Looking for the Right Paths to Use XAI in the Judiciary

Which Branches of Law Need Inherently Interpretable Machine Learning Models and Why?

Andrzej Porębski

Jagiellonian University, Gołębia 24, Kraków 31-007, Poland

Abstract

In a legal context, it is often particularly important to be able to trace the reasons why a decision was made; therefore, there may be an intuition that explainability is extremely important in judicial support systems. However, the standard of explainability (understandability, transparency) that machine learning technologies used to assist judges should meet has not yet been described. Defining this standard is even more complicated because, when considering it, it is necessary to take into account not only the specifics of the legal context in general but also of individual branches of law (for example, criminal or civil law). In this paper, I consider which branches of law, due to their specificity, seem to require the use of the most algorithmically transparent – and thus inherently interpretable – methods. Juxtaposing three general levels of explainability (white boxes, black boxes with post hoc explainers, and full black boxes) with legal values, I consider the paths that the development of machine learning models supporting judicial reasoning should follow in order to be tailored to the specifics of each legal field.

Keywords

AI & Law, right to a fair trail, XAI in the judiciary, inherent interpretability, decision support systems

1. Introduction

Machine learning came into use years ago in areas of life that can be best described as "critical contexts", that is, contexts in which a special role is attached to maintaining sufficiently high standards because of the need to ensure, among other things, the security or protection of individual rights. Machine learning, therefore, drives things like autonomous cars or various medical technologies. In these areas, the need to take into account the values of eXplainable Artificial Intelligence (XAI) [1] is rightly recognised [2, 3].

Another "critical context" in which machine learning methods can find application, and which is central to the presented paper, is in assisting the judiciary. Research on this topic clearly signals the possibility of using machine learning in supporting the operation of lawyers and the application of the law, for example, through risk prediction, prediction of likely court rulings, or warning of potential bias in judgements [4, 5]. Also, LLM models — including those designed for legal applications such as SaulLM-7B [6] — can hypothetically support lawyers' work on documents or related to jurisprudence. However, the application of the law by state

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

and.porebski@uj.edu.pl (A. Porębski)

https://www.researchgate.net/profile/Andrzej-Porebski-2 (A. Porębski)

0000-0003-0856-5500 (A. Porębski)

bodies, and in particular by the judiciary, is one of the particularly critical contexts of social life, especially as it often represents the last opportunity for an individual to obtain protection of his or her rights. This is why, for example, the COMPAS tool used in the United States to estimate the risk of recidivism — which operates in the absence of more precise regulation, although it may have had a real impact on the decisions and sentences handed down — has received far-reaching criticism in the prevailing literature [7, 8].

Machine learning, for all the potential disadvantages of ”outsourcing” legal inference to automated technologies developed by external parties, has one huge advantage: it can relieve the judicial system of a surfeit of repetitive, simple cases, thereby allowing judges to spend more time on those cases that require longer and deeper reflection. Because of this advantage, it is almost certain that over the coming decades, more European countries will introduce solutions based on machine learning (or, more broadly, artificial intelligence methods) into the judiciary, aiming to automate at least part of the judicial process of applying the law. However, in order for these implementations to take place without harming individual rights, it is necessary to first develop standards that should be met by machine learning-based technologies used in this context. These standards should ensure that algorithms assisting (or automating) judicial inferences do not jeopardise the key values that national legal systems (for example, state constitutions), acts of international law (for example, the European Convention on Human Rights or acts of EU law), and civil society require of the courts. Among the values that can be mentioned are the right to have a case decided by an independent and impartial court, the fairness of judicial proceedings, the appealability of decisions, and the right to public information. Due to these values, the law appears to be strongly linked to XAI, as noted by a number of researchers [4, 9, 10, 11]. Unfortunately, the subject of the requirements that lawyers and society should place on the technologies used within the judiciary is still very little explored.

This paper aims to address the topic of the standards that machine learning systems being used in the application of the law should meet in terms of the values associated with XAI, in particular, the explainability of systems. It further focuses on showing that the construction of this standard should take into account in which branch of law machine learning technologies would be used and contrasts the understandability of the selected classes of technologies with the axiology and functions of the two selected branches of law. For simplicity, a categorisation of machine learning technologies is used according to three levels of algorithmic transparency: black boxes, grey boxes (usually post hoc explained black boxes), and glass (white) boxes [12].

It is important to stress that in this work, I refer to a specific layer of explainability; ready-to-go systems. I do not, therefore, refer to separate chronologically previous levels — the posing of the problem that such systems are supposed to solve or the creation of the system (including data collection and model training). Properly regulating how the levels of problem posing and system creation should be disclosed or communicated is very important in order to achieve actual understandability (explainability) of the machine learning systems used in a legal context. However, these levels are largely independent of the class of technology on which the system was developed (that is, even if, for example, complex black box artificial neural networks were used, the process of creating such models can still be disclosed in detail, because the algorithmic black box does not prevent this). Hence, they constitute a field for separate discussion.

The research presented here is part of a larger research project aimed at setting a general standard of understandability (and related transparency and explainability) of machine learning-

based technologies used in the application of the law (see Acknowledgments). The project aims to determine to what extent the inherent interpretability of such technologies can be sacrificed in various legal applications of machine learning in the name of their effectiveness or efficiency.

2. Three Levels of Explainability — Three Paths of Regulation

The explainability of systems can, of course, be looked at in a great many ways [13], but for the purposes of this work and preliminary attempts to set a standard of explainability, I propose an approach relating to three relatively easily distinguishable levels of algorithmic transparency of (machine learning) models and associated regulatory paths. The proposed division is inspired by the work of Ali et al. [12], as well as the systematics included in the works [14, 15].

The first group of systems (I) are those based on machine learning techniques leading to inherently interpretable models[1] and uncomplicated enough to make model interpretation feasible in practice[2] (sometimes referred to as glass or white boxes). The fundamental advantage of such models is that the basis of the returned result can be accurately determined, provided the theoretical aspects of the method are known. Please note that, in order to narrow down the analysis, I refer in this paper to machine learning as the technology with the greatest scope of application in the legal context. However, there are also other directly interpretable classes of computational methods that could sometimes find application in law, in particular, rule-based or case-based methods, especially newer approaches implementing, for example, defeasible reasoning [16].

The second group of systems (II) are algorithmic black boxes for which adequate post hoc explanations (sometimes referred to as grey boxes) have been developed. Here, I consider black boxes to be algorithmically non-transparent models whose decision rules are not interpretable by humans, usually because they are overly complex due to either their ensembler nature, the complexity of the mathematical functions, or the far-reaching multidimensionality of the data. Typically, post hoc explanations, such as SHAP or LIME, do not allow full information about the basis of the results returned by the model, and the explanations can be unstable [17]. At the same time, properly constructed and validated explanations allow us to know at least approximately the characteristics of the inference made by the model.

The third group of systems (III) are algorithmic full black boxes, that means, those black box models where no explanation techniques have been used, either because explanations have not been created or because they cannot be created, for example, due to insufficient computing power. This group could also include those systems where post hoc explanations have been created but are known to be inadequate, that is, they do not capture most of the relationships present in the model.

[1]The group of such techniques is in fact not homogeneous, as what constitutes interpretability differs between, for example, linear regression and k-NN. This is an interesting topic for further research.

[2]The idea here is to exclude from the group of inherently interpretable models those cases which, although they have led to formally interpretable rules, are so elaborate that in practice they are incomprehensible to almost all people, even experts in machine learning. For example, a decision tree, classically regarded as one of the most algorithmically transparent methods requiring no additional explanation, will become practically uninterpretable if it has 10000 vertices.

The three groups indicated can be simply ordered in terms of the level of algorithmic transparency (level of explainability), let us call it T: $T(I) > T(II) > T(III)$. In this respect — assuming that the use of machine learning systems in judicial application of the law is allowed at all — it is, therefore, possible to distinguish at least three regulatory approaches: minimalist, deeming even the lowest $T(III)$ level sufficient; intermediate, requiring the $T(II)$ level; and restrictive, expecting the $T(I)$ level. These three routes of regulation are the starting point for the analysis that follows.

Before going on to characterise the two branches of law, I would like to draw attention to the important overall role of XAI-related values in the legal context. First and foremost, the law is based on transparency in the sense that the motives behind decisions should be knowable. Even if it is not possible to trace the exact neurobiological processes that led a person to take a certain decision, in a legal context it is expected to be possible to obtain at least the declared motives for the decision. Parties are served in many procedures by the right to a statement of reasons, which should firstly fulfil an informative function and secondly, and more importantly, enable the decision to be assessed by other courts, including higher instance courts. For all these reasons, and others not elaborated here, it is difficult to imagine the use in any judicial process of a technology with negligible transparency (that means, only its input and output would be interpretable). Thus, $T(III)$ referring to full black boxes, should be considered generally inadmissible in the judicial application of the law. Consequently, there should be a strong preference in the legal context for solutions created by XAI.

3. The Standard of Explainability in Criminal vs Civil Law

3.1. Overview of Criminal and Civil Law Axiology

I am consciously conducting the considerations in this section at a high level of generality. The axiological differences between civil and criminal law (discussed further below) are typical of much of the liberal democratic states, although the precise characteristics of the legal systems differ between them.

Criminal law is aimed at designating a certain set of behaviours (usually considered particularly reprehensible by society) that are prohibited under the threat of severe sanctions. The criminal process aims to punish those guilty of a crime and to acquit those who did not commit the crime. The effects that criminal court judgements can have are among the most severe for citizens. Only criminal convictions can legally imprison a person (and, in inhumane systems, also deprive them of their life). In addition, a criminal conviction can have severe consequences, for example, financial, but also related to one's career, among other things. At the same time, a conviction can trigger significant social sanctions such as a loss of reputation for the convicted person. However, criminal law has another extremely important function: the guarantee function [18, 19]. The guarantee function of criminal law (and criminal procedure) aims to protect citizens from abuse of the apparatus of state power and coercion that would unjustifiably interfere with their freedoms. The guarantee function can be fulfilled in the criminal law's precise delimitation of the set of penalised behaviours, strict rules for the interpretation of criminal laws, the guarantee of the right to defence, the principle of the presumption of innocence, or the possibility of conviction for a crime only by the courts. One should be aware

that in the case of a criminal trial, the individual stands against the state apparatus, which creates the risk of an undue advantage on the part of the prosecution. In the branch of law under consideration, speed of proceedings is not as important a value as preserving all the rights of the parties, allowing the accused to defend themselves, and — ultimately — correctly identifying the guilty or acquitting the wrongly accused.

Civil law[3] can be contrasted with criminal law, which regulates, inter alia, the sphere of contracts or liability for damages arising from non-criminal acts. Civil law normalizes and unifies legal relations in society and also prohibits certain behaviour (or contractual provisions), but a breach of such law may result in, at most, pecuniary liability. Civil law usually protects the individual interest rather than the public interest, which is protected by criminal law. The purpose of a civil trial is usually to resolve a dispute that has arisen (for example, to declare an obligation to pay), not to impose a punishment. If the parties succeed in resolving the dispute by other means (for example, through a settlement reached in mediation), a civil court trial becomes unnecessary. On the contrary, a criminal trial will usually continue even after the victim has forgiven the perpetrator for the incident. Although mistakes in a civil trial can be severe and are obviously not desirable, they usually create a much lower risk of a serious violation of civil rights compared to criminal trials. The outlined characteristics of civil law make it more flexible; one of the key values in this branch of law is the efficiency of proceedings, while the procedural guarantees of the parties, however present, are narrower.

The indicated axiology of the two fields of law clearly influences the admissibility of various solutions, including technological ones, in the courts. In Poland, for example, online hearings have been broadly permissible in civil trials since 2020. Meanwhile, in criminal trials, the possibility of remote participation is significantly limited. This difference is precisely due to the recognition that the increase in convenience for parties and lawyers and the speed of proceedings provided by the possibility of online hearings is a sufficient justification for the use of this technology in the case of civil procedure, but in view of the losses in the quality of communication that could hinder the pursuit of the truth, it cannot be allowed in criminal procedure. Similarly, axiological differences must be taken into account when setting the requirements that should be met by machine learning systems designed to assist judges in the application of the law.

3.2. Criminal Law as a Context Requiring a Higher Standard of Explainability

In this paper, I would like to identify criminal law as the area of law in which there should be the highest standard of algorithmic transparency for assistive machine learning technologies, which will allow the predictions they make to be directly explainable (interpretable). The guarantee function of criminal law could be insufficiently realised if any component with only partial understanding (for example, grey box models) is introduced into the criminal process. In the case of criminal justice, approximate rather than precise explanations appear insufficiently transparent and create an additional level of risk to the rights of the individual related to the uncertainty of their situation. On the other hand, grey boxes could lead to emerging doubts about the adequacy of prediction in a particular case. When it comes to criminal law, doubts

[3]Of course, in this text I refer to civil law in the sense of a larger branch of law and not a type of system of law.

arising from the uncertain operation of technology pose a significant problem because, according to the typically accepted criminal trial principle *in dubio pro reo*, irremovable doubts should be resolved in favour of the accused. In extreme cases, the use of insufficiently transparent technology to support a judge's decisions in a criminal trial could, therefore, lead to those decisions being undermined by raising doubts about the technology. If it could be possible to signal that these doubts are "reasonable" — I refer here to the frequently adopted "beyond a reasonable doubt" standard of proof [20] — this could even theoretically lead to acquittals based on flaws in the machine learning technology being used. For the aforementioned reasons, I propose that only inherently interpretable models should be used in criminal law, which is in line with the views presented by Rudin [21], among others.

4. The Two-Tier Structure of the Explainability Standard

The observations so far have been directed at indicating that, depending on the branch of law in which a machine learning system would be used, the requirements in terms of its explainability should be modified. The axiological differences between the branches are so significant that a single standard for each branch could lead to an unjustified shaping of the set of technologies acceptable for the given applications. However, the discussed differences between the branches of law do not mean that some part of the resulting standard is not common. Undoubtedly, the very nature of the institution of the courts, the right to a court and a fair trial, and other general values, mean that in every case of a judicial process, certain technological solutions should be unacceptable. Because of the issues discussed in Section 2, the occurrence of full black boxes (lacking even an element of post hoc explanation) can be identified as generally unacceptable for all branches of law application.

Since some requirements should be common to the judicial application of the law in general, and some adapted depending on the specific field of law, the standard of explainability (or algorithmic transparency) of the technological solutions in question can be viewed as two-tier, consisting of (I) *sine qua non* conditions, which are certain fundamental conditions, and (II) an element of increasing requirements, depending on the specificity of the subfield. This two-tier structure seems to have the potential to apply to other critical contexts as well, for example, to medicine, where there is (I) a certain general level of patient rights (which translates into the *sine qua non* conditions of physician support technology), but which is (II) realised in a different specific way in its individual subfields (for example, family medicine vs oncology). The structure of the two-tier standard is presented using the example of the judiciary support technology in Figure 1.

The main advantage of the standard presented is the systematization of the discussion on the requirements relating to XAI. Its application makes it possible to conduct the debate on the requirements at two separate levels. According to the standard, it is appropriate to first distinguish a general level of transparency (explainability, understandability) that must be guaranteed in a certain field (for example, application of law, medicine, or finances) and then to find those subfields that are distinctive from an axiological point of view and therefore should have a higher level of the standard. This could help to address problems with the heterogeneity of XAI-related standards in different branches of law (noticed in the literature [9]).

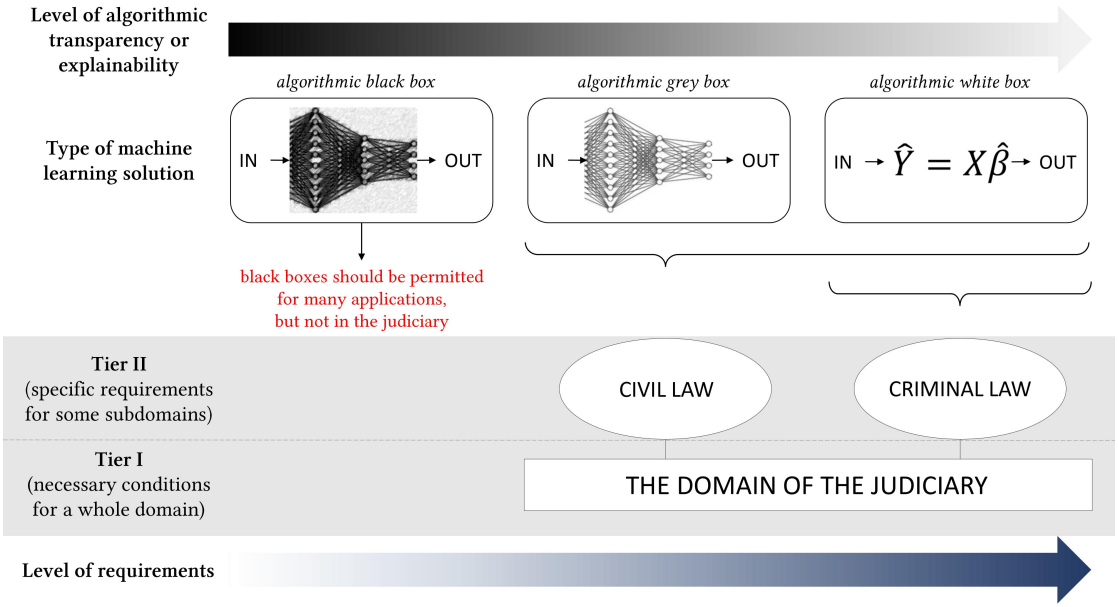

Figure 1: The relationship between applications in the judiciary (*Tier I*) and, particularly, in civil and criminal law (*Tier II*) and the different levels of algorithmic transparency. Source: own elaboration.

Acknowledgments

This research was funded by the National Science Center, Poland, and is the result of the research project "The Understandability Requirement of Machine Learning Systems Used in the Application of Law" (no. 2022/45/N/HS5/00871).

References

[1] L. Longo, M. Brcic, F. Cabitza, et al., Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions, Information Fusion 106 (2024) 102301. doi:10.1016/j.inffus.2024.102301.

[2] C. Combi, B. Amico, R. Bellazzi, et al., A manifesto on explainability for artificial intelligence in medicine, Artificial Intelligence in Medicine 133 (2022) 102423. doi:10.1016/j.artmed.2022.102423.

[3] S. Umbrello, R. Yampolskiy, Designing AI for explainability and verifiability: A value sensitive design approach to avoid artificial stupidity in autonomous vehicles, International Journal of Social Robotics 14 (2022) 313–322. doi:10.1007/s12369-021-00790-w.

[4] A. Porębski, Machine learning and law, in: B. Brożek, O. Kanevskaia, P. Pałka (Eds.), Research Handbook on Law and Technology, Edward Elgar Publishing, Cheltenham, UK, 2023, pp. 450–467. doi:10.4337/9781803921327.00037.

[5] R. Berk, J. Hyatt, Machine Learning Forecasts of Risk to Inform Sentencing Decisions, Federal Sentencing Reporter 27 (2015) 222–228. doi:10.1525/fsr.2015.27.4.222.

[6] P. Colombo, T. P. Pires, M. Boudiaf, et al., SaulLM-7B: A pioneering large language model for law, 2024. arXiv:2403.03883.
[7] C. Rudin, C. Wang, B. Coker, The age of secrecy and unfairness in recidivism prediction, Harvard Data Science Review 2 (2020). doi:10.1162/99608f92.6ed64b30, https://hdsr.mit-press.mit.edu/pub/7z10o269.
[8] J. Dressel, H. Farid, The accuracy, fairness, and limits of predicting recidivism, Science Advances 4 (2018) eaao5580. doi:10.1126/sciadv.aao5580.
[9] K. M. Richmond, S. M. Muddamsetty, T. Gammeltoft-Hansen, et al., Explainable AI and law: an evidential survey, Digital Society 3 (2024) 1. doi:10.1007/s44206-023-00081-z.
[10] Z. Ződi, Algorithmic explainability and legal reasoning, The Theory and Practice of Legislation 10 (2022) 67–92. doi:10.1080/20508840.2022.2033945.
[11] A. Bibal, M. Lognoul, A. De Streel, B. Frénay, Legal requirements on explainability in machine learning, Artificial Intelligence and Law 29 (2021) 149–169. doi:10.1007/s10506-020-09270-4.
[12] S. Ali, T. Abuhmed, S. El-Sappagh, et al., Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence, Information Fusion 99 (2023) 101805. doi:10.1016/j.inffus.2023.101805.
[13] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Information Fusion 76 (2021) 89–106. doi:10.1016/j.inffus.2021.05.009.
[14] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115. doi:10.1016/j.inffus.2019.12.012.
[15] S. Maksymiuk, A. Gosiewska, P. Biecek, Landscape of R packages for explainable artificial intelligence, 2021. arXiv:2009.13248.
[16] L. Rizzo, L. Longo, A qualitative investigation of the degree of explainability of defeasible argumentation and non-monotonic fuzzy reasoning, in: Proc. of 26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science, 2018, pp. 138–149. doi:10.21427/tby8-8z04.
[17] C. Burger, L. Chen, T. Le, "Are your explanations reliable?" Investigating the stability of LIME in explaining text classifiers by marrying XAI and adversarial attack, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proc. of Conference on Empirical Methods in Natural Language Processing, 2023, Association for Computational Linguistics, Singapore, 2023, pp. 12831–12844. doi:10.18653/v1/2023.emnlp-main.792.
[18] E. Maculan, A. Gil Gil, The rationale and purposes of criminal law and punishment in transitional contexts, Oxford Journal of Legal Studies 40 (2020) 132–157. doi:10.1093/ojls/gqz033.
[19] A. Cornford, The aims and functions of criminal law, The Modern Law Review 87 (2024) 398–429. doi:10.1111/1468-2230.12846.
[20] K. M. Clermont, E. Sherwin, A Comparative View of Standards of Proof, The American Journal of Comparative Law 50 (2002) 243–275. doi:10.2307/840821.
[21] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215. doi:10.1038/s42256-019-0048-x.