XAI

Explainable AI—especially explainable machine learning—will be essential if future war fighters 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). 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-explain ability trade space.

A picture containing television, screen, sitting, monitor

Description automatically generated

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)..