The Research Monograph Series in Computing, Electrical & Communication Networks
Vol. 1 (01), June 2023, pp. 5–9

CHAPTER 2

STATE OF ART

2.1 AI AND TRUST MANAGEMENT

Trust is a critical component in the adoption and success of AI applications. Users must trust that AI systems are reliable, accurate, and secure. However, trust in AI can be difficult to establish and maintain, given the complexity and opacity of these systems.

One way to manage trust in AI is through transparency. AI systems should be designed to be explainable, meaning that the decision-making process should be clear and understandable. This allows users to understand how the AI arrived at its conclusions and to identify potential biases or errors in the system.

Another way to manage trust in AI is through regulation and oversight. Governments and regulatory bodies can establish guidelines and standards for AI development and use, ensuring that AI applications meet certain requirements for reliability, accuracy, and security.

Additionally, organizations can build trust in AI by engaging with stakeholders and involving them in the development process. This includes gathering feedback from users and addressing their concerns, as well as collaborating with experts in the field to ensure that the AI system is reliable and accurate.

Ultimately, building trust in AI requires a holistic approach that includes both technical and social considerations. AI developers must consider the ethical implications of their work and be transparent about their processes and decisions. Users must also be educated about the limitations and potential biases of AI systems, so that they can make informed decisions about their use.

2.2 ARTIFICIAL INTELLIGENCE IN CYBERSECURITY

Cybersecurity and Artificial Intelligence (AI) are two critical areas that are increasingly becoming interdependent in the modern world. As AI technologies continue to advance, they are being incorporated

into more and more aspects of our daily lives, including in the realm of cybersecurity. Here are some key points to consider when discussing the relationship between cybersecurity and AI:

AI for Cybersecurity: AI technologies are being used to improve cybersecurity in several ways. For example, AI algorithms can analyze large amounts of data to detect and prevent cyber attacks. Machine learning models can be trained to identify patterns in data that may indicate a cyber threat. Additionally, AI can be used to automate security operations, such as monitoring network traffic and responding to security incidents.

Cybersecurity for AI: As AI systems become more prevalent, they also become targets for cyber attacks. Adversaries may seek to exploit vulnerabilities in AI systems to gain unauthorized access to data or to manipulate the behavior of the AI system itself. Therefore, it is essential to implement robust cybersecurity measures to protect AI systems from cyber threats.

Ethical Considerations: The use of AI in cybersecurity also raises important ethical considerations. For example, AI algorithms may make decisions that have significant consequences for individuals or organizations. It is essential to ensure that these algorithms are transparent, explainable, and free from biases to prevent unintended harm.

Human Expertise: While AI can automate many aspects of cybersecurity, human expertise remains critical. Cybersecurity professionals must understand how to configure and monitor AI systems effectively to ensure that they are providing accurate and reliable results. Additionally, human experts are needed to interpret the output of AI systems and make decisions about how to respond to security incidents.

2.3 TRANSITION TO EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

A subfield of artificial intelligence called Explainable Artificial Intelligence (XAI) places a strong emphasis on how crucial transparency and interpretability are for machine learning algorithms. Humans can make informed decisions and spot any biases or errors thanks to the use of XAI techniques that strive to build models that are easily understandable by humans.

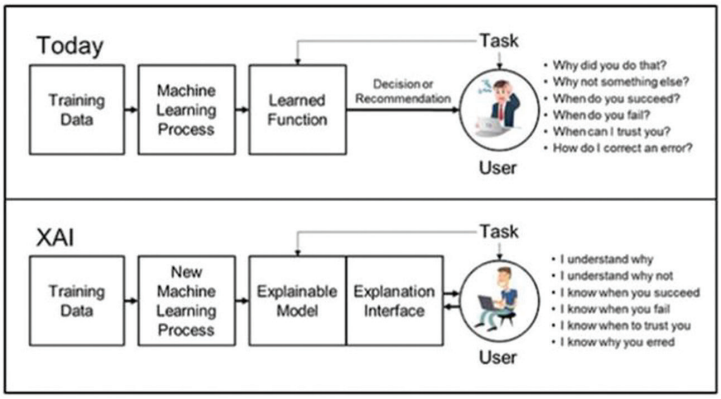

Figure 5. Explainable artificial intelligence (xAI) project proposed by DARPA [2, 3].

These models are intended to be integrated with cutting-edge human-computer interactive interface approaches that can result in explanation dialogues for the end user that are clear and helpful (Figure 5).

If you want to transition to XAI, there are several steps you can take:

Understand the basics of XAI: Before you can start implementing XAI techniques, it's important to understand what it is and how it works. There are many resources available online to help you learn more about XAI, including research papers, blogs, and tutorials.

Choose the right tools: There are many XAI tools available, ranging from simple visualization tools to complex machine learning frameworks. Depending on your needs, you may need to use a combination of different tools to achieve the level of transparency and interpretability you require.

Evaluate your existing models: If you have existing machine learning models, you can start by evaluating their transparency and interpretability. This will help you identify potential areas for improvement and determine which XAI techniques to use.

Implement XAI techniques: Once you have chosen the right tools and evaluated your models, you can start implementing XAI techniques. This may involve creating visualizations to help explain how your models work, using feature importance techniques to identify which features are most important, or using techniques like counterfactual explanations to help explain why a particular decision was made.

Monitor and evaluate: Once you have implemented XAI techniques, it's important to monitor and evaluate their effectiveness. This will help you identify areas for improvement and ensure that your models remain transparent and interpretable over time.

Transitioning to XAI requires a combination of technical expertise, domain knowledge, and a commitment to transparency and interpretability. By following these steps, you can help ensure that your machine learning models are more transparent and interpretable, which can lead to better decision-making and improved outcomes.

Explainable Artificial Intelligence (XAI) is a branch of artificial intelligence that emphasizes the importance of transparency and interpretability in machine learning models. XAI techniques aim to create models that can be easily understood by humans, allowing us to make informed decisions and identify potential biases or errors.

Explainable AI (XAI) is a set of techniques and methods used to create artificial intelligence (AI) systems that can be easily understood by humans. XAI aims to increase transparency and trust in AI systems by providing insights into how they make decisions and why they arrived at a particular output.

XAI is particularly important in applications where AI systems are used to make critical decisions, such as healthcare, finance, and legal systems. In these domains, it is essential to have a clear understanding of the factors that influenced the AI system's output, as it may have significant consequences for people's lives.

XAI techniques include visualization methods, natural language generation, and interpretable models. Visualization techniques use charts, graphs, and other graphical representations to help users understand how the AI system arrived at a particular decision. Natural language generation generates explanations in plain language, making it easier for users to understand the system's decision-making process. Interpretable models, such as decision trees and linear models, provide insight into how the AI system makes predictions and decisions.

XAI seeks to create AI systems that are transparent, interpretable, and explainable, so that humans can understand the reasoning behind their decisions and trust the output.

Indeed, the success of machine learning (ML) and deep learning (DL) models can be attributed to their ability to efficiently learn from large amounts of data and extract complex patterns and relationships from it. The large parametric space, consisting of numerous layers and millions of parameters, allows these models to capture and represent a vast range of features and patterns in the data.

However, the complexity of these models also poses challenges in understanding how they make decisions and predictions. ML and DL models are often viewed as black-box models because their internal workings are not easily interpretable. This lack of transparency can be problematic in many real-world applications, where the decision-making process needs to be transparent and explainable.