Explainable Artificial Intelligence (XAI) Program

Defense Advanced Research Projects Agency (DARPA) Information Innovation Office (I2O)
Dave Gunning, Program Manager, david.gunning@darpa.mil
xai@darpa.mil

Introduction

Dramatic success in machine learning (ML) has led to a new wave of AI applications (e.g., transportation, security, medicine, finance, and defense) that offer tremendous benefits, but cannot explain their decisions and actions to human users. DARPA's XAI program endeavors to create AI systems whose learned models and decisions can be understood and appropriately trusted by human users. Realizing this requires methods for learning more explainable models, designing effective explanation interfaces, and understanding the psychological requirements for effective explanations. XAI's eleven system developer teams are addressing the first two challenges by creating new ML techniques and developing principles, strategies, and human-computer interaction techniques for generating effective explanations. Another XAI team is addressing the third challenge by summarizing, extending, and applying psychological theories of explanation to assist the XAI Evaluator with defining a suitable evaluation framework to test the XAI systems.

Problem Statement

The XAI program addresses two operationally relevant challenge problem areas: 1) **data analytics** (e.g., classification of events of interest in heterogeneous multi-media data), and 2) **autonomy** (e.g., decision policies of autonomous systems).

The **data analytics challenge** was motivated by the need for intelligence analysts to effectively utilize information provided by AI systems. For example, presented with recommendations/decisions from an AI system's algorithms, analysts must determine which to include as supporting evidence in their analyses/reports, which to pursue further, and perhaps which to ignore. AI systems that produce effective explanations have the potential to help intelligence analysts better assess evidence that might support or contradict their hypotheses, and more readily identify potential issues such as 'concept drift' in an AI system's algorithmic processes.

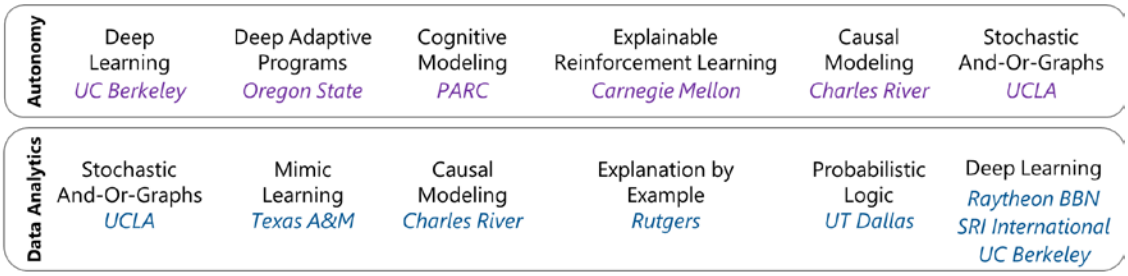
The **autonomy challenge** was motivated by the need for military operators to effectively manage autonomous AI partners. For example, as the military continues to adopt the use of AI systems with increasing levels of autonomy, operators will need to understand the behavior of these systems under various conditions. AI systems that produce effective explanations have the potential to help operators better determine their applicability and utility in future mission scenarios.

Teams, Approaches, and Techniques

Eleven XAI teams are investigating a diverse range of techniques and approaches for developing **explainable models** and effective **explanation interfaces**. Three system developer teams are addressing both program challenge problem areas of autonomy and data analytics; three team are addressing autonomy only; and five teams are addressing data analytics only.

Approved for Public Release, Distribution Unlimited

Figure 1. XAI System Developers, Approaches, and Challenge Problem Areas



One XAI research team (led by the Florida Institute for Human and Machine Cognition, with team members Macrocognition, LLC and the Michigan Technological University) is summarizing, extending, and applying psychological theories of explanation (e.g., What is an explanation? How do humans explain things? How might intelligent systems help humans better understand and explain?) The team is constructing a model of explanation to support the eleven XAI system developer teams (e.g., by informing explanation interface designs and user study protocol designs) as well as the government evaluator team (e.g., to define measures of explanation effectiveness for the evaluation framework). The team has authored a series of published articles on "explaining explanation" as part of this effort (Theoretical Foundations, Empirical Foundations, The Causal Landscape, A Deep Dive on Deep Nets).

Key Objectives and Current Status

The overarching goal of this fundamental research program is to create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance; and enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI systems. The high-level, key program metric objective is **explanation effectiveness** (as measured by user satisfaction, mental model, trust, and task performance) measured under various experimental conditions: without explanation (e.g., the AI system is used to perform a task without providing an explanation to the user), with explanation (e.g., the AI system is used to perform a task and generates explanations for every recommendation or decision it makes, and every action it takes), partial explanation (the AI system is used to perform a task and generates only partial or ablated explanations to the user, and control (e.g., a baseline state-of-the-art non-explainable AI system is used to perform a task). The program includes three evaluations over the course of two phases:

- Phase 1: developer team-specified challenges in data analytics and/or autonomy and developer team-executed user studies and evaluations (Evaluation 1); and
- Phase 2: developer team-specified challenges in data analytics/autonomy and government team-executed user studies and evaluations (Evaluation 2); government team-specified challenges in data analytics/autonomy and government team-executed user studies and evaluations (Evaluation 3).

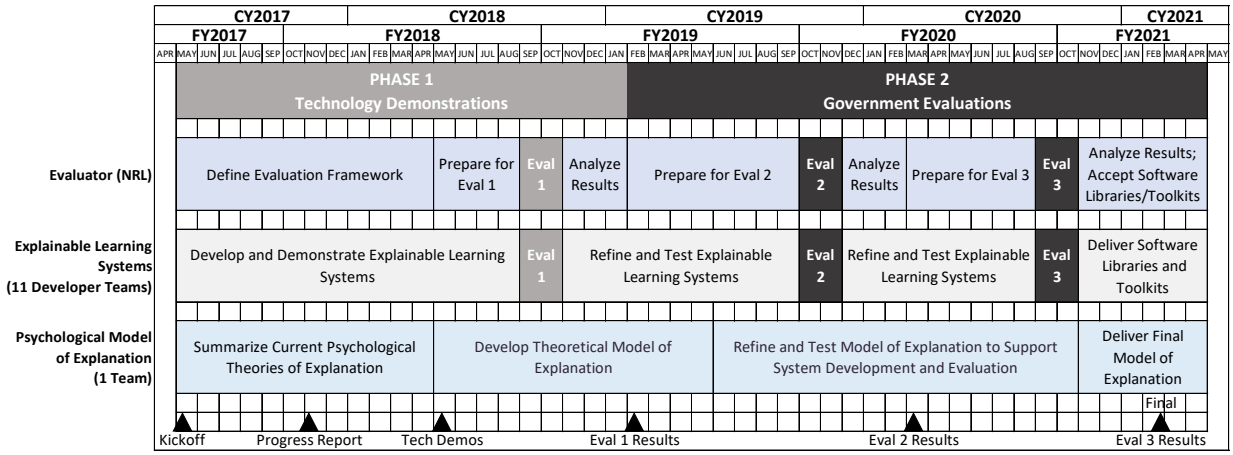
High-level (across all eleven system developer teams) measures of explanation effectiveness achieved to date are as follows (as of February 2019):

- Phase 1 – XAI system-generated explanations increase user satisfaction, improve user mental model, engender appropriate user trust/reliance, and improve user-system task performance; and
- Phase 2 – not applicable/nothing to report - Phase 2 is underway as of March 2019.

Approved for Public Release, Distribution Unlimited

Program Schedule and Milestones

Figure 2. XAI Program Schedule and Milestones



Evaluation

The XAI program Evaluator (Naval Research Laboratory) is the government team tasked with evaluation of XAI systems to measure explanation effectiveness and ML model performance. The Evaluator team is currently (as of April 2019) in the process of summarizing the Phase 1 results from the various developer teams, based on their individual challenge problems, user studies, and evaluations with the prototype XAI systems.

The Evaluator team is also currently in the process of defining milestones and requirements leading up to Evaluation 2 (e.g., developer team challenge problem plan, pilot study design, and pilot study execution and analysis).

Approved for Public Release, Distribution Unlimited