

EXPLAINABLE AI: BRIDGING THE GAP BETWEEN HUMANS AND MACHINE

Vinnytsia National Technical University

Анотація

Оскільки системи штучного інтелекту (ШІ) відіграють дедалі значущішу роль у суспільстві, природа складних моделей як «чорний ящик» створила критичну потребу в прозорості. Пояснений ШІ слугує сполучною ланкою між машинною логікою та людським розумінням. У цій статті досліджуються фундаментальні принципи поясненого штучного інтелекту, а методи класифікуються на вбудовані та модельно-агностичні. Вона висвітлює технічний компроміс між точністю прогнозування та інтерпретованістю, водночас розглядаючи етичну необхідність довіри та справедливості.

Ключові слова: пояснений штучний інтелект, моделі чорного ящика, моделі білого ящика, транспарентність, інтерпретованість, людинноорієнтований ШІ, алгоритмічна неупередженість.

Abstract

As Artificial Intelligence (AI) systems take on more significant roles in society [1], the "black box" nature of complex models has created a critical need for transparency. Explainable AI (XAI) serves as the bridge to close the gap between machine logic and human understanding. This paper explores the fundamental principles of XAI, categorizing methods into intrinsic and model-agnostic approaches. It highlights the technical trade-off between predictive accuracy and interpretability while addressing the ethical necessity of trust and fairness.

Keywords: explainable AI (XAI), black-box models, white-box models, transparency, interpretability, user-centric AI, algorithmic fairness.

Introduction

A "black box" is defined as any system where the internal workings are hidden from or ignored by an observer, who instead focuses only on what goes in (inputs) and what comes out (outputs). In contrast, white-box models provide results that are understandable to domain experts. As AI technology advances, the demand for transparency has surged, particularly in crucial sectors like healthcare, finance and transportation, where decisions must be justified and understandable to human stakeholders [2].

Explainable AI (XAI) follows three core principles: transparency, interpretability, and explainability. While interpretability is typically associated with the inherent design and structure of a model (allowing users to see how features directly influence outputs), explainability encompasses a broader range of approaches and additional techniques used to articulate the behavior of complex systems like neural networks. The ultimate goal of these systems is to increase user trust and address concerns regarding a lack of fairness or discriminatory effects [7].

Results of a survey

Early research quantified explainability based on how understandable model decisions are to humans, noting that complex models often must sacrifice interpretability for accuracy. This led to the categorization of methods into two main types: intrinsic and model-agnostic. Intrinsic models, such as decision trees and linear regression, integrate interpretability directly into their architecture. Decision trees provide a series of if-then rules that users can easily follow, while linear regression offers a straightforward relationship between inputs and outputs that prioritizes clarity. However, while these models are transparent, they often lack the predictive power required for high-dimensional data found in modern AI applications.

To solve the limitations of opaque models, researchers developed model-agnostic techniques that do not rely on the internal workings of specific algorithms. Two main tools are LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations). LIME, introduced in 2016, works by approximating a complex model with a simpler, interpretable one in a local area to explain a specific decision. SHAP, introduced in 2017 and based on game theory, assigns a mathematically consistent

contribution value to every feature regarding a model's prediction [4]. These methods have found heavy use in finance and medicine because they allow the use of powerful deep learning models without losing the ability to justify individual results. Specialized tools like Layer-wise Relevance Propagation (LRP) and Grad-CAM have been developed to demystify deep learning by producing heatmaps that highlight which parts of an input were most influential to the machine [6].

A persistent finding throughout the survey is the "Accuracy-Interpretability Trade-off," where the most accurate models (like deep neural networks) tend to be the least transparent [5]. Most tools provide generic mathematical outputs that do not account for the diverse backgrounds, experience, or cognitive capabilities of the end-users. For example, a data scientist may find a SHAP value useful, but a doctor or a bank customer may find it unintelligible [4]. This has led to the emergence of "Causability," which argues that an explanation is only effective if it is comprehended and regarded as credible by the specific person receiving it. Human accuracy on CLEVR-X explanations is impressive, hovering around 92%. However, this isn't a baseline for all tasks. When complexity increases, the ability to parse these explanations diminishes rapidly [3].

Conclusion

Explainable AI is a critical component in the evolution of intelligent systems, serving to foster trust and understanding between humans and machines. However, the path forward requires more than just better algorithms; it requires a concerted effort to balance accuracy with explainability while addressing ethical considerations and regulatory standards [7]. Future research must shift from generic explanations to personalized frameworks that respect the diverse needs of human stakeholders [5]. Only by making AI transparent and fair can we ensure its responsible use in society.

REFERENCES

1. Understanding Explainable AI: Bridging the Gap Between Humans and Machines URL: <https://www.oreateai.com/blog/understanding-explainable-ai-bridging-the-gap-between-humans-and-machines/3f77101c895fe239b8222eced27b2a0a>.
2. Demystifying Explainable AI: Bridging the Gap Between Humans and Machines URL: <https://www.linkedin.com/pulse/demystifying-explainable-ai-bridging-gap-between-kanishka-venishetty-9onje/>.
3. Explainable AI: Bridging the Gap Between Machine Intelligence and Human Understanding URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5040613.
4. Explainable AI Bridging the Gap Between Machine Learning Models and Human Interpretability URL: https://www.researchgate.net/publication/389987305_Explainable_AI_Bridging_the_Gap_Between_Machine_Learning_Models_and_Human_Interpretability.
5. How Explainable AI is Bridging the Gap Between Humans and Algorithms URL: <https://medium.com/@preetikapuria587/how-explainable-ai-is-bridging-the-gap-between-humans-and-algorithms-1de2342c3ba3>.
6. Explainable AI: the new 42? URL: https://www.academia.edu/55602958/Explainable_AI_The_New_42?auto=download.
7. xxAI – Beyond Explainable AI. URL: https://www.researchgate.net/publication/36003662_xxAI_-_Beyond_Explainable_Artificial_Intelligence.

Затолокіна Софія Вадимівна – студентка групи СА-25б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м. Вінниця. E-mail: zatolokinasofia74@gmail.com.

Науковий керівник: Крютченко Олена Олександрівна, викладач англійської мови, кафедра іноземних мов, Вінницький національний технічний університет. E-mail: olena.kriut@gmail.com.

Zatolokina Sofia – a student of Faculty of Intelligent Information Technologies and Automation, SA-25b group, Vinnytsia National Technical University, Vinnytsia. E-mail: zatolokinasofia74@gmail.com

Scientific supervisor Kriutchenko Olena Oleksandrivna – a teacher of English, a Department of the Foreign Languages, Vinnytsia National Technical University. E-mail: olena.kriut@gmail.com.