AI* at **unibz** can be divided into three main research streams: (1) integration of symbolic and sub-symbolic reasoning for semantics-aware explanations of black box models, (2) explainable recommender systems, and (3) trustworthy process mining.

We have built on and extended existing approaches approximating the behaviour of black-box models using interpretable symbolic representations such as, for instance, decision trees. In particular, in [Confalonieri *et al.*, 2020b], we proposed Trepan Reloaded, a knowledge-driven algorithm for post-hoc global explanations of black box models. Trepan Reloaded extends Trepan [Craven e Shavlik, 1995] using ontologies to guide the extraction of *semantics-aware explanations*. Linking explanations to structured knowledge, in the form of ontologies, brings multiple advantages. It does not only enrich explanations (or the elements therein) with semantic information—thus facilitating effective knowledge transmission to users—, but it also creates a potential for supporting the customisation of the levels of specificity and generality of explanations to specific user profiles. Whilst this provides us with an example of the benefits of integrating symbolic and non-symbolic ML artefacts, and decision trees are considered interpretable models, there is the need to use richer representations to support more advanced forms of common-sense reasoning. Besides, global explanations are difficult to build and to be understood by human users. Local explanations are simpler and can be beneficial in different contexts. To this end, we started to explore the integration of linear models with logical knowledge bases via so-called threshold operators [Porello *et al.*, 2019; Galliani *et al.*, 2019], with an eye towards practical applications for locally explaining richer (black-box) models and integrating them in logical knowledge bases [Confalonieri *et al.*, 2020a].

Other forms of explanability techniques are applied to explaining black box algorithms for *explainable recommender systems*. Model-based explanation algorithms such as Explainable Matrix Factorisation [Coba *et al.*, 2019] were proposed to generate recommendations that were explainable by design. This was achieved by constraining the loss function of Matrix Factorisation algorithm with an interpretable component. Among the activities carried out in this research line, we have also addressed the challenge of replicability of prior algorithms when creating and evaluating new approaches. Aiming to address the resulting need, we have proposed em recoXplainer [Coba *et al.*, 2022]. *recoXplainer* is a unified, extendable and easy to use library that includes several state-of-the-art explainability methods, and evaluation metrics that are useful for various groups of stakeholders.

A common problem of the two lines of research, and also for Responsible and Trustworthy AI, is how to evaluate the perceived understandability of the explanations by human users. Previous work attempting to measure the understandability of symbolic decision models (e.g., [Huysmans *et al.*, 2011]) proposed syntactic complexity measures based on the model's structure. The syntactic complexity of an explanation can be measured, for instance, in the case of decision trees, by counting the number of internal nodes or leaves, or in the case of logical formulas, by counting the number of symbols adopted. Having a measure like syntactic complexity, that can be easily computed, is useful from an application perspective. E.g., it may be used to prevent excessive complexity in building decision trees and threshold expressions when explaining a black box. On the other hand, the syntactic complexity does not necessarily capture precisely the understandability of explanations by users. A direct measure of user understandability is how accurately a user can employ a given explanation to perform a decision. Another measure of cognitive difficulty is the reaction time (RT) or response latency [Donders, 1969]. RT is a standard measure used by cognitive psychologists and has become a staple measure of complexity in the domain of design and user interfaces [William Lidwell, 2003]. Understandability depends on the cognitive load experienced by users, e.g., in using the decision model to classify instances and in understanding the features in the model itself. However, for practical processing human understandability needs to be approximated by an objective measure. In [Confalonieri *et al.*, 2021] we assessed the understandability of explanations in two ways: i) implicitly, based on the syntactic complexity of an explanation (number of internal nodes, leaves, symbols used in a weighted formulas, etc.) ii) explicitely, based on users' performances and subjective ratings, reflecting, for instance, the cognitive load by users in carrying out tasks using a given explanation format.

In a parallel line of research, we are investigating how AI techniques can be used towards explainable, trustworthy *process mining*. Process mining stems from the increasing availability of digital traces recording the execution of processes in contemporary organizations, with the goal of extracting factual insights about such processes. These, in turn, can be used to reduce operational frictions and take informed decisions for continuous process improvement. Solid process mining techniques have been produced to handle the automated discovery of process models from event data, conformance checking of event data to detect deviations between the expected and the observed behavior, and predictive monitoring to predict what will likely happen. However, contemporary techniques suffer of two key limitations: i) the process mining lifecycle lacks documentation and traceability, due to the heterogeneity of its steps, the presence of ad-hoc procedures, and the usage of black-box components that do not provide interpretable results; ii) process mining pipelines do not integrate and use domain knowledge to empower learning and inference algorithms towards more meaningful results.

Research in this context has so far been conducted within **unibz** along two directions. On the one hand, an extension of the ontology-based data access framework [Calvanese *et al.*, 2018] has been employed to provide more transparent methodologies and tools for extracting event logs from legacy data sources [Calvanese *et al.*, 2017]. On the other hand, logic-based techniques have been investigated in the context of conformance checking and monitoring, towards comparing recorded with expected behaviors and detecting the sources of non-conformance [Maggi *et al.*, 2011; Felli *et al.*, 2021].

This research will be intensified in the near future through the *PINPOINT PRIN2020 Italian Project*, which indeed focuses on explanations in symbolic systems, in neural-symbolic learning and reasoning, and via knowledge extraction towards *exPlaInable kNowledge-aware PrOcess INTelligence*. The project is coordinated by **unibz**, with four additional participating partners (Sapienza University of Rome, University of Milano-Bicocca, the National Research Council of Italy, and the University of Calabria), as well as Fondazione Bruno Kessler and three private companies as supporting institutions.

2 Involved People

Research on Responsible and Trustworthy AI is carried out by members from two research groups at the Faculty of Computer Science of **unibz**: i) Information and Database Systems Engineering (IDSE): Roberto Confalonieri and Markus Zanker; Research Centre for Knowledge and Data (KRDB): Pietro Galliani, Oliver Kutz, Guendalina Righetti, Nicolas Troquard, and Marco Montali.

3 Applications

Explaining black box models in the form of symbolic explanations can be applied in a variety of applications. Trepan Reloaded have been applied to explain decisions made in sensitive domains such as health and finance, for instance in predicting a disease or granting a loan. Experiments based on the Gene Ontology annotations on yeast proteins (published in [Galliani *et al.*, 2020]) showed that even simple such expressions can capture non-trivial properties with accuracy comparable to that of much more complex (and less interpretable) machine learning approaches.

Identifying and mitigating algorithmic fairness is an important requirement for explainable, and transparent AI aided decision. Extracting decision trees can be useful to identifying and justifying algorithmic bias [Hajian *et al.*, 2016], in particular, to understand if any (undesirable) discrimination features are used in a black-box classifier. Explanations of black-boxes, e.g., in the form of extracted decision trees, could provide a means to identify biases in black-box models. Preliminaries results have been presented in the context of granting a loan in [Mariotti *et al.*, 2021], and for managing sex and gender bias in AI models for healthcare in [Confalonieri *et al.*, 2022].

4 Libraries and Tools

In relation to the research lines carried out and described above, the following tools and libraries have been developed:

- *recoXplainer*[1] is a Python library for development and evaluation of explainable recommender systems. The library is an easy-to-use, unified and extendable library that supports the development and evaluation of explainable recommender algorithms. *recoXplainer* includes several state-of-the-art black box algorithms, model-based and post-hoc explainability techniques, as well as offline evaluation metrics in order to assess the quality of the explanation algorithms [Coba *et al.*, 2022].
- *Trepan Reloaded*[2] is a prototype that extends Trepan, a model-agnostic algorithm written in C, Java and Python that extracts global explanations, in the form of decision trees, of pre-trained black box models (e.g., MLP and Random Forest). The extraction of global post-hoc explanations takes into account a domain ontology, modeled in OWL. The ontology defines the semantics of the features in the dataset. The extraction process is guided by the ontology and decision trees are generated by giving priority to features associated to more general concepts defined in the ontology.
- *ToothLearner*: an experimental Java library to learn threshold expressions from data and add it to ontologies, together with a framework to translate (with a polynomial overhead) the resulting expressions into OWL language.

5 Challenges and Perspectives

Much work remains to be done, on the theoretical and the practical side. We envisage the following future research activities:

Framework for the evaluation of perceived understandability of explanations by human users. Assessing the perceived understandability of explanations by human users is usually done through user studies. Designing and running a user study is however a difficult task. One has to decide what kind of questions to include, what

[1] https://www.inf.unibz.it/~rconfalonieri/recoxplainer/

[2] https://github.com/rconfalonieri/trepan_reloaded

participant to look for etc. Assessing explanations and their perceived understandability depends on the application domain and on the explanation goals. For instance, in recommender systems it is of interest to evaluate whether an explanation is persuasive, whereas in precision medicine whether it can be trusted. Another open issue is how to measure the causal understanding achieved by explanations, e.g., in terms of effectiveness, efficiency, satisfaction related to causal understanding and its transparency for a user [Holzinger *et al.*, 2019]. Providing a framework able to guide researchers in the design of a user study, for instance, recommending which kind of questions to include and which analysis techniques to adopt, could be a useful asset for the community.

Adapting explanations to user profiles. An important requirement of explanations is that they have to be user-centered, accommodating different user profiles [Ribera e Lapedriza, 2019]. For instance, an explanation served to a technical person, i.e., a doctor, containing a lot of technical details, is not usually suitable for a lay audience, that generally seeks explanations closer to its knowledge background. Being able to change the level of abstraction and adapt to different user profiles is a desirable property of explanations. For this purpose we will adopt the idea of concept refinement operators, which were applied in the context of ontology repair [Troquard *et al.*, 2018; Porello *et al.*, 2018], to refine the theory based on which explanations are distilled.

Integrating existing knowledge graphs. In its current form, Trepan Reloaded is based on a predefined ontology onto which the features used by our algorithm should be mapped. An interesting improvement is to automatically construct the most appropriate ontology to be mapped onto. Such a process could be achieved by automatically mapping sets of features into pre-existing general domain ontologies (e.g., MS Concept Graph [Wu *et al.*, 2012], DBpedia [Lehmann *et al.*, 2015], or WordNet [Miller, 1995]).

More expressive threshold expressions. Threshold expressions already provide a natural and interpretable link between statistical learning of concept from data and logical reasoning in knowledge bases, but they may be further extended in order to increase their expressivity. A particularly promising avenue of research in this second sense is given by the study of threshold connectives able to count role-relation successors. This will extend the known framework of threshold expressions in Description Logic in a natural and useful way: as a practical use case, one could in this way represent the driving license score ('patente a punti') of a person by expressing, in a natural and easily readable and yet computationally adequate way, the fact that six points are lost for each single traffic light infraction. Such a statement can then be directly integrated into a rich formal ontology. [Galliani *et al.*, 2021] contains some results about such more expressive threshold operators.

Comparing and integrating threshold expressions and decision trees. As discussed in [Confalonieri *et al.*, 2020a], decision trees and threshold expressions are two different and to some degree complementary approaches to integrating concepts learned from data into knowledge bases. We aim to study how these two frameworks interact on a theoretical level, and secondly, to investigate use-cases in ML and AI in a comparative manner, specifically user-studies that help determine human understandability of explanations generated using these two frameworks by human users.

Integrating background and commonsense knowledge in process mining. As pointed out in [Calvanese *et al.*, 2021], a particularly challenging, open problem in process mining is how background knowledge (both domain-specific and stemming from common sense) can be integrated within process mining techniques so as to improve the quality and trustworthiness of the produced results.

References

[Calvanese *et al.*, 2017] Diego Calvanese, Tahir Emre Kalayci, Marco Montali, and Stefano Tinella. Ontology-based data access for extracting event logs from legacy data: The onprom tool and methodology. In *Proc. of the 20th International Conference on Business Information Systems (BIS 2017)*, volume 288 of *Lecture Notes in Business Information Processing*, pages 220–236. Springer, 2017.

[Calvanese *et al.*, 2018] Diego Calvanese, Tahir Emre Kalayci, Marco Montali, Ario Santoso, and Wil M. P. van der Aalst. Conceptual schema transformation in ontology-based data access. In *Proc. of the 21st International Conference on Knowledge Engineering and Knowledge Management (EKAW 2018)*, volume 11313 of *Lecture Notes in Computer Science*, pages 50–67. Springer, 2018.

[Calvanese *et al.*, 2021] Diego Calvanese, Sanja Lukumbuzya, Marco Montali, and Mantas Simkus. Process mining with common sense. In *Proc. of the International Workshop on BPM Problems to Solve Before We Die (PROBLEMS 2021)*, volume 2938, pages 45–50. CEUR-WS.org, 2021.

[Coba *et al.*, 2019] Ludovik Coba, Panagiotis Symeonidis, and Markus Zanker. Personalised novel and explainable matrix factorisation. *Data and Knowledge Engineering*, 2019.

[Coba *et al.*, 2022] Ludovik Coba, Roberto Confalonieri, and Markus Zanker. recoxplainer: A library for development and offline evaluation of explainable recommender systems. *IEEE Computational Intelligence Magazine*, 17(1):46–58, 2022.

[Confalonieri *et al.*, 2020a] Roberto Confalonieri, Pietro Galliani, Oliver Kutz, Daniele Porello, Guendalina Righetti, and Nicolas Troquard. Two knowledge-driven approaches to explaining black-box models. In *Proc. of Explainable Logic-Based Knowledge Representation (XLoKR 2020)*, 2020.

[Confalonieri *et al.*, 2020b] Roberto Confalonieri, Tillman Weyde, Tarek R. Besold, and Fermín Moscoso del Prado Martín. Trepan Reloaded: A Knowledge-driven Approach to Explaining Black-box Models. In *Proc. of the 24th*

European Conference on Artificial Intelligence, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2457–2464. IOS press, 2020. **Distinguished Paper Award**.

[Confalonieri *et al.*, 2021] Roberto Confalonieri, Tillman Weyde, Tarek R. Besold, and Fermín Moscoso del Prado Martín. Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. *Artificial Intelligence*, 296, 2021.

[Confalonieri *et al.*, 2022] Roberto Confalonieri, Federico Lucchesi, Giovanni Maffei, and Silvina Catuara Solarz. A unified framework for managing sex and gender bias in AI models for Healthcare. In *Sex and Gender Bias in Technology and Artificial Intelligence*. Elsevier, 2022. To appear.

[Craven e Shavlik, 1995] Mark W. Craven and Jude W. Shavlik. Extracting tree-structured representations of trained networks. In *NIPS 1995*, pages 24–30. MIT Press, 1995.

[Donders, 1969] Franciscus Cornelis Donders. On the speed of mental processes. *Acta Psychologica*, 30:412–31, 1969.

[Felli *et al.*, 2021] Paolo Felli, Alessandro Gianola, Marco Montali, Andrey Rivkin, and Sarah Winkler. Cocomot: Conformance checking of multi-perspective processes via SMT. In *Proc. of the 19th International Conference on Business Process Management (BPM 2021)*, volume 12875 of *Lecture Notes in Computer Science*, pages 217–234. Springer, 2021.

[Galliani *et al.*, 2019] Pietro Galliani, Oliver Kutz, Daniele Porello, Guendalina Righetti, and Nicolas Troquard. On knowledge dependence in weighted description logic. In *GCAI 2019. Proceedings of the 5th Global Conference on Artificial Intelligence*, volume 65, pages 68–80, 2019.

[Galliani *et al.*, 2020] Pietro Galliani, Guendalina Righetti, Oliver Kutz, Daniele Porello, and Nicolas Troquard. Perceptron connectives in knowledge representation. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 183–193. Springer, 2020.

[Galliani *et al.*, 2021] Pietro Galliani, Oliver Kutz, and Nicolas Troquard. Perceptron operators that count. In *Proc. of the 34th International Workshop on Description Logics (DL 2021). CEUR Workshop Proceedings, CEURWS. org*, 2021.

[Hajian *et al.*, 2016] Sara Hajian, Francesco Bonchi, and Carlos Castillo. Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 2125–2126. ACM, 2016.

[Holzinger *et al.*, 2019] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

[Huysmans *et al.*, 2011] Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154, 2011.

[Lehmann *et al.*, 2015] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2):167–195, 2015.

[Maggi *et al.*, 2011] Fabrizio Maria Maggi, Michael Westergaard, Marco Montali, and Wil M. P. van der Aalst. Runtime verification of ltl-based declarative process models. In Sarfraz Khurshid and Koushik Sen, editors, *Proceedings of the 2nd International Conference on Runtime Verification (RV 2011)*, volume 7186 of *Lecture Notes in Computer Science*, pages 131–146. Springer, 2011.

[Mariotti *et al.*, 2021] Ettore Mariotti, José M. Alonso, and Roberto Confalonieri. A framework for analyzing fairness, accountability, transparency and ethics: A use-case in banking services. In *30th IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2021, Luxembourg, July 11-14, 2021*, pages 1–6. IEEE, 2021.

[Miller, 1995] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[Porello *et al.*, 2018] Daniele Porello, Nicolas Troquard, Rafael Peñaloza, Roberto Confalonieri, Pietro Galliani, and Oliver Kutz. Two Approaches to Ontology Aggregation Based on Axiom Weakening. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 1942–1948. ijcai.org, 2018.

[Porello *et al.*, 2019] Daniele Porello, Oliver Kutz, Guendalina Righetti, Nicolas Troquard, Pietro Galliani, and Claudio Masolo. A toothful of concepts: Towards a theory of weighted concept combination. In *Proceedings of the 32nd International Workshop on Description Logics, Oslo, Norway, June 18-21, 2019*. 2019.

[Ribera e Lapedriza, 2019] Mireia Ribera and Àgata Lapedriza. Can we do better explanations? a proposal of user-centered explainable ai. In *IUI Workshops*, volume 2327 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.

[Troquard *et al.*, 2018] Nicolas Troquard, Roberto Confalonieri, Pietro Galliani, Rafael Peñaloza, Daniele Porello, and Oliver Kutz. Repairing Ontologies via Axiom Weakening. In *Proc. of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1981–1988. AAAI Press, 2018.

[William Lidwell, 2003] Jill Butler William Lidwell, Kritina Holden. *Universal. Principles of Design.* Rockport, 2003.

[Wu *et al.*, 2012] Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. Probase: A Probabilistic Taxonomy for Text Understanding. In *Proc. of the 2012 ACM SIGMOD Int. Conf. on Management of Data*, SIGMOD '12, pages 481–492. ACM, 2012.