Nguyễn Chí Thanh-64TTNT2

2251262639

**Demystifying AI: A Robust and Comprehensive Approach to Explainable AI**

Vấn đề: Sự gia tăng sử dụng Trí tuệ Nhân tạo (AI) và Học máy (ML) trong nhiều nền tảng và lĩnh vực máy tính đòi hỏi sự phát triển của các kỹ thuật AI Giải thích được (XAI) mạnh mẽ. Hầu hết các mô hình AI hiện tại thiếu minh bạch về quy trình ra quyết định của chúng, cản trở niềm tin (trust), gỡ lỗi (debugging) và cải thiện[**Abstract—The adoption of Artificial Intelligence (AI) and Machine Learning (ML) in various computing platforms and areas, necessitates the development of strong Explainable AI (XAI) techniques. Most current AI models are opaque about their decision-making process thereby impeding trust, debugging, and improvement. The goal of this research is to develop comprehensive robust XAI methods capable of explaining the reasoning and decision-making processes in Autonomic, Edge, Server-less, Quantum computing platforms and IoT, Business Automation, Service Innovation domains where these AI models are deployed .This study comprehensively addresses the opacity in AI models through solutions for balanced test-train splits, model evaluation, feature importance, metric imbalances, ROC curve and precision-recall curve analysis, accuracy and statistical metrics, benefits of manual review. This research aims at increasing transparency and trustworthiness within AI systems through developing as well as applying such XAI methods that can detect and mitigate biases while enhancing ethical debugging; responsible development for AI enabled computing purposes.**]-[1].Đặc biệt, điều này trở nên quan trọng trong các bối cảnh rủi ro cao như tuyển dụng và chăm sóc sức khỏe, nơi sự công bằng được nhấn mạnh. Sự thiếu minh bạch này cũng gây khó khăn trong việc duy trì các định kiến (biases) có trong tập dữ liệu huấn luyện[**I. INTRODUCTION As the use of Artificial Intelligence (AI) and Machine Learning (ML) exhibits a steady increase, there is an increasing pressure on the development of Explainable AI (XAI) as a form of assistance in understanding why certain AI models make certain choices. Because of the way contemporary AI models are structured, people find it hard to trust them, they cannot tell if the model is performing well, or even seek to improve it, and this is more so because of the fear of extending the deficits found in the training set B. As AI technology increases acceptance and usage in context such as hiring [6] and healthcare [12], where fairness is highly emphasized, and where general decisions have high stakes, the demand for XAI methods that eliminate bias, and enforcement of fairness and accountability is even more helpful.**]

Mục tiêu: Nghiên cứu này nhằm phát triển các phương pháp XAI toàn diện và mạnh mẽ có khả năng giải thích lý do và quy trình ra quyết định trong các nền tảng máy tính Autonomic, Edge, Server-less, Quantum computing và các lĩnh vực IoT, Business Automation, Service Innovation, nơi các mô hình AI này được triển khai[1]. Mục tiêu là tăng tính minh bạch (transparency) và độ tin cậy (trustworthiness) trong các hệ thống AI thông qua việc phát triển và áp dụng các phương pháp XAI có thể phát hiện và giảm thiểu định kiến (detect and mitigate biases) đồng thời nâng cao gỡ lỗi đạo đức (ethical debugging) và phát triển có trách nhiệm[1].... XAI cũng giúp AI đưa ra quyết định tốt hơn và dễ bảo vệ hơn, khuyến khích mọi người tin tưởng AI hơn[**According to the stand of the European Data Protection Supervisor (EDPS), the use of XAI is precisely through It is illegal to discriminate against customers and also any purpose. Also, more and more studies have proved that XAI is a useful tool in enabling AI to make better and more defendable decisions and encourages people to trust AI more .But even with all progress there is still a long**

**way to go in creating adequate, efficient XAI that can serve the purpose without excessive sculpting.**]....

Phương pháp luận: Nghiên cứu áp dụng một phương pháp luận toàn diện để xây dựng và đánh giá các kỹ thuật XAI[ **The research applies a holistic methodology to build and evaluate Explainable AI (XAI) techniques for divergent computing architectures and contexts, in response to major obstacles and shortcomings of existing XAI approaches. The process includes literature review to inform framework development, data collection and preprocessing from diverse sources, development of new XAI algorithms through evaluation, gathering user feedback on effectiveness and usability through user studies, iterative refining and optimization based on performance metrics and user feedbacks, deployment and integration into real-world applications.**]-[2]. Quy trình bao gồm đánh giá tài liệu hiện có, thu thập và tiền xử lý dữ liệu từ nhiều nguồn khác nhau, phát triển các thuật toán XAI mới, đánh giá chúng thông qua các nghiên cứu người dùng (user studies) để thu thập phản hồi về hiệu quả và khả năng sử dụng, tinh chỉnh và tối ưu hóa lặp đi lặp lại dựa trên các chỉ số hiệu suất và phản hồi của người dùng, sau đó triển khai và tích hợp vào các ứng dụng thực tế[2]

Mô hình hệ thống: Mô hình hệ thống XAI được đề xuất bao gồm các thành phần: Module Nạp dữ liệu (Data Ingestion), Module Phát triển Kỹ thuật XAI (Module for Developing XAI Techniques) (có thể bao gồm LIME, SHAP, bản đồ nổi bật (saliency maps), tầm quan trọng của đặc trưng (feature importance), hoặc các phương pháp kết hợp), Module Đánh giá Mô hình (Model Evaluation), Module Nghiên cứu Người dùng (User Study), và Module Triển khai (Deploying Module)[**The system model of Explainable AI (XAI) proposed to facilitate the design, evaluation, and deployment of XAI strategies across various computing platforms and domains. The system model has several components**]

Kết quả: Khung XAI được đề xuất đã được đánh giá và cho thấy hiệu suất vượt trội so với mô hình nền (baseline model) và các công trình trước đó[**We assessed our proposed framework with a dataset comprising of 1000 samples, dividing 500 for training purpose and reserving 500 for testing purpose. The results are presented in below:**

**In all the metrics depicted in the table as in terms of accuracy, F1 score, AUC ROC, transparency, fairness among other pertinent metrics, our proposed framework beats the baseline model. For instance, the proposed framework has an accuracy of 95.2%, which is 3.1% above the baseline model. Similarly, there have been improvements in other parameters such as the F1 score and AUC ROC, where the framework managed to attain 94.5% and 97.1% respectively.**]-[3]Cụ thể, khung này đạt độ chính xác 95.2%, vượt trội so với baseline và các công trình trước đó[3].... Các chỉ số khác cũng được cải thiện như F1 score 94.5% và AUC ROC 97.1%[3]. Về mặt minh bạch và công bằng, khung đề xuất đạt minh bạch 92.5% và công bằng 90.2%, vượt trội so với các công trình trước đó[**Our proposed XAI framework has surpassed earlier contributions in the area of explainable AI. For instance, the work of Hofeditz et al. has an accuracy of 92.1 percent implementation which is less than that reached by our proposed framework of 95.2 percent. Similarly, the work of Zhang et al. has an accuracy of F1-Score of 91.3 percent that is much lower than our proposed F1-Score complimentary to the framework work, which is 94.5 percent.**

**In the same way, intervening under rule-based approaches manages to result in fair and undistorted outcomes. The work by Thiebes et al [6] achieves transparency of only 85.1% where as, the proposed framework achieves transparency of 92.5 %. The same case applies to work by Mittelstadt et al.where a fairness score of and only 84.5% is reached whereas, a busting fairness score of 90.2% is attained with the aid of the proposed framework.**].

Kết luận: Nghiên cứu đã đề xuất một mô hình AI Giải thích được (XAI) mới linh hoạt, có thể điều chỉnh và rõ ràng hơn các mô hình trước[ **A new Explainable AI (XAI) model was thus proposed in this study that is more flexible, adjustable and clear than the previous models. Thus, our model outperforms other methods that exist in terms of interpretability, accuracy, and explainability, transparency, fairness. so it can be used for many real life applications.Therefore, our approach is efficient enough to show how decisions are made by AI models.Thus, the proposed model is superior to existing ones due to its ability to handle complex data distributions, adapt to new data as well as provide transparent explanations.Such strengths thus imply that health care and finance could be among areas best suited for application of the model.**]-[4]. Mô hình này vượt trội so với các phương pháp hiện có về khả năng giải thích (interpretability), độ chính xác, tính minh bạch và công bằng, cho thấy hiệu quả trong việc hiển thị cách các mô hình AI đưa ra quyết định[4]. Khung đề xuất phù hợp cho nhiều ứng dụng thực tế, đặc biệt là trong các lĩnh vực như chăm sóc sức khỏe và tài chính do khả năng xử lý dữ liệu phức tạp, thích ứng với dữ liệu mới và cung cấp giải thích minh bạch[4].

Cải tiến tương lai: Các hướng cải tiến trong tương lai có thể bao gồm phát triển giải thích đa phương thức (Multi-model Explanations) (ví dụ: kết hợp hình ảnh hóa, giải thích bằng ngôn ngữ tự nhiên), giải thích trong thời gian thực (Explainability in Real Time) cho dữ liệu luồng, tích hợp con người vào vòng lặp (Human-in-the-Loop), giải thích cho các mô hình Học sâu (Deep Learning Models), Hệ thống Đa tác nhân (Multi-Agent Systems), AI Biên (Edge AI), và Học chuyển giao (Transfer Learning)[**Multi-model Explanations:Currently, our model explains through feature importance scores. The future studies can thus focus on developing multimodal explanations involving visualizations, natural language explanations as well as interactive dashboards.**

**Explainability in Real Time: Our hypothesis explains the batch data. In future, real-time explainability could be developed to allow streaming data explanations by the model.**]