Addressing High-Risk Contexts with Transparent AI

Marco Pota[1], Erica Vaccaro[2]

1. CNR, Naples, Italy | 2. University of Naples Federico II

Abstract. Artificial Intelligence (AI) is increasingly used in high-risk fields like healthcare and public administration, where balancing accuracy and interpretability is crucial. The AI Act and Ethics Guidelines for Trustworthy AI stress the need for transparency, explainability, and accountability. However, traditional AI models often sacrifice one aspect for the other, limiting their adoption.This study explores whether Symbolic Neural Networks can achieve a better balance between performance, transparency, and robustness compared to deep learning and symbolic AI. Special focus is given to Kolmogorov-Arnold Networks (KAN), a Neuro-Symbolic AI architecture that decomposes AI decisions into interpretable mathematical functions. The research assesses their potential in Decision Support Systems (DSS) for medical diagnostics and geo-hazard monitoring, ensuring compliance with ethical and regulatory standards.

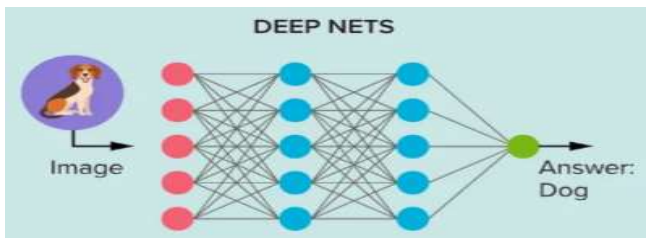

Fig. 1. Represents a Deep Network

INTRODUCTION

The widespread adoption of AI in decision-making processes has raised concerns about trust, transparency, and accountability, particularly in high-risk fields where human oversight is essential. Traditional deep learning models, while highly accurate, often function as black boxes, making their decisions difficult to interpret. This lack of transparency limits trust and hinders their integration in sectors such as healthcare, Public Administration.To address these challenges, the AI Act and the Ethics Guidelines for Trustworthy AI highlight the necessity of explainability and robustness in AI systems. Symbolic AI methods, such as decision trees and rule-based models, provide greater interpretability but often at the cost of reduced predictive accuracy. Neuro-Symbolic AI, particularly Symbolic Neural Networks, aims to combine the adaptability of neural models with structured symbolic reasoning, creating AI systems that are both powerful and interpretable.Among recent advancements, Kolmogorov-Arnold Networks (KAN) propose a unique approach to explainability by

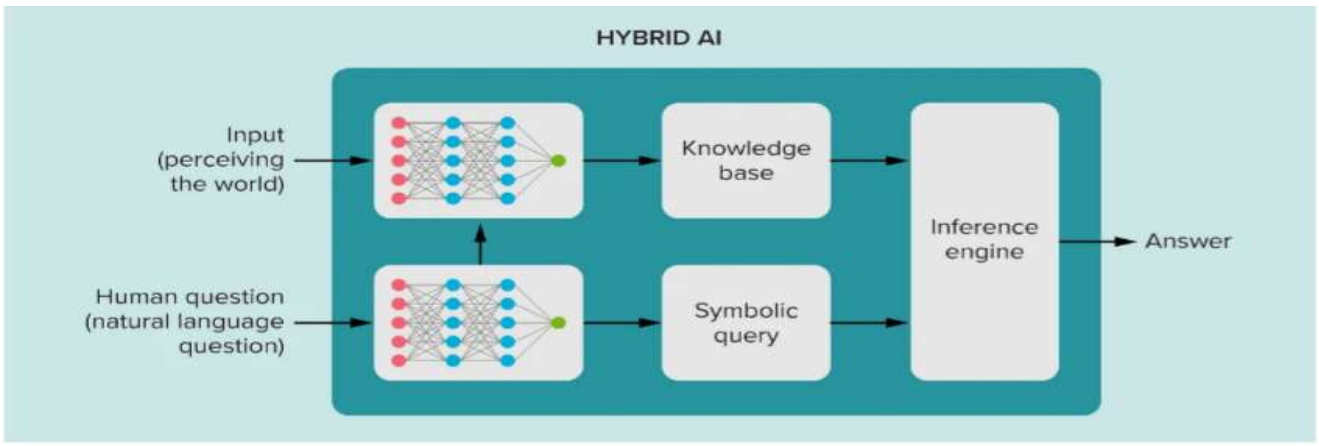

Fig. 2. Represents Hybrid models of AI

decomposing complex AI function into interpretable mathematical transformations. This hybrid method could significantly enhance AI transparency in Decision Support Systems, reducing the opacity of traditional deep learning approaches.

LITERATURE REVIEW

Existing research on AI interpretability highlights two main approaches: post-hoc explanation techniques and inherently interpretable models. Methods such as SHAP and LIME attempt to explain black-box models after predictions are made, but they do not improve the model's intrinsic transparency. Conversely, Symbolic AI methods, including decision trees and rule-based systems, offer high interpretability but often sacrifice predictive accuracy, limiting their applicability in complex tasks. Recent studies in Neuro-Symbolic AI explore hybrid models that combine deep learning with symbolic reasoning to balance accuracy and interpretability. Kolmogorov-Arnold Networks (KAN), in particular, enhance transparency by decomposing complex functions into structured mathematical representations. The literature suggests that KAN and similar architectures improve explainability while maintaining strong performance, making them suitable for medical diagnostics and risk assessment.

RESEARCH METHODOLOGY

The methodology involves a comparative analysis of Symbolic Neural Networks, Deep Neural Networks (DNNs), and Symbolic AI methods, with particular attention to Kolmogorov-Arnold Networks (KAN) as a potentially more interpretable alternative. Selected models will be benchmarked on standardized AI datasets, assessing accuracy, uncertainty estimation, and computational efficiency. Explainability will be evaluated through mathematical function decomposition, feature attribution techniques (SHAP, LIME), and rule extraction to ensure compliance with Trustworthy AI guidelines. Finally, domain-specific validation will be conducted through expert reviews in medical and geo-hazard applications, assessing the practical usability of interpretable AI models in real-world decision-making.

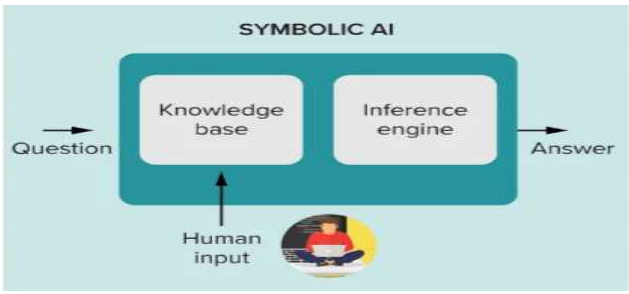

Fig. 3. Represents Symbolic AI Model

CONCLUSIONS

This research introduces a Neuro-Symbolic AI framework to assess whether Symbolic Neural Networks can bridge the gap between accuracy and interpretability in high-risk AI applications. By leveraging KAN and similar architectures, this study aims to determine if these models provide a more transparent, explainable, and ethically compliant alternative to traditional deep learning approaches. Preliminary findings suggest that KAN-based AI models could significantly improve the interpretability of Decision Support Systems without sacrificing predictive performance. However, further validation is required to evaluate real-world applicability in sensitive domains with high impact on human rights.

References.

[1] Xu, K., Chen, L., & Wang, S. (2024). Kolmogorov-Arnold Networks for Time Series: Bridging Predictive Power and Interpretability.

[2] Delfosse, Q., Shindo, H., Dhami, D., & Kersting, K. (2023). Interpretable and Explainable Logical Policies via Neurally Guided Symbolic Abstraction

[3] Marcus, G. (2020). The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence.

[4] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A survey of methods for explaining black box models. ACM Computing Surveys

[5] Pota, M., Esposito, M., & De Pietro, G. (2016). Interpretability indexes for fuzzy classification in cognitive systems.

Acknowledgements. The HCAIM (the Human-Centred AI Master's Programme).This poster was created as part of the Blended Intensive Programme organized under the Erasmus + Programme of the European Union