Defense Advanced Research Projects Agency   Our Research   Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) (Archived)

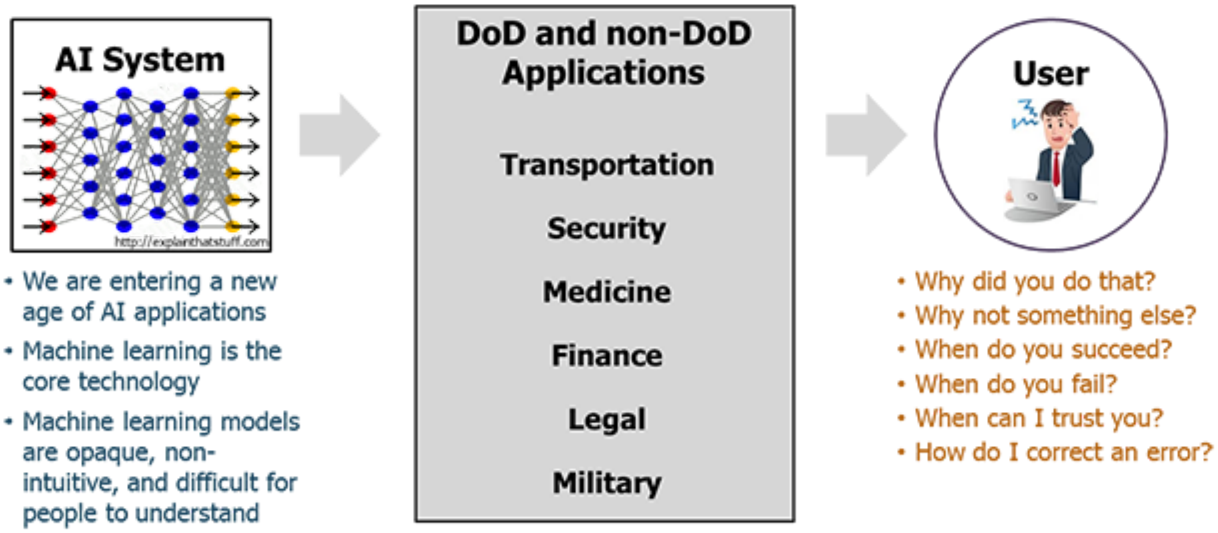
AI System
http://explainthatstuff.com
We are entering a new age of AI applications
Machine learning is the core technology
Machine learning models are opaque, non-intuitive, and difficult for people to understand
DoD and non-DoD Applications
Transportation
Security
Medicine
Finance
Legal
Military
User
Why did you do that?
Why not something else?
When do you succeed?
When do you fail?
When can I trust you?
How do I correct an error?

Figure 1. The Need for Explainable AI

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 (Figure 1). The Department of Defense (DoD) 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 (Figure 2). 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.

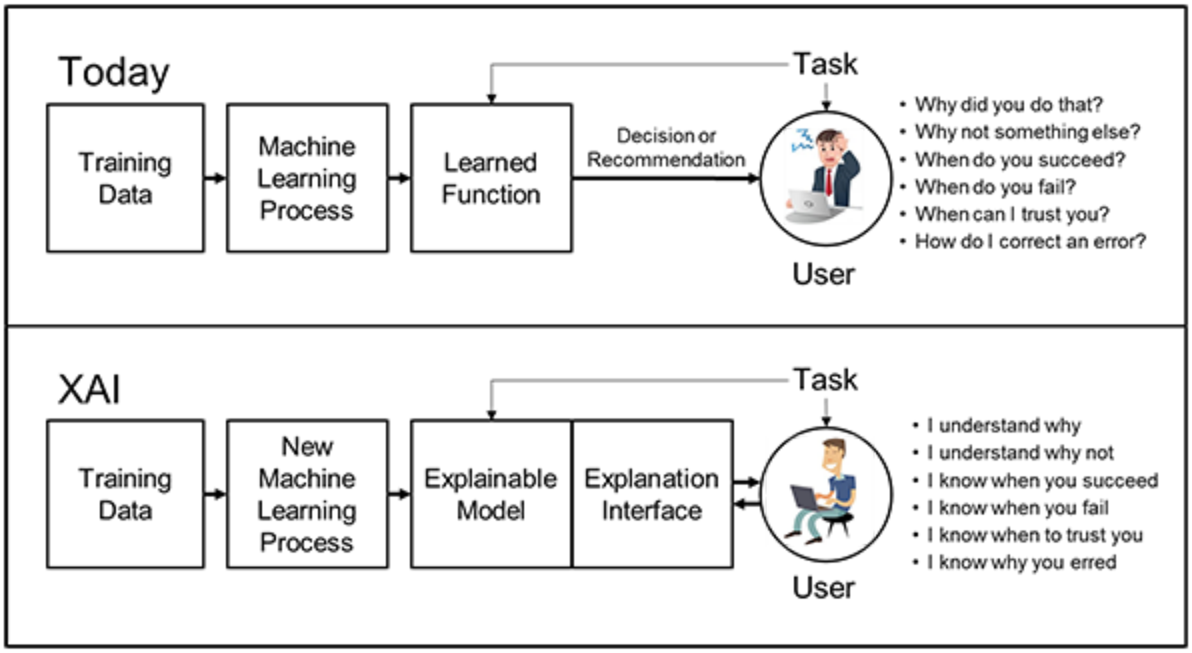

Figure 2. XAI Concept

XAI is one of a handful of current DARPA programs expected to enable “third-wave AI systems”, where machines understand the context and environment in which they operate, and over time build underlying explanatory models that allow them to characterize real world phenomena.

The XAI program is focused on the development of multiple systems by addressing challenge 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 DoD (intelligence analysis and autonomous systems).

In addition, researchers are examining the psychology of explanation.

XAI research prototypes are tested and continually evaluated throughout the course of the program. In May 2018, XAI researchers demonstrated initial implementations of their explainable learning systems and presented results of initial pilot studies of their Phase 1 evaluations. Full Phase 1 system evaluations are expected in November 2018.

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.

XAI RESOURCES AVAILABLE TODAY