UDC 34:004.8:340.12
DOI https://doi.org/10.24144/2307-3322.2025.90.3.63

THE «BLACK BOX» PROBLEM IN LEGAL AI: ARE EXISTING XAI METHODS SUFFICIENT TO MEET DUE PROCESS REQUIREMENTS?

Shamov O.A.,
Intelligent systems researcher,
head of Human Rights Educational Guild
ORCID: 0009-0009-5001-0526
e-mail: yursprava@gmail.com

Shamov O.A. The «Black box» problem in legal AI: Are existing XAI methods sufficient to meet due process requirements?

The integration of artificial intelligence (AI) in legal practice creates the «black box» problem, where opaque models challenge the principles of due process. This opacity conflicts with the right to a reasoned and contestable decision.

This article investigates the fundamental gap between the technical explanations provided by modern XAI methods (like LIME, SHAP) and the normative justifications required by the doctrine of due process. The goal is to critically evaluate XAI's limitations and, based on these findings, to formulate a new standard for transparency in legal AI.

The study is based on a comprehensive application of legal and scientific methods, including comparative-legal and formal-logical analysis. This approach was used to examine the interplay between AI technology and legal doctrines, analyzing legal acts and scientific works to identify inconsistencies between technical explanations and legal requirements.

The article demonstrates that popular post-hoc XAI methods are fundamentally insufficient for legal needs. They generate unstable technical approximations of a model's behavior, not legally meaningful justifications. These explanations can be unreliable and create a "Rashomon effect" (multiple contradictory explanations for one outcome), making the right to an effective appeal illusory. The law requires justification, not just explanation.

The article proposes a new standard: «Justifiable AI» (JAI). This concept shifts focus from explaining opaque models to designing inherently interpretable hybrid systems that combine rule-based components with data-driven models. Future research should focus on developing JAI architectures, creating legal certification standards, and studying user trust in such systems.

Key words: artificial intelligence, due process, Explainable AI (XAI), black box problem, legal justification.

Шамов О.А. Проблема «Чорної скриньки» в юридичному ШІ: чи достатньо існуючих методів XAI для задоволення вимог належного правового процесу?

Інтеграція штучного інтелекту (ШІ) в юридичну практику створює проблему "чорної скриньки", де непрозорі моделі ставлять під сумнів принципи належної правової процедури. Ця непрозорість суперечить праву на обґрунтоване та оскаржуване рішення.

Стаття досліджує фундаментальний розрив між технічними поясненнями, які надають сучасні методи XAI (LIME, SHAP), та нормативними обґрунтуваннями, що вимагаються доктриною належного правового процесу. Метою є критична оцінка обмежень XAI та, на основі цих висновків, формулювання нового стандарту прозорості для юридичного ШІ.

Дослідження базується на комплексному застосуванні правових та наукових методів, включаючи порівняльно-правовий та формально-логічний аналіз. Цей підхід використовувався для вивчення взаємозв'язку між технологією ШІ та правовими доктринами, аналізуючи правові акти та наукові праці для виявлення розбіжностей між технічними поясненнями та юридичними вимогами.

У статті доведено, що популярні пост-хок методи XAI є принципово недостатніми для юридичних потреб. Вони генерують нестабільні технічні апроксимації поведінки моделі, а не юридично значущі обгрунтування. Такі пояснення можуть бути ненадійними та створювати «ефект Расьомона» (кілька суперечливих пояснень для одного результату), що робить право на ефективну апеляцію ілюзорним. Право вимагає обгрунтування, а не лише пояснення.

У статті пропонується новий стандарт: «Обгрунтований ШІ» (JAI). Ця концепція зміщує фокус з пояснення непрозорих моделей на проєктування інтерпретованих гібридних систем, що поєднують компоненти на основі правил з моделями на основі даних. Подальші дослідження мають зосередитися на розробці архітектур JAI, створенні правових стандартів сертифікації та вивченні довіри користувачів до таких систем.

Ключові слова: штучний інтелект, належна правова процедура, пояснюваний ШІ (XAI), проблема «чорної скриньки», юридичне обгрунтування.

Introduction. In recent decades, artificial intelligence (AI) has transformed from a theoretical concept into a practical tool that is reshaping key areas of public life, including the justice system. Algorithmic systems are now used for crime prediction, recidivism risk assessment (such as the infamous COMPAS system in the US), analysis of evidence, and even to assist in judicial decision-making [5]. The potential benefits are clear: increased efficiency, reduced costs, and, possibly, a reduction in human bias. However, alongside these prospects, a fundamental problem arises that threatens the foundations of the rule of law, the «black box» problem.

The «black box» is a metaphor describing an AI system whose internal logic is so complex (especially in the case of deep neural networks) that it becomes opaque and incomprehensible even to its creators. When such a system makes a legally significant decision for example, recommending denial of parole a legal vacuum is created. The individual whose fate is decided by the algorithm is deprived of the fundamental right to know the basis on which the decision was made. This directly contradicts the principles of due process, which require transparency, reasoned motivation, and the possibility of an effective appeal In response to this challenge, the tech community has developed the field of Explainable AI (XAI), which aims to create tools for interpreting the decisions of opaque models. However, as researchers note, there is a tension between what is considered an «explanation» in computer science and what constitutes a «justification» in law [1]. A technical explanation may point to correlations in the data, but it does not provide the normative reasoning required by law. This article aims to investigate this tension by analyzing whether modern XAI methods are sufficient to meet the constitutional requirements of due process.

The objectives are to critically evaluate popular XAI tools, analyze their limitations in the legal field, and develop a proposal for a new, legally-oriented standard of transparency.

Analysis of Recent Research and Publications. The problem of algorithmic transparency in justice is the subject of active interdisciplinary debate. A key contribution to the critique of post-hoc explainability has been made by Cynthia Rudin, who argues that instead of trying to explain «black boxes,» we should create inherently interpretable models from the outset. She emphasizes that post-hoc explanations are often not faithful to the model's true logic and can be misleading, which is unacceptable in high-stakes fields like medicine or justice [2].

The theoretical understanding of the problem is largely shaped by the works of Frank Pasquale and Mireille Hildebrandt. In his monograph «The Black Box Society,» Pasquale analyzes how corporate and state secrecy regarding algorithms undermines accountability and democratic control, creating new forms of power without responsibility [6], which is a manifestation of the broader phenomenon of «algorithmic regulation» [7]. Hildebrandt, in turn, explores the concept of «Law as Computation,» pointing to the fundamental incompatibility between the statistical logic of modern AI and the normative-argumentative nature of the rule of law. She stresses that law is not just about prediction but also about justification based on legal norms, which data-driven models cannot provide [8].

Research also focuses on the empirical perception of fairness. People tend to view algorithmic decisions as less fair than similar human decisions, even when the outcomes are identical, indicating a deep distrust of automated systems in justice.

Review works, such as «Explainable AI and Law: An Evidential Survey,» systematize existing approaches and challenges, classifying XAI methods and their applicability in the legal domain. The authors of this review emphasize that most existing methods do not account for the specific requirements of legal reasoning, such as the need for causality rather than mere correlation [1].

Despite the significant volume of research, a key part of the problem remains unresolved: the lack of a coherent standard and a practical framework that would bridge the gap between technological capabilities and legal requirements.

There is a disconnect between how AI developers understand «explanation» and what lawyers understand by «justification»[10]. This article is dedicated to this unresolved issue the search for ways to create legally meaningful transparency that would satisfy the requirements of due process.

Materials and Methods. This study is theoretical-legal and interdisciplinary, which determined the choice of its methodological basis. The primary materials for analysis were: 1) legal acts establishing the principles of due process and the right to a fair trial (the Constitution of Ukraine, the European Convention on Human Rights) [4]; 2) scientific publications from legal, computer, and ethical sciences, indexed in international scientometric databases (Scopus, Web of Science), as well as posted in open academic archives; 3) technical and reference literature describing the principles of operation and limitations of Explainable AI (XAI) methods.

In the first stage, an analysis of the doctrine of due process was conducted to identify its key components relevant to algorithmic decision-making: the right to be heard, the right to a reasoned decision, and the right to an effective appeal.

In the second stage, using formal-logical and system-structural methods, the most common XAI techniques were analyzed, particularly LIMB (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Additive exPlanations), and counterfactual explanations. The analysis aimed to identify their fundamental limitations when applied in a legal context, such as problems of fidelity, stability, and the multiplicity of explanations.

In the third stage, using comparative-legal and dialectical methods, the requirements of the legal system were contrasted with the capabilities of technological solutions. This allowed for the identification of the fundamental conflict between the law's need for normative justification and XAI's ability to provide only technical, correlational explanations.

The synthesis of the obtained data formed the basis for formulating the author's hypothesis and proposal for the development of a new standard of «Justifiable AI,» which would shift the focus from post-hoc explanations to the design of inherently transparent and legally auditable systems.

Results and Discussion. The central result of the research is the establishment of the fact that existing XAI methods are fundamentally inadequate to meet the requirements of due process due to an ontological gap between the concepts of «explanation» and «justification.»

1. The Legal Requirement: Justification, not Explanation. Due process of law requires that a decision affecting an individual's rights be motivated. Legal motivation is not just a description of how a result was reached but a normative act that connects established facts with specific legal norms. For example, a court decision is justified not by the «judge's feeling of certainty of guilt,» but by the fact that «on the basis of evidence A, B, C, recognized as proper and admissible, the court has established facts X, Y, Z, which correspond to the disposition of an article of the Criminal Code of Ukraine, which formed the basis for the verdict.» This justification is public, stable, and can be appealed point by point: one can challenge the evidence, the established facts, or the correctness of the application of the norm.

2. The Technical Proposal: LIME, SHAP Explanations and their Shortcomings. XAI methods like LIME and SHAP do not provide such justification. They operate as post-hoc tools that attempt to approximate the behavior of a «black box» around a specific decision, LIME, for example, creates a simple, locally interpretable model (e.g., a linear regression) that mimics the behavior of the complex model for only one specific case. The explanation might sound like this: «In this case, the decision for a high risk of recidivism was most influenced by the factors 'number of prior arrests' and 'age at first offense'.»

The problem is that this explanation is only an approximation. As Cynthia Rudin proves, it may not be faithful to the true logic of the model [2]. Furthermore, it is unstable: a minor change in the input data can drastically change the explanation without changing the result itself. Most dangerous is the «Rashomon effect», described in reference literature: for the same complex model, one can find many different but equally accurate approximating simple models [9]. This means that one can generate several different explanations for the same decision, which makes the appeal process meaningless. Which of the explanations should be trusted?

Counterfactual explanations («the decision would have been different if your income were 10% higher') also have flaws [3]. They may point to plausible but not necessarily legally relevant or causally justified changes.

3. The Theoretical Gap and the Proposal for a New Standard. This gap between the legal requirement and the technical proposal is aptly described by Mireille Hildebrandt, who speaks of a shift from «rule-based knowledge» to «data-based experience» [8]. The legal system is based on the former, while modern AI is based on the latter. Trying to «explain» the second with the first is a palliative.

Therefore, instead of trying to impose legal logic on a fundamentally a-normative system, it is proposed to change the design approach itself. It is necessary to move from Explainable AI (XAI) to Justifiable AI (JAI). This standard should be based on the principle of «transparency by design» and require that legal AI systems be inherently interpretable.

Practically, this can be implemented through hybrid models. For example, a system could consist of two parts:

- A rule-based component: This part contains formalized legal norms, standards of proof, and procedural rules. It is completely transparent and auditable.
- A data driven component: This part uses machine learning to identify patterns in evidence, but not to make the final verdict, but to form «fact candidates.»

The decision-making process in such a JAI system would look like this: the machine learning model analyzes the data and passes the detected facts (e.g., «the object in the video is 95% likely to be a weapon») to the rule-based component. This component then applies legal norms to these facts and generates a final decision along with a step-by-step, logical, and legally justified trace of its reasoning. Such a justification can be verified and appealed according to the same standards as a human lawyer's decision.

Conclusions. The «black box» problem in legal AI is not just a technical but a fundamental legal challenge that threatens the principles of due process, particularly the right to a reasoned decision and its effective appeal. Existing methods of Explainable AI (XAI), such as LIME, SHAP, and counterfactual explanations, are insufficient to solve this problem. They provide technical, post-hoc, correlational explanations of a model's behavior, not the normative, stable, and legally meaningful justifications required by law. Their propensity for infidelity, instability, and multiplicity of explanations makes them unsuitable for use in high-stakes legal contexts. To overcome this gap, a paradigm shift is necessary from attempts to «explain» opaque systems to the creation of inherently «justifiable» systems. The author proposes the concept of a new standard Justifiable AI (JAI). This standard requires the design of hybrid systems that integrate transparent, rule-based, logical components with data-driven models. The goal of JAI is to ensure that every decision is accompanied by a complete, auditable, and legally contestable reasoning trace that complies with traditional standards of legal argumentation.

Prospects for further research lie in several directions. First, the development of specific technical architectures for JAI systems and their pilot implementation in less risky legal areas is necessary. Second, legal and technical standards for the certification and audit of such systems for «legal justifiability» should be developed. Third, it is important to conduct empirical research on the trust and interaction of legal professionals (judges, lawyers) with JAI systems to ensure their effective and ethical integration into the practice of justice.

REFERENCES:

1. Richmond, K. M., Muddamsetty, S. M., Gammeltoft-Hansen, T., Olsen, H. P., & Moeslund, T.B. Explainable AI and Law: An Evidential Survey. *Digital Society*. 2024. 3(1). URL: https://doi.org/10.1007/s44206-023-00081-z.
2. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell.* 2019. No 1. pp. 206–215. URL: https://doi.org/10.1038/s42256-019-0048-x.
3. Wachter, S., Mittelstadt, B., & Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*. 2018. 31(2). pp. 841-887. URL: https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf.
4. The Council of Europe. The European Convention on Human Rights. *The Council of Europe Website*. 1953. URL: https://www.coe.int/en/web/portal.
5. Završnik, A. Criminal justice, artificial intelligence systems, and human rights. *ERA Forum.* 2020. No 20. pp. 567–583. URL: https://doi.org/10.1007/s12027-020-00602-0.

6. Pasquale, F. The Black Box Society: The Secret Algorithms That Control Money and Information. *Cambridge: Harvard University Press*. 2015. 260P. URL: https://www.jstor.org/stable/j.ctt13x0hch.
7. Yeung, K. Algorithmic regulation: A critical interrogation. *King's College London Law School Research Paper*. 2017. No. 2017-27. URL: https://papers.ssrn.com/sol3/papers.cfm? abstract_id=2972505.
8. Hildebrandt, M. Law as computation in the era of artificial legal intelligence: Speaking law to the power of statistics. *University of Toronto Law Journal*. 2018. Vol. 68. No 1. URL: https://doi.org/10.3138/utlj.2017-0044.
9. Molnar, C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2022. URL: https://christophm.github.io/interpretable-ml-book/.
10. Selbst, A. & Barocas, S. The Intuitive Appeal of Explainable Machines. *Fordham Law Review.* 2018. 87/1085. URL: http://dx.doi.org/10.2139/ssrn.3126971.