The Explanation Dialogues: Understanding How Legal Experts Reason About XAI Methods

Laura State[1,2], Alejandra Bringas Colmenarejo[3], Andrea Beretta[4], Salvatore Ruggieri[1], Franco Turini[1] and Stephanie Law[3]

[1]Department of Computer Science, University of Pisa, Largo B. Pontecorvo, 3, 56127 Pisa, Italy
[2]Scuola Normale Superiore, Piazza dei Cavalieri, 7, 56126 Pisa, Italy
[3]School of Law, University of Southampton, 4 University Rd, Southampton SO17 1BJ, United Kingdom
[4]ISTI-CNR, Via G. Moruzzi, 1, 56124 Pisa, Italy

Abstract

The *Explanation Dialogues* project is an expert focus study that aims to uncover expectations, reasoning, and rules of legal experts and practitioners towards explainable artificial intelligence (XAI). We examine legal perceptions and disputes that arise in a fictional scenario that resembles a daily life situation - a bank's use of an automated decision-making (ADM) system to decide on credit allocation to individuals. Through this simulation, the study aims to provide insights into the legal value and validity of explanations of ADMs, identify potential gaps and issues that may arise in the context of compliance with European legislation, and provide guidance on how to address these shortcomings.

Keywords

Explainability, AI, Automated Decision-Making, General Data Protection Regulation

The field of *explainable artificial intelligence* (XAI) provides tools to understand *automated decision-making* (ADM) systems, which, in turn, are often based on not interpretable machine learning (ML) models. While a significant number of approaches are already put forward, we know little about the perception of these explanations through legal experts. This is a critical drawback, given that one of the main motivating factors to construct these explanations is existing regulation such as the European *General Data Protection Regulation* (GDPR)[1], emerging regulative frameworks on AI such as the *Proposal for the European AI Act* [2], or, as an example from outside of Europe, the American *Blueprint for an AI Bill of Rights* [3].

With this work, we are aiming to close this gap: we present the *Explanation Dialogues*, an expert focus study that is designed to reveal legal expectations, reasoning, and rules of legal experts and practitioners towards explanations of ADM systems. We strongly expect that this novel knowledge can help jurists and computer scientists judge the legal value and validity of explanations, i.e., identifying the potential gaps and problems XAI can pose in the context of GDPR compliance and pointers on how to account for these shortcomings.

Explanations Dialogues is designed to offer insights regarding the following questions: i) How do legal experts reason about explanations for ADM systems, and how do they judge the

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland

[1]https://gdpr.eu/
[2]https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206
[3]https://www.whitehouse.gov/ostp/ai-bill-of-rights/

legal compliance of existing methods? ii) Do legal experts understand and trust explanations for ADM systems, and what are the steps identified to go forward?

In concrete, we have developed a focus study where legal experts will be questioned about a constructed, real-case scenario involving a private bank, an ADM system and an internal consultant. The bank provides explanations about the ADM process used to assess its customers' creditworthiness and, acting as internal consultants of the bank, the legal experts are expected to evaluate compliance with the information and explanations concerning the interest and duties of the bank and the interest and rights of the data subjects.

1. Legal Background

The project has been developed under the legal basis of the *right to not be subject to an automated decision-making system*, as referred to in Article 22 of the GDPR, and the safeguards and information duties established upon it in the same European Regulation. For the project, we hold the view that the *rights to information and an explanation* about automated decision-making arise from different Articles of the GDPR, Article 22(3), and Articles 13(2)(h), 14(2)(g) and 15(1)(h) respectively. Firstly, Articles 13(2)(h) and 14(2)(g) of the GDPR establish a duty for the data controller to provide information to the data subject regarding the existence of automated decision-making, meaningful information about the logic involved, and the significance and the envisaged consequences. As a result of this duty, the data subject has an ex-ante right to information which does not need to be actively exercised. Additionally, Article 15(1)(h) assesses the rights of the data subject to the access and requirement of information, including information regarding the existence of automated decision-making, meaningful information about the logic involved, and the significance and the envisaged consequences. As the exercise of this right resides in the active exercise of it by the data subject, it is arguable that it can entail ex-ante or ex-post information about the decision. Given that it is the subject who has to exercise that right, it is apparent that Article 15(1)(h) covers situations where either no automated decision affecting the subject has yet been taken or a decision has been taken and the subject sought to corroborate whether it is an automated or not. In the first case, the information received by the data subject shall be the same as for Articles 13(2)(h) and 14(2)(g). In the second case, on the contrary, the information shall be as set out in Article 22 (3) and its Recital 71. Thus, this ex-post information about the particular decision shall explain the decision reached after such assessment to contest the decision and express the point of view of the data subject. Indeed, the first case will involve a right to information, while the second will entail a right to explanation. Finally, Article 22 establishes both a right to information and a right to an explanation. The first will arise from the ex-ante necessity of the data subject to consent to and enter into a contract with automated decision-making (systems) as referred to in the exceptions of Article 22(2). The second will arise from the ex-post effective exercise of the safeguards set out in Article 22(3), namely, to express his or her point of view and to contest the decision.

This assessment of the rights to information and an explanation presents the legal motivation behind our simulation. However, we will refrain from transmitting this view to our project participants in order to avoid influencing their responses and their approach to the explanations and information provided. We expect participants to evaluate whether the provided examples

comply with the provisions of the GDPR and to identify the possible gaps or problems, wherefore, respecting their own rights' assessment and judgment is of special relevance.

2. Technical Background

In the experiment, we present five different global and local explanation methods. While global explanation methods focus on the full model, local methods describe the model behavior only close to the data instance in focus (here: an application instance). Also, all presented methods are model-agnostic, i.e. can generally be used to explain any type of underlying ML model [1].

We present global and local explanations in the form of SHAP values [2], global explanations as PDPs [3], local explanations in the form of counterfactual data instances computed by DICE [4], and (counter-)factual rules computed by LORE [5].

SHapley Additive exPlanations [2]: SHAP values rely on a game-theoretic approach. Local SHAP values represent contributions of features towards the ADM risk score value for a single application instance. Global SHAP values are calculated as averages over local SHAP values.

Partial Dependence Plot [3]: PDPs show the marginal effect of a feature towards the ADM risk score. PDPs are calculated by varying the range of feature values and then computing the average ADM risk score by keeping that value fixed over a set of application instances.

LOcal Rule-based Explanations [5]: LORE provides if-then rules as explanations. Both a factual rule ("Why did the model decide that way?") and a contrastive rule ("What has to change in which way such that the decision will change?") is presented.

DIverse Counterfactual Explanations [4]: DICE provides single contrastive examples in the form of data points. Such points are computed based on a minimum distance to the original data instance (the application instance), however, they must receive the opposite decision (different risk score) by the ADM system.

3. Experiment and Interview Details

The *Explanation Dialogues* project is realized via expert interviews where participants play the role of internal legal consultants requested by a bank to analyse a set of explanations about automated decisions on the granting and refusal of credit. For the project, participants are presented with two different cases: explanations for customers that applied for credit and were correctly rejected from the credit ("true positive"), and explanations for customers that applied for credit and were falsely rejected, i.e. customers that were respectively considered correctly or incorrectly as "high risk" creditors by the ADM system ("false positive"). For each case scenario, five different types of explanations are developed. Questions focus on the analysis of single explanations, and to compare between different methods and cases.

We decided to rely on an hypothetical case, rather than observing and analyzing the process followed by a real company to provide information and explanation to its customers, as due to

time and resources constrains the later would require to take under consideration too many variants and explanations, whether the later allow us to set a quasi-controlled scenario.

Randomization and Design The explanations are randomly sampled per participant - we sample three out of five (one global, two local). This is to cover different explanation methods between different participants (subjects) but avoid an overload per participant. In each of the two cases, therefore, the same explanation method is presented to the same participant, i.e., if for a participant PDP explanations were sampled, they are presented in both cases. This is to compare answers within the same participant. Also, the order of the presentation of the two cases is random, i.e., whether a participant first learns about the "true positive" or the "false positive" case. Thus, we use a mix of within- and between-subject design.

Between the two cases, participants are asked to answer questions to facilitate the comparison of answers within each case. After the online (written) interview, we will offer the opportunity to all participants to engage in follow-up interviews, and aim to clarify and gain more insights about the answers provided.

Participants Selection Since *Explanation Dialogues* seeks to gather the expertise and knowledge of academics and professionals with reputable and renowned careers in legal matters and compliance with AI systems, we will contact around thirty participants –including academics, researchers, and professors- selected through purposive sampling, on the criteria that they are legal experts on the GDPR of the European Union, particularly on explainability and interpretability of ADM systems.

Evaluation The analysis of the responses will be carried out through qualitative and quantitative methods. We will analyze the open-ended questions of the (written) interview through a qualitative analysis of the experts' judgments about the explanations provided, with a focus on understanding whether they consider the explanations to comply with the law and which improvements should be made. We will also ask some questions about the general understanding of the method - but this is not the main focus point of this study. We will conduct a thematic analysis based on grounded theory (following Glaser and Strauss), from which we intend to obtain knowledge that cannot be explored with closed questions. On the other hand, we will use computational tools for summarization, exploration, and visualization in the quantitative analysis of response data.

4. Outlook

It is critical to understand that while our work is closely related to those concerning users' interactions and experiences towards XAI [6, 7], as well as their perceptions of justice on automated decisions [8], our main purpose is distinct. While we also care about the understandability of explanations, *our* central concern is the assessment of explanations by legal experts. In other words, the study is set up to investigate how explanations about ADM systems as provided by XAI tools are perceived and disputed by legal experts and scholars in a fictional scenario resembling a daily life situation.

This work is a highly interdisciplinary contribution between the law, the computer sciences, the social sciences, and the cognitive sciences. Further - by putting forward an expert focus study based on a fictions use-case in the credit domain and establishing links towards the European legislative framework - we strongly acknowledge the need for *contextual* work in the domain of XAI.

Acknowledgments

Work supported by the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Actions for the project NoBIAS (g.a. No. 860630), and under the Excellent Science European Research Council (ERC) programme for the XAI project (g.a. No. 834756). 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] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2019) 93:1–93:42.

[2] S. M. Lundberg, S. Lee, A unified approach to interpreting model predictions, in: NIPS, 2017, pp. 4765–4774.

[3] C. Molnar, Interpretable Machine Learning, 2019. https://christophm.github.io/interpretable-ml-book/.

[4] R. K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: FAT*, ACM, 2020, pp. 607–617.

[5] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, F. Giannotti, Local rule-based explanations of black box decision systems, CoRR abs/1805.10820 (2018).

[6] Q. V. Liao, K. R. Varshney, Human-centered explainable AI (XAI): from algorithms to user experiences, CoRR abs/2110.10790 (2021).

[7] Y. Rong, T. Leemann, T. Nguyen, L. Fiedler, T. Seidel, G. Kasneci, E. Kasneci, Towards human-centered explainable AI: user studies for model explanations, CoRR abs/2210.11584 (2022).

[8] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, N. Shadbolt, 'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions, in: Proceedings of the 2018 Chi conference on human factors in computing systems, 2018, pp. 1–14.