МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ХАРЧОВИХ ТЕХНОЛОГІЙ

ВСП «ПОЛТАВСЬКИЙ ФАХОВИЙ КОЛЕДЖ НУХТ»

KHARKIV IT CLUSTER

Перша всеукраїнська

науково-практична конференція

«Комп’ютерна інженерія: сучасний стан та особливості підготовки фахівців»

15 – 16 квітня 2026 р.

Полтава ВСП ПФК НУХТ 2026

УДК 004

Наукові праці Першої всеукр. наук.-практ. конф. «Комп’ютерна інженерія: сучасний стан та особливості підготовки фахівців» (CE-2026), 15-16 квітня 2026 р. (Полтава, Україна). П. : ВСП ПФК НУХТ, 2026. 163 с.

У працях конференції наведено доповіді за напрямами:

- сучасні тенденції, технології та архітектурні рішення у сфері комп'ютерної інженерії та інформаційних технологій
- інноваційні педагогічні технології та моделі формування професійних компетентностей фахівців з комп’ютерної інженерії та інформаційних технологій
- екосистема розвитку комп’ютерної інженерії: індустріальні практики, інновації та інтеграція науки, освіти і бізнесу

Праці конференції будуть корисні науковим та інженерно-технічним працівникам, студентам ЗПФО, ЗВО та всім, хто цікавиться сучасними інформаційними системами та телекомунікаційними технологіями.

Подано в авторській редакції.

Автори матеріалів несуть повну відповідальність за достовірність наведеної інформації та відповідність матеріалів нормам законодавства, моралі й етики.

ISBN

УДК 004

ETHICAL AND TECHNICAL CHALLENGES OF EXPLAINABLE AI (XAI)

Hrama M.
National University of Food Technologies, Kyiv, Ukraine
E-mail: gramamp@nuft.edu.ua

Ethical and Technical Challenges of Explainable AI (XAI)

The rapid integration of artificial intelligence into critical sectors such as healthcare, finance, and criminal justice has underscored a fundamental tension between model performance and human understanding. This thesis explores the dual nature of Explainable Artificial Intelligence (XAI), analyzing the technical hurdles of post-hoc interpretability and the ethical dilemmas inherent in making complex algorithmic decisions transparent. By examining the trade-off between predictive accuracy and model simplicity, this work argues that true explainability is not merely a technical feature but a socio-technical necessity that requires reconciling mathematical fidelity with human cognitive limitations. The discussion extends to the risks of adversarial explanations and the potential for "fair-washing," concluding that the future of XAI must balance computational rigor with robust ethical frameworks to ensure accountability in autonomous systems.

The evolution of machine learning from simple linear regressions to sophisticated deep neural networks has created a paradox where the most capable models are often the least understandable. This "black box" nature of modern AI presents a significant barrier to the widespread adoption of automated systems in environments where the cost of error is high. Explainable Artificial Intelligence (XAI) has emerged as a critical field of research aimed at bridging this gap, yet it faces a daunting array of technical and ethical challenges that complicate its implementation. Deep learning architectures, particularly large-scale transformers and convolutional neural networks, derive their power from millions of parameters and non-linear interactions that defy simple human intuition. When researchers attempt to extract explanations from these models, they often rely on post-hoc methods that approximate the model's behavior rather than reflecting its internal logic directly. This leads to the problem of fidelity, where the explanation may be a simplified narrative that sounds plausible to a human observer but fails to accurately represent the actual decision-making process of the underlying algorithm.

The technical landscape of XAI is further complicated by the distinction between global and local interpretability. Global interpretability attempts to explain the overall logic of a model across all possible inputs, which is nearly impossible for high-dimensional deep learning systems without sacrificing their predictive power. Consequently, much of the field has shifted toward local explanations, which clarify why a specific decision was made for a single input. However, techniques such as Local Interpretable Model-agnostic Explanations (LIME) or Shapley Additive Explanations (SHAP) are often sensitive to noise and small perturbations in the data. If a minor change in an input leads to a radically different explanation while the model's output remains the same, the reliability of the XAI method is called into question. This instability creates a technical vulnerability where explanations can be manipulated, leading to a false sense

of security for end-users who might trust a flawed model simply because it provides a confident-sounding justification for its actions.

Beyond the purely mathematical and computational hurdles, the ethical implications of XAI are profound and multifaceted. One of the most pressing concerns is the phenomenon of "deceptive transparency" or "fair-washing." Because XAI methods are themselves models or approximations, they can be designed—intentionally or unintentionally—to hide the biased or discriminatory criteria used by the primary algorithm.

For instance, a loan-granting AI might rely on proxies for race or socioeconomic status, but its explanation module might highlight benign factors like "employment history" to satisfy regulatory requirements. This creates an ethical trap where transparency is used as a tool for obfuscation rather than accountability. The illusion of understanding can be more dangerous than a known black box, as it discourages critical oversight and allows systemic biases to persist under the guise of technological objectivity.

Furthermore, the "right to explanation," as enshrined in various legal frameworks like the General Data Protection Regulation (GDPR) in the European Union, poses an ethical and practical challenge regarding the audience of the explanation. An explanation that satisfies a data scientist might be entirely incomprehensible to a doctor, a judge, or a layperson. This subjectivity of understanding means that XAI must be tailored to the cognitive load and domain expertise of the specific user.

Accountability and liability represent another critical ethical frontier in the XAI discourse. In a scenario where an AI-driven medical diagnostic tool provides an incorrect recommendation along with a plausible-sounding explanation, who is responsible for the resulting harm? If the explanation appeared logical, the human practitioner might be exonerated for following the AI’s lead, yet the patient still suffers. This "responsibility gap" is exacerbated by XAI because it blurs the line between human judgment and machine suggestion. The explanation serves as a bridge, but if that bridge is built on flawed approximations, it may lead to a diffusion of responsibility where neither the software developers nor the end-users are held fully accountable for the system's failures.

In conclusion, while Explainable AI is a vital step toward creating trustworthy autonomous systems, it is not a panacea. The technical challenges of ensuring high-fidelity, stable, and secure explanations are inextricably linked to the ethical imperatives of fairness, privacy, and accountability. Moving forward, the development of XAI must move beyond post-hoc approximations toward "inherently interpretable" models that are transparent by design. This requires a paradigm shift in the AI community, prioritizing not just the maximization of accuracy, but the optimization of the human-AI partnership. Only by addressing both the mathematical limitations of interpretability and the sociological complexities of human trust can we build AI systems that are truly beneficial to society and consistent with our ethical values.