3rd World Conference on
ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND DATA SCIENCE
May 23, 2024 | Rome, Italy
https://doi.org/10.62422/978-81-970290-9-7-003

Survey on XAI: Different Approaches and Aspects

Raimondo Fanale
Universitas Mercatorum, Rome, Italy

Deep learning has gained popularity due to its ability to learn from vast amounts of data, but its complexity often makes it difficult for humans to understand. "Reliable" AI, therefore, requires robustness, interpretability, and explainability. The field of eXplainable Artificial Intelligence (XAI) focuses on developing tools to design and understand complex AI models.

It's important to consider three aspects: whether the explanations provided are reliable, whether there is a risk of misrepresentation or misinterpretation, and whether any vulnerabilities can be exploited. Different contexts create different needs for explainability, and system design may need to balance these competing needs. Transparency and explainability in AI methods are only the first steps in creating trustworthy systems. This may require both technical approaches and other measures, such as guaranteeing certain properties. Designers and developers of AI must consider how its use fits into a broader technical and social context.

There are many reasons why interpretability and explainability in AI systems are desirable or necessary. These include giving users confidence, safeguarding against bias, adhering to regulatory standards, helping developers understand why a system works, assessing its vulnerabilities, verifying its outputs, and meeting society's expectations about decision-making processes.

In the future, as shown by this recent analysis and trends, explainable and interpretable AI must always be applied. It should be enhanced with methods such "attention" to include causality and measure the quality of explanations.

Biography:

Data architect/scientst/enthusiast, developer, professor and founder of 3 companies. Started to code for fun at 9 and the fun has become a job.

B.Sc. Hons – First class honour at UniDerby (UK): Multilevel Spam filtering using NLP

M.Sc. at UniDerby: Increase efficiency in clinics with Augmented Analytics

Currently involved in PhD study and research in Big Data and AI at Universitas Mercatorum.

I would like to acknowledge my PhD supervisors Prof. R. Caldelli, Prof.ssa B. Martini, Prof. F. Sciarrone; this work is part of the PhD course in Big Data and AI.