Constructing Meaningful Explanations: Logic-based Approaches

Laura State
laura.state@di.unipi.it
University of Pisa and Scuola Normale Superiore
Pisa, Italy

ABSTRACT

Machine learning (ML) models are ubiquitous: we encounter them when using a search engine, behind online text translation, etc. However, these models have to be used with care, as they are susceptible to social biases. Further, most ML models are inherently opaque, another obstacle to understand and verify them.

Being concerned with *meaningful* explanations, this work is putting forward two research paths: constructing counterfactual explanations with prior knowledge, and reasoning over explanations and time. Prior knowledge has the potential to significantly increase explanation quality, whereas time dimensions are necessary to track changes in ML models and explanations. The proposal builds on *(constraint) logic programming* and *meta-reasoning*. While situated in the computer sciences, it strives to reflect the interdisciplinary character of the field of *eXplainable Artificial Intelligence*.

1 INTRODUCTION

Machine learning (ML) systems are susceptible to social biases, potentially increasing and systematizing harm done to already marginalized groups. Further, many high performing ML systems are not interpretable. Algorithmic auditing, adopting explanations, can help us to uncover and quantify these harms. The field providing these explanations is *eXplainable Artificial Intelligence* (XAI).

Constructing a *meaningful* explanation tool, i.e. a tool that provides clear and easily understandable, possibly interactive, explanations to end users, is the central concern of my thesis work. It builds around four key challenges: definition of explanations in XAI, prior knowledge integration, time dimensions and evaluation of explanations [5]. While challenges are interconnected, I primarily focus on the integration of prior knowledge and time dimensions. To this end, I use *formal logic*. It has several important properties [1, 3, 4]: inherent interpretability and verifyability, straightforward integration of prior knowledge and support of reasoning under mixed settings (deductive, abductive, inductive, meta-reasoning).

2 MAIN RESEARCH PATHS

In my work, I understand an explanation as "an exchange of information", being rather active than passive. It matters who the target audience of an explanation is (most often lay users) and why the explanation is provided (problem and purpose) [5]. As such, explanations are *context-dependent*.

Counterfactuals Specifically for counterfactual (CF) explanations, I am focusing on the integration of prior knowledge. Here, prior knowledge refers to any knowledge that is important in the decision making process and the respective domain but that does not emerge from the ML model or data itself. It has potential to significantly enhance the quality of explanations. In a loan application scenario (LAS) that involves a ML model in the decision, a CF is able to answer questions such as: "What do I need to do to succeed in the next loan application?" CF were introduced to the field of XAI in 2017 [6]. Prior knowledge can be integrated in the form of constraints, and can be adapted depending on the intended purpose of the CF explanation. An example of such a constraint in the LAS is the minimum loan amount. To generate the CF explanations, I exploit the expressive power of constraint logic programming.

Time Dimensions ML models are rarely static, but likely to change over time, e.g., by retraining on the incoming data stream. I am examining logic structures such as the meta-interpreter, combined with theories [2], to reason over explanations. It allows integrating time dimensions into explanations, i.e. to track and compare how ML models and explanations evolve. In the LAS, the framework would be able to answer questions such as "Why did the rating of my loan application change, compared to last month?" Potentially, this line builds on the CF approach as presented above, and can be augmented by space dimensions, relevant to compare ML models and explanations in different contexts (countries, neighborhoods, individuals, etc.). Relevant use cases are redlining, individual fairness (space), and feedback loops in ML systems (time).

ACKNOWLEDGMENTS

I want to thank my supervisors S. Ruggieri and F. Turini. This work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Actions (grant agreement number 860630) for the project "NoBIAS - Artificial Intelligence without Bias" (*nobias-project.eu*). This work reflects only the authors' views and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] Krzysztof Apt. 1997. *From logic programming to Prolog.* Prentice Hall.
[2] Antonio Brogi, Paolo Mancarella, Dino Pedreschi, and Franco Turini. 1991. Theory Construction in Computational Logic. In *ICLP Workshop on Construction of Logic Programs.* Wiley, 241–250.
[3] Andrew Cropper and Sebastijan Dumancic. 2020. Inductive logic programming at 30: a new introduction. *CoRR* abs/2008.07912 (2020).
[4] Stuart J. Russell and Peter Norvig. 2003. *Artificial Intelligence: A Modern Approach* (2 ed.). Pearson Education.
[5] Laura State. 2021. Logic Programming for XAI: A Technical Perspective. In *ICLP Workshops (CEUR Workshop Proceedings, Vol. 2970).* CEUR-WS.org.
[6] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *CoRR* abs/1711.00399 (2017).

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08.
https://doi.org/10.1145/3514094.3539544