DARPA's Explainable Artificial Intelligence Program

David Gunning, David W. Aha

■ *Dramatic success in machine learning has led to a new wave of AI applications (for example, transportation, security, medicine, finance, defense) that offer tremendous benefits but cannot explain their decisions and actions to human users. DARPA's explainable artificial intelligence (XAI) program endeavors to create AI systems whose learned models and decisions can be understood and appropriately trusted by end users. Realizing this goal requires methods for learning more explainable models, designing effective explanation interfaces, and understanding the psychologic requirements for effective explanations. The XAI developer teams are addressing the first two challenges by creating 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 psychologic theories of explanation to help the XAI evaluator define a suitable evaluation framework, which the developer teams will use to test their systems. The XAI teams completed the first of this 4-year program in May 2018. In a series of ongoing evaluations, the developer teams are assessing how well their XAM systems' explanations improve user understanding, user trust, and user task performance.*

Advances in machine learning (ML) techniques promise to produce AI systems that perceive, learn, decide, and act on their own. However, they will be unable to explain their decisions and actions to human users. This lack is especially important for the Department of Defense, whose challenges require developing more intelligent, autonomous, and symbiotic systems. Explainable AI will be essential if users are to understand, appropriately trust, and effectively manage these artificially intelligent partners. To address this, DARPA launched its explainable artificial intelligence (XAI) program in May 2017. DARPA defines explainable AI as AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future. Naming this program explainable AI (rather than interpretable, comprehensible, or transparent AI, for example) reflects DARPA's objective to create more human-understandable AI systems through the use of effective explanations. It also reflects the XAI team's interest in the human psychology of explanation, which draws on the vast body of research and expertise in the social sciences.