Proceedings of the 2nd eXplainable AI in LAw Workshop XAILA 2019

(http://xaila.geist.re)

Extended Preface

Michał Araszkiewicz, Grzegorz J. Nalepa, Martin Atzmueller, Paulo Novais

During the last few years a whole new area of investigation has emerged. First identified through the term of explainable AI (XAI), currently it discusses a broad set of interrelated concepts such as interpretability, transparency, responsibility or trustworthiness of AI[1]. This highly interdisciplinary field encompasses and attempts to integrate conceptual, ethical, legal and engineering perspectives and methods, The very notion of explanation has become intensively investigated to bring results revealing its multi-faceted nature. At the same time, the topics of trustworthiness, transparency and explainability of AI has become the subject of interest not only of the academia and business, but also of general public and of political bodies. In particular, on 8 April 2019 the High-Level Expert Group on AI presented Ethics Guidelines for Trustworthy Artificial Intelligence. It is expected that more guidelines and standards concerning the said topics will be developed in the near future. The development of the normative framework concerning XAI may eventually result in a binding legislative act. However, the creation of any regulative framework in the said area requires thorough analysis of the basic concepts, technical solutions and potential legal mechanism that might be used for the purpose of understanding of AI operations to particular groups of actors.

These considerations require, first and foremost, solid conceptual foundations. The discussion concerning explanations of AI has only a minor intersection with the philosophical debate on the notion of explanation. To recall, explanatory reasoning consists in forming hypotheses that remain in certain relations with sentences describing the results of the observation. There is no rigid boundary between the observation sentences and the theoretical sentences (including hypotheses). The relation of explanation is one of the most

1 For a recent in-depth survey on the topics of XAI see: Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*, Information Fusion, Volume 58, 2020, pp. 82-115.

controversial topics in the methodology of science. Generally speaking it is assumed that hypotheses should be falsifiable, that they should encompass causal relations and that they should fit into some coherent whole forming a scientific theory. However, unlike natural phenomena, intelligent systems are artifacts. Even if in many cases their structure is enormously complex, their general features are known, at least to their designers. Moreover, they are developed to perform certain tasks and are evaluated with regard to the level of this performance. Therefore, we do not actually have to discover how do they function on a general level.

The problem arises because numerous intelligent systems are based on the machine learning models that are not transparent. Transparency is a complex concept which encompasses the criteria of simulatability, decomposability and algorithmic transparency. Simulatability means that the operation of the system may be reproduced by a human. Decomposability is a possibility to characterize what particular elements are responsible for in the activity of the system. Finally, algorithmic transparency means that it is generally possible to clearly present how the systems' output is generated on the basis of the input. Obviously, many types of the learning models used in the different branches of industry do not satisfy any criterion of transparency to a satisfiable extent. This concerns not only the so-called deep models (mostly multi-layer neural networks) but also such models as Support Vector Machines or Random Forests. The operations of nontransparent systems may be attempted to be explained post hoc. Many XAI techniques have been developed to attain this goals, including the generators of textual explanations, visual explanations, explanations by example, feature relevance models or simplifications. The latter category, in fact, pertains to all types of nontransparent models explanators, while their main function is to describe the operation of the system through complexity reduction. We expect that the operations of the XAI model will be transparent, because otherwise we would enter into a regress of explanations.

Transparency is an important feature, because it directly contributes to model understandability. The latter concept, also referred to as intelligibility, is a gradual feature of a system which represents the possibility to grasp the function of the system by a human. However, it is important to keep in mind that what should be understood by a human is the explained model, and not its radically reduced representation.

Explainabiity may in turn be defined as a relative balance between understandability on the one hand and the accuracy of representation on the other hand. It should be thus emphasized that the notion of explainability is auditorium relative. Different auditoria will expect or require different levels of description accuracy and will also differ in their capacities

to process the explanation on a given level of complexity.

For some auditoria and some types of tasks, symbolic explanations will be required. This pertains in particular to the area of automated decision making where human rights and obligations may be affected by a decision. This is a particularly challenging area, because it is expected that a decision is supported by an appropriate reasons rather than simply extracted from the existing data. Practical normative reasoning is interested in what should be done rather than in what decisions have been made so far, even though in many situations the earlier decisions may be treated as adequate reasons to act similarly in the current state of affairs. However, practical reasoning is open in the sense that the existing practices may be questioned on normative grounds, shifts of preferences may be argued for, and entirely new propositions may be subjected to debate. This open character of practical reasoning is also characteristic for its important sub-area, legal reasoning.

The question arises, then, how the novel issues of AI understandability, comprehensibility, transparency and last but not least explainability should be absorbed by a necessarily open (in the sense described above) legal discourse. These considerations are particularly important from the point of view of accountability of the potentially responsible legal entities involved in the design, development, evaluation, exploitation and use of the intelligent systems. Accountability has become the standard criterion of assessment of the behavior of data protection controllers in the GDPR regime, but its significance is broader. The question arises in particular what features of an intelligent system should be emphasized in the design and how the development process should be prepared and documented to enable the potentially liable entity to become exculpated? Should the foreseeability of harm be used in the context be used in connection with liability ascription to the operator of an intelligent systems? A natural candidate for the standard used in this context is risk-based approach required by the GDPR in connection with data protection. This methodology may be considered to become generalized approach in the field of AI-related liability, however, it may be criticized because it adds complexity to the process. Rather than application of clear rules and standards, it requires a concrete assessment of risks and there is more than one methodology for the performance of such analysis. These considerations may lead to the conclusion that civil liability related to the AI related systems will eventually be based on risk. However, regulatory approach characteristic for the European regulation emphasizes compliance with objective standards, and not liability based on harm. Therefore, the issues of accountability become relevant again in the context of administrative liability.

It is reasonable to assume that we will need numerous standards of accountability of intelligent systems, taking into considerations not only the differences between the used

technological solutions, but the specificities of particular areas of their use as well. The operation of energy industry, transport, medical diagnostics, online marketing and, last but certainly not least, automated prediction of practical decisions, including judicial ones. The development of such standards is a complex challenge, and currently it is the time to consider what factors, interests, values and principles should be taken into consideration in the preliminary stage of the process of their formation.

In this context, the XAILA (eXplainable AI and Law) workshop was proposed two years ago in 2018. We believed, that it was the intersection of Law and AI that made the perfect choice to discuss the questions of XAI and their broader social context. Together, the work of legal specialists and AI engineers lays foundations and provides a conceptual framework for ethical concepts and values in AI systems. Therefore, when discussing social consequences and considerations of transparent and explainable AI systems, we should focus on the legal conceptual framework. A significant part of AI and Law research during the last two decades was devoted to operationalization of legal thinking with values. These results may now be reconsidered in a broader context, concerning the development of XAI systems and with their social impact. As such we realized it was a very timely issue for the AI and Law community to discuss together[2]. Therefore, our objective with XAILA has been to bring people from AI interested in XAI topics (possibly with broader background than just engineering) and create an ample space for discussion with people from the field of legal scholarship and/or legal practice.

The first edition of XAILA was organized at the JURIX 2018 conference in Groningen and was acclaimed as very successful both in terms of quality of papers and attendance. One year later, we held the second edition of the XAILA workshop on December 11 2019 at the 32nd International Conference on Legal Knowledge and Information Systems – JURIX 2019 (https://jurix2019.oeg-upm.net) in Madrid, Spain. The workshop was devoted to the discussion of the above mentioned and similar topics. The event attracted significant attendance (more than 30 participants) and 7 papers from which 5 papers were accepted in the comprehensive review process. Upon invitation from the organizers, María Jesús González-Espejo from the Instituto de Innovacion Legal kindly agreed to deliver an invited talk entitled *Drivers for Adopting Legal AI*. The remaining part of the volume presents revised versions of papers that were discussed during the workshop.

In their paper Francesco Sovrano, Fabio Vitali and Monica Palmirani discuss upon the difference between Explainable and Explaining, specifically on requirements and challenges

2 For a recent, and possibly the first book on AI dedicated to legal professionals see: María Jesús González-Espejo, Juan Pavón (Eds.), *An Introductory Guide to Artificial Intelligence for Legal Professionals*, Wolters Kluwer, 2020.

under the GDPR. Next, Grzegorz J. Nalepa, Michał Araszkiewicz, Sławomir Nowaczyk and Szymon Bobek present technical and legal perspectives for building trust into AI systems through explainability. After that, Ramon Ruiz-Dolz, José Alemany, Stella Heras and Ana Garcia-Fornes discuss the automatic generation of explanations to prevent privacy violations. Finally, Michal Klincewicz and Lily Frank focus on the healthcare domain, and tackle emerging ethical and legal issues in healthcare machine learning.

The editors wish to thank the organizers of Jurix 2019 as well as the members of the international program committee for their support of XAILA!

Editors:

Grzegorz J. Nalepa, Jagiellonian University, AGH University of Science and Technology

Martin Atzmueller, Tilburg University

Michał Araszkiewicz, Jagiellonian University

Paulo Novais, University of Minho Braga

Program Committee of XAILA 2019