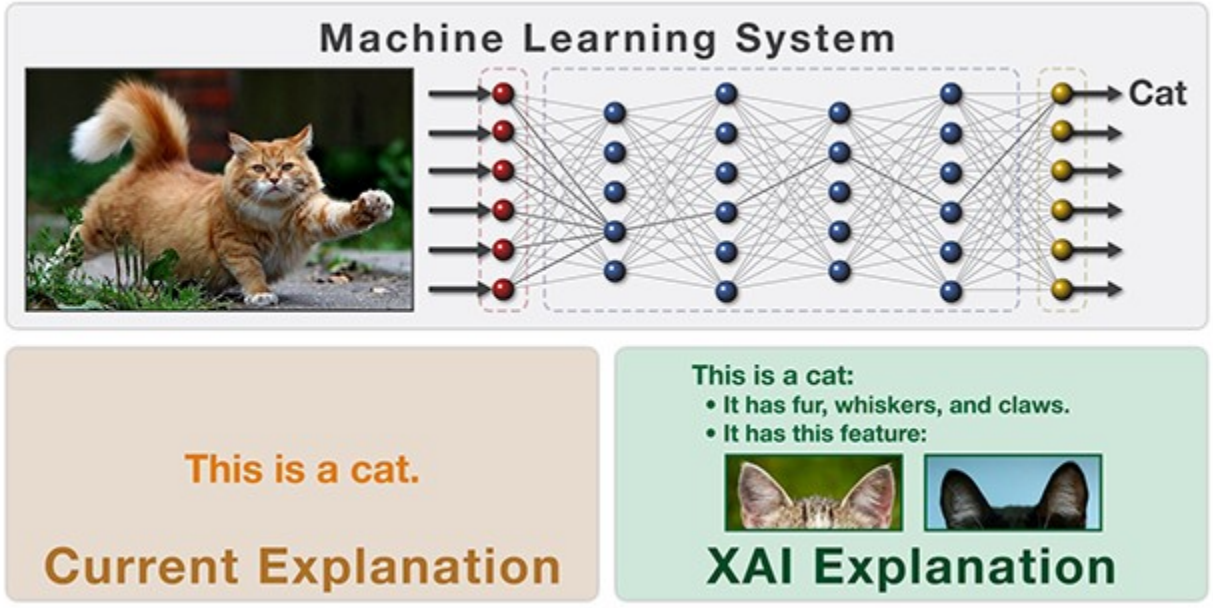
Defense Advanced Research Projects Agency Program Information

Explainable Artificial Intelligence (XAI)

Mr. David Gunning



Dramatic success in machine learning has led to a torrent of Artificial Intelligence (AI) applications. Continued advances promise to produce autonomous systems that will perceive, learn, decide, and act on their own. However, the effectiveness of these systems is limited by the machine's current inability to explain their decisions and actions to human users. The Department of Defense is facing challenges that demand more intelligent,

autonomous, and symbiotic systems. Explainable AI—especially explainable machine learning—will be essential if future warfighters are to understand, appropriately trust, and effectively manage an emerging generation of artificially intelligent machine partners.

The Explainable AI (XAI) program aims to create a suite of machine learning techniques that:

Produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and

Enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.

New machine-learning systems will have the ability to explain their rationale, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future. The strategy for achieving that goal is to develop new or modified machine-learning techniques that will produce more explainable models. These models will be combined with state-of-the-art human-computer interface techniques capable of translating models into understandable and useful explanation dialogues for the end user. Our strategy is to pursue a variety of techniques in order to generate a portfolio of methods that will provide future developers with a range of design options covering the performance-versus-explainability trade space.

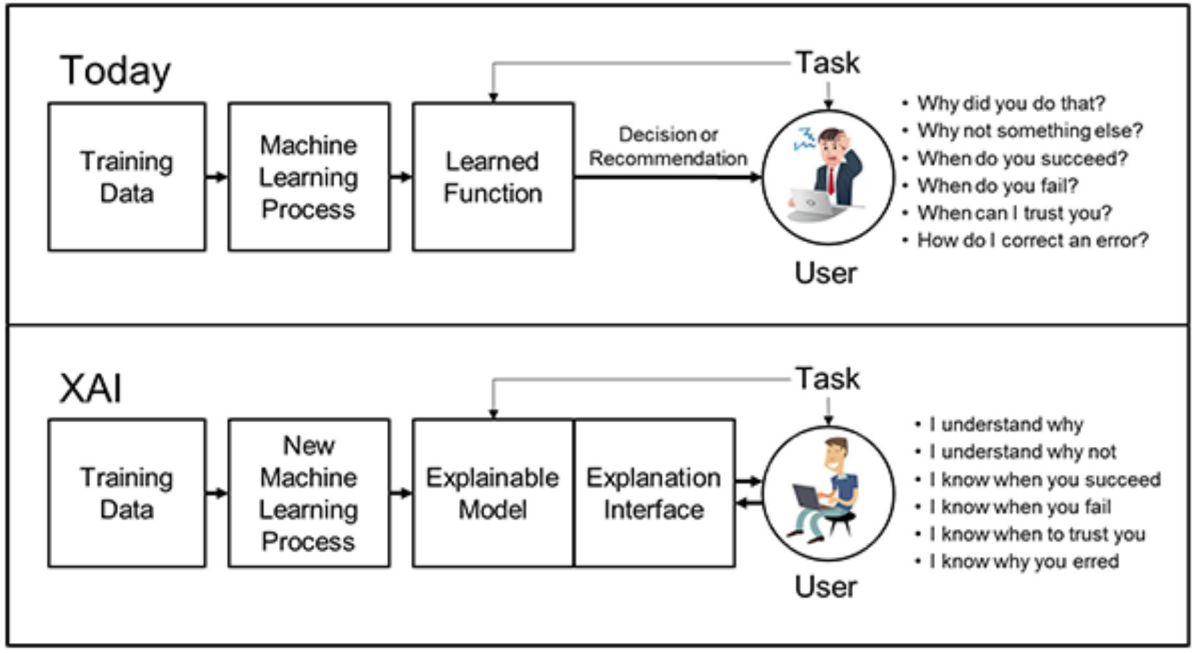


Figure 1: XAI Concept

The XAI program will focus the development of multiple systems on addressing challenges problems in two areas: (1) machine learning problems to classify events of interest in heterogeneous, multimedia data; and (2) machine learning problems to construct decision policies for an autonomous

system to perform a variety of simulated missions. These two challenge problem areas were chosen to represent the intersection of two important machine learning approaches (classification and reinforcement learning) and two important operational problem areas for the Department of Defense (intelligence analysis and autonomous systems).

XAI research prototypes will be tested and continually evaluated throughout the course of the program. At the end of the program, the final delivery will be a toolkit library consisting of machine learning and human-computer interface software modules that could be used to develop future explainable AI systems. After the program is complete, these toolkits would be available for further refinement and transition into defense or commercial applications.

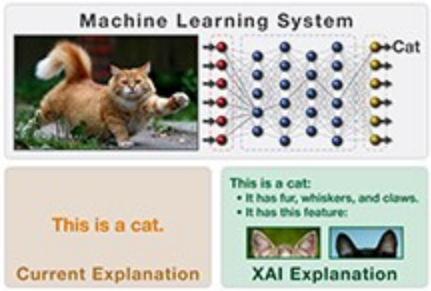
TAGS

| AI | Data | Interface | Programming |

SIMILARLY TAGGED CONTENT

TRAnsformative DESign (TRADES) Proposers Day
New Tools for Human-Machine Collaborative Design
DARPA Seeks to Remove Communication Barrier Between Humans and Computers
Big Mechanism Seeks the “Whys” Hidden in Big Data
DARPA Envisions the Future of Machine Learning

IMAGES



Explainable AI (XAI)