Impact of Legal Requirements on Explainability in Machine Learning

Adrien Bibal * 1 Michael Lognoul * 2 Alexandre de Streel 2 Benoît Frénay 1

1. Legal Requirements on Explainability

The requirements on explainability imposed by European laws and their implications for machine learning (ML) models are not always clear. In that perspective, our research (Bibal et al., Forthcoming) analyzes explanation obligations imposed for private and public decision-making, and how they can be implemented by machine learning techniques.

For decisions adopted by firms or individuals, we mainly focus on requirements imposed by general European legislation applicable to all the sectors of the economy. The obligations of the General Data Protection Regulation (GDPR) (art. 13-15 and 22) as interpreted by the European Data Protection Board (EDPB) require the processors of personal data to provide "the rationale behind or the criteria relied on in reaching the decision," under certain circumstances, when a fully automated decision is made (EDPB Guidelines of 3 October 2017 on Automated individual decision-making and Profiling, p. 25; see also (Edwards & Veale, 2018; Wachter et al., 2017)). Consumer protection law imposes to online marketplaces to provide their consumers with "the main parameters determining ranking [...] and the relative importance of those parameters" (art. 6(a) of Directive 2011/83). The Online Platforms Regulation imposes very similar obligations to online intermediation services and search engines towards their professional users (art. 5 of Regulation 2019/1150).

Sectoral rules are also analyzed. For instance, financial regulators "may require the investment firm to provide [...] a description of the nature of its algorithmic trading strategies, details of the trading parameters or limits to which the system is subject, the key compliance and risk controls that it has in place [...]. The competent authority [...] may, at any time, request further information from an investment firm about its algorithmic trading and the systems used for that trading" (art. 17(2) of Directive 2014/65 on Markets in Financial Instruments).

For decisions adopted by public authorities, two stronger requirements are studied: motivation obligations for administrations and for judges (imposed by European Convention on Human Rights). For administrative decisions, all factual and legal grounds on which the decision is based should be provided. For judicial decisions, judges have in addition to answer the arguments made by the parties in the litigation.

The objectives of those explanation requirements are twofold: first, allowing the recipients of a decision to understand it and act accordingly; second, allowing the public authority, before which a decision is contested, to exercise a meaningful effective control on the legality of the decision (European Commission White Paper of 19 February 2020 on Artificial Intelligence, p. 14).

2. Legal Requirements and Machine Learning

As explained in the previous section, legal texts do not always clearly identify the focus of the requirements. In private decision making, we identified that the explainability of four levels of machine learning entities or concepts are mentioned in legal texts (Bibal et al., Forthcoming): the main features used for a decision, all features used for a decision, how the features are combined for reaching a decision and the whole model (see Table 1).

The first and weaker level of requirements is to provide the main features used for a decision. Note that the main parameters mentioned in the legal texts refer to the features used by a ML model. While the main features used are natively provided by interpretable models such as linear models and decision trees, some works go further and provide weakly and strongly relevant features in linear models (John et al., 1994; Kohavi & John, 1997). In the context of black-box models, the feature importance provided by the out-of-bag error of random forests can pass these requirements, as well as the feature importance provided through the perturbation of input feature values (Fisher et al., 2019).

The second level of requirements is to provide all features involved in a decision. While providing all features used is again natively proposed by interpretable models, this requirement can be difficult to achieve when the number of

*Equal contribution [1]PReCISE, Faculty of Computer Science, NADI, University of Namur, Belgium [2]CRIDS, Faculty of Law, NADI, University of Namur, Belgium. Correspondence to: Adrien Bibal <adrien.bibal@unaur.be>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020. Copyright 2020 by the author(s).

Main features
• Directive 2011/83 on Consumer Rights, art. 6(a): obligation to provide the "main parameters" and their "relative importance" • Regulation 2019/1150 on promoting fairness and transparency for business users of online intermediation services, art. 5: obligation to provide "the main parameters" and "the relative importance of those parameters"
All features
• Guidelines on automated individual decision-making and profiling: obligation to provide "the criteria relied on in reaching the decision" • Belgian law of 4 April 2014 on insurances, art. 46: obligation to provide "the segmentation criteria"
Combination of features
Guidelines on Automated individual decision-making and Profiling: obligation to provide "the rationale behind the decision"
Whole model
Directive 2014/65 on Markets in Financial Instruments, art. 17: obligation to provide "information [...] about its algorithmic trading and the systems used for that trading"

Table 1. Table reproduced from (Bibal et al., Forthcoming) containing the legal texts used as examples in this paper.

features used by the model is huge. Sparsity penalties such as Lasso may be necessary to satisfy the requirement.

The third level of explainability requirements is to provide the combination of features that led to a particular decision. Again, interpretable models make it possible to check how the features have been combined to lead to a decision. In the context of black-box models, techniques like LIME (Ribeiro et al., 2016) have been developed to get insights on how models behave locally, i.e. for a particular decision.

Finally, the strongest requirement is to provide the whole model. In this case of strong requirement, only interpretable models can be used, as, by definition, black-box models cannot be provided (e.g. if the model is non-parametric) or understood (e.g. in the case of neural networks).

In addition to these four levels of explainability requirements for private decisions, requirements for public decisions impose two additional constraints. For administrative decisions, the legal motivation should also be provided with the decision. This means that all factual and legal grounds on which the decision is based must be provided. In the case of judicial decisions, in addition to the facts of the case and the motivation, which was already needed for administrative decisions, answers to the arguments of the parties to the litigation must also be provided. While some works try to tackle these requirements (e.g. (Ashley & Brüninghaus, 2009) explain decisions with facts only; (Zhong et al., 2018) introduce multi-task learning for dealing with legal articles, as well as facts; and (Ye et al., 2018) use sequence-to-sequence learning to propose answers to the arguments of the parties), legal requirements on the explainability of public decisions remain a challenge in machine learning, because ML algorithms are not designed to manipulate factual and legal grounds, as well as arguments, directly.

In conclusion, we call for an interdisciplinary conversation between the legal and AI research communities. In particular, legal scholars could benefit from better understanding the potential and the limitations of ML models and AI scholars from better understanding the objectives and ambiguities of the law.

References

Ashley, K. D. and Brüninghaus, S. Automatically classifying case texts and predicting outcomes. *Artificial Intelligence and Law*, 17(2):125–165, 2009.

Bibal, A., Lognoul, M., de Streel, A., and Frénay, B. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, Forthcoming.

Edwards, L. and Veale, M. Enslaving the algorithm: From a "right to an explanation" to a "right to better decisions"? *IEEE Security & Privacy*, 16(3):46–54, 2018.

Fisher, A., Rudin, C., and Dominici, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *JMLR*, 20(177):1–81, 2019.

John, G. H., Kohavi, R., and Pfleger, K. Irrelevant features and the subset selection problem. In *Proceedings of ICML*, pp. 121–129, 1994.

Kohavi, R. and John, G. H. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

Ribeiro, M. T., Singh, S., and Guestrin, C. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of ACM SIGKDD*, pp. 1135–1144, 2016.

Wachter, S., Mittelstadt, B., and Floridi, L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017.

Ye, H., Jiang, X., Luo, Z., and Chao, W. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of NAACL*, pp. 1854–1864, 2018.

Zhong, H., Zhipeng, G., Tu, C., Xiao, C., Liu, Z., and Sun, M. Legal judgment prediction via topological learning. In *Proceedings of EMNLP*, pp. 3540–3549, 2018.