Explainable AI is Responsible AI:

A survey of transparency in AI approaches

**Nika Gedenidze** (ngedenidze@caldwell.edu)

**Vladislav D. Veksler** (vveksler@caldwell.edu)

School of Business and Computer Science

Caldwell University

Caldwell, NJ, USA

1. **INTRODUCTION**

In recent years, software engineers and scientists made advances into machine learning and deep learning. AI systems are reaching and sometimes even exceeding the human level on an increasing number of complex tasks involving big datas. In domains such as strategic game playing, image classification, sentiment analysis, speech understanding AI has already shown an impressive development. In spite of these benefits, AI-based systems still have the major issue -the lack of transparency. “Black box” in machine learning makes it almost impossible to interpret the inner workings of the system. It is crucial for an organization to have a full understanding of the AI decision-making processes with model monitoring and accountability of AI and not to trust them blindly. The great solution for this problem is Explainable AI[[1]](#footnote-0)(XAI), which can help humans understand and explain machine learning (ML) algorithms, deep learning and neural networks.

XAI is still a new concept - the earliest work on it could be found in the literature published forty years ago**[1]**, where expert systems explained their results via the applied rules. It was mainly overlooked in recent decades as the AI was not getting as much attention from businesses and users. However, after people started widespread adoption of AI, Explainable AI has become a new attraction for modern deep learning engineers. “Explainability is an easy problem to sweep under the carpet, especially when AI is in the middle of a major hype cycle — but it is the Achilles heel of AI/ML implementations[2]” - Says Michael Lukianoff in the DERPA’s XAI article. His published document promises new advances towards transparency of machine learning and gives insight of the XAI advantages and how predictability is correlated to transparency. The Fig.1 below shows that Explainability of a machine learning model is usually inverse to its prediction accuracy - the higher the prediction accuracy, the lower the model explainability. The focus of this research is to compare multiple AI designs, like DERPA, for transparency to single out more responsible, trustworthy, and safe approaches.

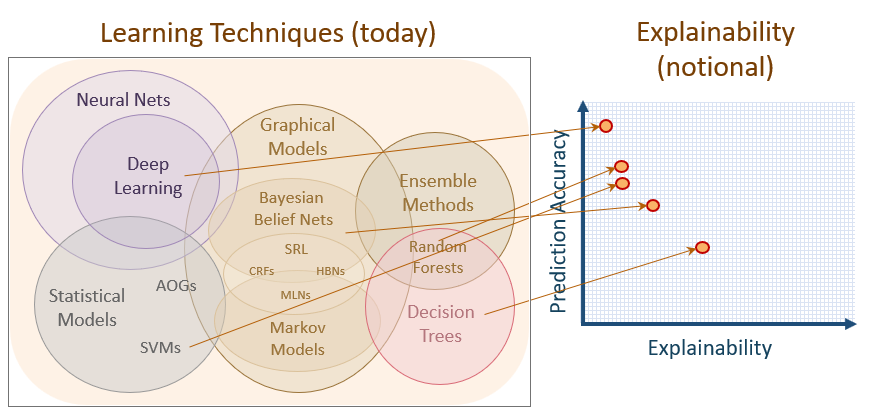


Fig.1 Graph of

1. **RELEVANCE OF XAI**

Applicability of AI is growing in a fast paced economical world, so does the burden of responsibility it has. It became part of people’s everyday life purposes ranging from music recommendation engines in mobile music apps to NLP applications[[2]](#footnote-1) in healthcare systems.

1. **MAJOR XAI APPROACHES**
2. **Interpretable Models**

Interpretability is the degree to which humans can predict the model's outcome. The higher interpretability of machine learning, the easier it is for engineers or users to understand why certain decisions or predictions have been made by the AI agent. It is a core characteristic for machine learning to start working towards removing the “black box” from it. Sometimes, when you work on achieving this interpretability you have to make the tradeoff between the knowledge of predictivity and information on why the prediction was made(possibly pay for the interpretability with a drop in predictive performance). “In some cases, the user does not care why a decision was made - knowing that the predictive performance on a test dataset was good enough. On the other hand, knowing the 'why' can help you learn more about the problem, the data and the reason why a model might fail .**[4]**” Some models may not require explanations because they are used in a low-risk environment, meaning a mistake will not have serious consequences or the method has already been extensively studied and evaluated. The need for interpretability arises from an incompleteness in problem formalization, which means that for certain problems or tasks it is not enough to get the prediction. Algorithms like Linear regression, Logistic regression, Decision trees, RuleFit, and KNN have their own advantages and disadvantages in interpretability. The prediction is often a linear combination of

features, which is both its greatest strength and its greatest limitation. Linear effects are easy to quantify, describe, and interpret, but they have their own limitations. Sometimes engineers have to use models which don't have easily interpretable designs, for example, in computer vision or in natural language processing. If we want to use the method that works on all types of models we will have to use Model-Agnostic explainability models(SHAP, Perturbation, Lime).

1. **Model-Agnostic Methods**

(TEXT NEEDS TO BE INSERTED)

1. **Model-Specific Methods**

(TEXT NEEDS TO BE INSERTED)

1. **Example-Based Methods**

(TEXT NEEDS TO BE INSERTED)

1. **Neural Representations**

(TEXT NEEDS TO BE INSERTED)

1. **CONCLUSION**

(TEXT NEEDS TO BE INSERTED)

1. **REFERENCES**

[1]Scott, A.C., Clancey, W.J., Davis, R., Shortliffe, E.H.: Explanation capabilities of production-based consultation systems. American Journal of Computational Linguistics 62 (1977)

[2]Lukianoff, M. (2019, December 13). *Explainable artificial intelligence (XAI) is on DARPA's agenda - why you should pay attention*. Medium. Retrieved April 2, 2022, from https://towardsdatascience.com/explainable-artificial-intelligence-xai-is-on-darpas-agenda-why-you-should-pay-attention-b63afcf284b5

[3]<https://www.youtube.com/watch?v=Yg3q5x7yDeM>

[4]Molnar, Christoph. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. United States, Lulu.com, 2020.

1. Explainable AI (XAI), or Interpretable AI, is artificial intelligence (AI) in which the results of the solution can be understood by humans. [↑](#footnote-ref-0)
2. Natural language processing- applications such as speech recognition, text analysis, translation and other goals related to language [↑](#footnote-ref-1)