DARPA's Explainable AI (XAI) program: A retrospective

David Gunning[1], Eric Vorm[2], Yunyan Wang[3], and Matt Turek[1]

[1]DARPA
[2]US Naval Research Laboratory
[3]Quantitative Scientific Solutions

November 15, 2021

Abstract

DARPA formulated the Explainable Artificial Intelligence (XAI) program in 2015 with the goal to enable end users to better understand, trust, and effectively manage artificially intelligent systems. In 2017, the four-year XAI research program began. Now, as XAI comes to an end in 2021, it is time to reflect on what succeeded, what failed, and what was learned. This article summarizes the goals, organization, and research progress of the XAI Program.

Dave Gunning, Eric Vorm, Jennifer Wang, Matt Turek

Abstract

DARPA formulated the Explainable Artificial Intelligence (XAI) program in 2015 with the goal to enable end users to better understand, trust, and effectively manage artificially intelligent systems. In 2017, the four-year XAI research program began. Now, as XAI comes to an end in 2021, it is time to reflect on what succeeded, what failed, and what was learned. This article summarizes the goals, organization, and research progress of the XAI Program.

Creation of XAI

Dramatic success in machine learning has created an explosion of new Artificial Intelligence (AI) capabilities. Continued advances promise to produce autonomous systems that perceive, learn, decide, and act on their own. These systems offer tremendous benefits, but their effectiveness will be limited by the machine's inability to explain its decisions and actions to human users. This issue is especially important for the United States Department of Defense (DoD), which faces challenges that require the development of more intelligent, autonomous, and reliable systems. Explainable AI will be essential for users to understand, appropriately trust, and effectively manage this emerging generation of artificially intelligent partners.

The problem of explainability is, to some extent, the result of AI's success. In the early days of AI, the predominant reasoning methods were logical and symbolic. These early systems reasoned by performing some form of logical inference on (somewhat) human readable symbols. Early systems could generate a trace of their inference steps, which could then become the basis for explanation. As a result, there was significant work on how to make these systems explainable (Shortliffe & Buchanan, 1975; Swartout, Paris, & Moore, 1991; Johnson, 1994; Lacave & Díez, 2002; Van Lent, Fisher, & Mancuso, 2004).

Yet these early AI systems were ineffective; they proved too expensive to build and too brittle against the complexities of the real world. Success in AI came as researchers developed new machine learning techniques that could construct models of the world using their own internal representations (e.g., support vectors, random forests, probabilistic models, and neural networks). These new models were much more effective, but necessarily more opaque and less explainable.

2015 was an inflection point in the need for XAI. Data analytics and machine learning had just experienced a decade of rapid progress (Jordan & Mitchell, 2015). The deep learning revolution had just begun, following the breakthrough ImageNet demonstration in 2012 (Krizhevsky, Sutskever, & Hinton, 2012). The popular press was alive with animated speculation about Superintelligence (Bostrom, 2014) and the coming AI Apocalypse (Gibbs, 2017, Cellan-Jones, 2014, Marr, 2018). Everyone wanted to know how to understand, trust, and manage these mysterious, seemingly inscrutable, AI systems.

2015 also saw the emergence of initial ideas for providing explainability. Some researchers were exploring deep learning techniques, such as the use of deconvolutional networks to visualize the layers of convolutional networks (Zeiler & Fergus, 2014). Other researchers were pursuing techniques to learn more interpretable models, such as Bayesian Rule Lists (Letham, Rudin, McCormick, & Madigan, 2015). Others were developing model-agnostic techniques that could experiment with a machine learning model—as a black box—to infer an approximate, explainable model, such as LIME (Ribeiro, Singh, & Guestrin, 2016). Yet others were evaluating the psychological and human-computer interaction aspects of the explanation interface. (Kulesza, Burnett, Wong, & Stumpf, 2015)

DARPA spent a year surveying researchers, analyzing possible research strategies, and formulating the goals and structure of the program. In August 2016, DARPA released DARPA-BAA-16-53 to call for proposals.

XAI Program Goals

The stated goal of Explainable Artificial Intelligence (XAI) was*to create a suite of new or modified machine learning techniques that produce explainable models that, when combined with effective explanation techniques, enable end users to understand, appropriately trust, and effectively manage the emerging generation of AI systems* .

The target of XAI was an end user who depends on decisions or recommendations produced by an AI system, or actions taken by it, and therefore needs to understand the system's rationale. For example, an intelligence analyst who receives recommendations from a big data analytics system needs to understand why it recommended certain activity for further investigation. Similarly, an operator who tasks an autonomous system needs to understand the system's decision-making model to appropriately use it in future missions. The XAI concept was to provide users with explanations that enable them to understand the system's overall strengths and weaknesses; convey an understanding of how it will behave in future/different situations; and perhaps permit users to correct the system's mistakes.

The XAI program assumed an inherent tension between machine learning performance (e.g., predictive accuracy) and explainability, a concern that was consistent with the research results at the time. Often the highest performing methods (e.g., deep learning) were the least explainable and the most explainable (e.g., decision trees) were the least accurate. The program hoped to create a portfolio of new machine learning and explanation techniques to provide future practitioners with a wider range of design options covering the performance-explainability trade space. If an application required higher performance, the XAI portfolio would include more explainable, high performing, deep learning techniques. If an application required more explainability, XAI would include higher performing, interpretable models.

XAI Program Structure

The program was organized into three major technical areas (TAs), as illustrated in Figure 1: (1) the development of new XAI machine learning and explanation techniques for generating effective explanations;

(2) understanding the psychology of explanation by summarizing, extending and applying psychological theories of explanation; and (3) evaluation of the new XAI techniques in two challenge problem areas: data analytics and autonomy.

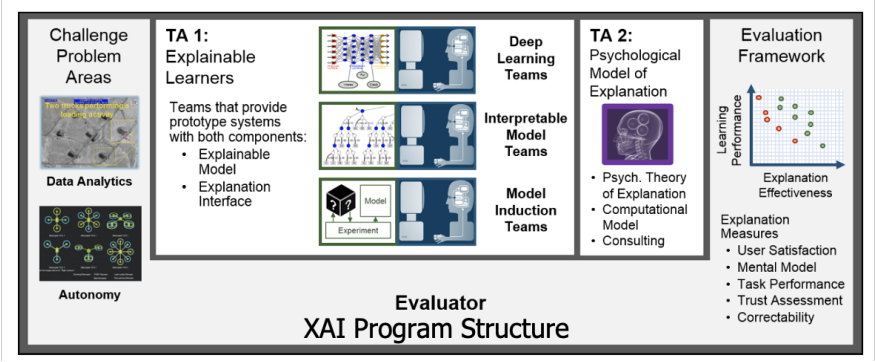

Figure 1: DARPA XAI Program Structure, including technical areas (TAs) and evaluation framework

The original program schedule consisted of two phases: Phase 1, Technology Demonstrations (18 months); and Phase 2, Comparative Evaluations (30 months). During Phase 1, developers were asked to demonstrate their technology against their own test problems. During Phase 2, the original plan was to have developers test their technology against one of two common problems (Figure 2) defined by the government evaluator. At the end of Phase 2, the developers were expected to contribute prototype software to an open source XAI toolkit.

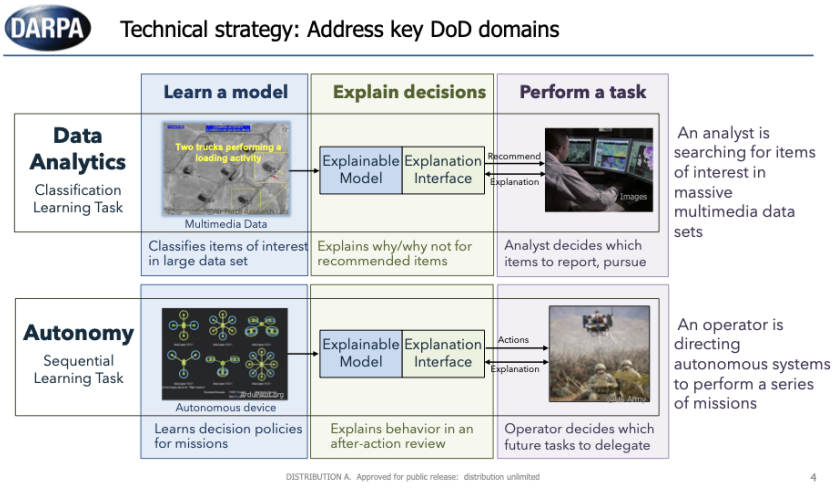

Figure 2: Challenge Problem Areas

XAI Program Development

In May, 2017, XAI development began. Eleven research teams were selected to develop the Explainable Learners (TA1) and one team was selected to develop the Psychological Models of Explanation. Evaluation was provided by the Naval Research Lab. The following summarizes those developments and the final state of this work at the end of the program. An interim summary of the XAI developments at the end of 2018 is given in Gunning and Aha, 2019.

XAI Explainable Learner Approaches

The program anticipated that researchers would examine the training process, model representations, and, importantly, explanation interfaces. Three general approaches were envisioned for model representations. Interpretable model approaches would seek to develop ML models that were inherently more explainable and more introspectable for machine learning experts. Deep explanation approaches would leverage deep learning or hybrid deep learning approaches to produce explanations in addition to predictions. Finally, model induction techniques would create approximate explainable models from more opaque, black-box models. Explanation interfaces were expected to be a critical element of XAI, connecting a user to the model to enable them to understand and interact with the decision making process.

As the research progressed, eleven XAI teams explored a number of machine learning approaches, such as tractable probabilistic models (Roy et al. 2021) and causal models (Druce et al. 2021) and explanation techniques such as state machines generated by reinforcement learning algorithms (Koul et al. 2019, Danesh et al. 2021), Bayesian teaching (Yang et al. 2021), visual saliency maps (Petsiuk 2021, Li et al. 2021, Ray et al. 2021, Alipour et al. 2021, Vasu et al. 2021), and network and GAN dissection (Ferguson et al. 2021). Perhaps the most challenging and most unique contributions came from the combination of machine learning and explanation techniques to conduct well-designed psychological experiments to evaluate explanation effectiveness.

As the program progressed, we also gained a more refined understanding of the spectrum of users and development timeline (Figure 3).

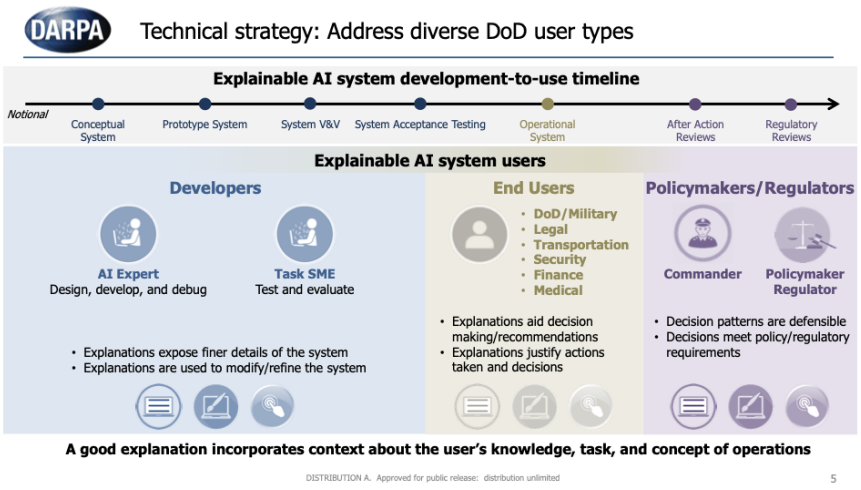

Figure 3: XAI Users and Development Timeline

Psychological Models of Explanation

The program structure anticipated the need for a grounded psychological understanding of explanation. One team was selected to summarize current psychological theories of explanation to assist the XAI developers and the evaluation team. This work began with an extensive literature survey on the psychology of explanation and previous work on explainability in AI (. . . reference for IHMC literature survey). Originally, this team was asked to (1) produce a summary of current theories of explanation, (2) develop a computational model of explanation from those theories; and (3) validate the computational model against the evaluation results from the XAI developers. Developing computational models proved to be a bridge too far, but the team did gain a deep understanding of the area and successfully produced descriptive models. These descriptive models were critical to supporting the effective evaluation approaches, which involved carefully designed

user studies, carried out in accordance with DoD human subject research guidelines. Figure 2 illustrates a top-level descriptive model of the XAI explanation process.

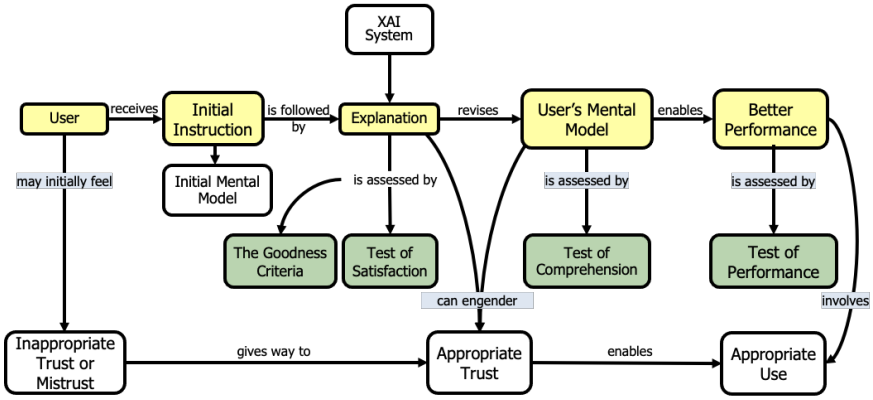

Figure 4: Psychological model of explanation. Yellow boxes illustrate the underlying process. The green boxes illustrate measurement opportunities. White boxes illustrate potential outcomes.

Evaluation

Evaluation was originally envisioned to be based on a common set of problems, within the data analytics and autonomy domains. However, it quickly became clear that it would be more valuable to explore a variety of approaches across a breadth of problem domains. In order to evaluate performance in the final year of the program, the evaluation team led by the U.S. Naval Research Laboratory (NRL) developed an explanation scoring system (ESS). Based on recommendations from a group of domain experts and validated using content validity ratio (CVR), the ESS provides a quantitative mechanism for assessing the design of a XAI user study. The ESS assesses multiple elements of the user study, including the task, domain, explanations, explanation interface, users, hypothesis, data collection, and analysis. XAI evaluation measures are shown in Figure 3, including functional measures, learning performance measures, and explanation effectiveness measures. It is critical to carefully design a user study in order to accurately evaluate the effectiveness of an explanation. Often times, multiple types of measures (cf. performance, functionality, explanation effectiveness) will be necessary to evaluate the performance of an XAI algorithm. XAI user study design can be tricky and typically the teams that were most effective on the program had colleagues with significant psychology expertise.

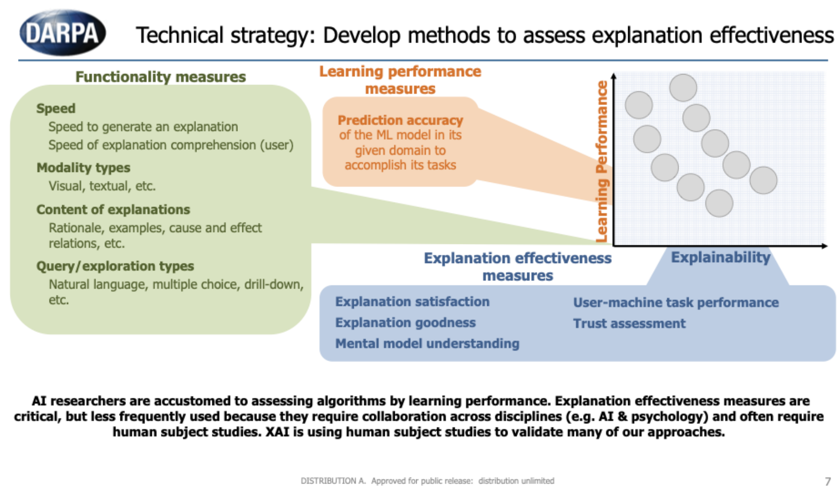

Figure 5: Evaluation measures for XAI algorithms.

XAI Development Approaches

The XAI program explored many approaches, as shown in Table 1.

Team	Approach	Application
UC Berkeley	Salience-map attention mechanisms implemented in DNNs Petsiuk 2021 Vasu et al. 2021	Saliency maps for object detectors allow users to identify the detector which will be more accurate by reviewing sample detections & maps Petsiuk 2021
	Transduction of DNN states into natural language explanations Hendricks et al. 2021	Explainable and advisable autonomous driving systems to fill in knowledge gaps. Humans can evaluate AI-generated explanations for navigation decisions. Kim et al. 2021 Watkins et al. 2021
Charles River Analytics	Causal models of deep reinforcement learning policies to enable explanation-enhanced training by answering counterfactual queries Druce et al. 2021 Witty et al. 2021	Human-machine teaming gameplay in StarCraft2
		Developed a distilled version of a pedestrian detection model, which used convolutional auto encoders to condense the activations into user understandable "chunks".

Carnegie Mellon University	Robustified classifiers with salient gradients Yeh and Ravikumar 2021	Interactive debugger interface for visualizing poisoned training datasets. Work is applied on the IARPA TrojAI dataset. Sun et al. 2021
Oregon State University	iGOS++ visual saliency algorithm Khorram et al. 2021	Debugging of COVID-19 diagnosis chest x-ray classifier
	Quantized bottleneck networks for deep RL algorithms	Understanding recurrent policy networks through extracted state machines and key decision points in video games and control Danesh et al. 2021
	Explanation analysis process for reinforcement learning systems Dodge et al. 2021	After-action review of AI decisions mirror the army's after action review system to understand why AI made its decisions to improve explainability and AI trust Mai et al. 2020
	Reinforcement learning model via embedded self-predictions	Contrastive explanations of action choices in terms of human understandable properties of future outcomes Lin et al., 2021
Rutgers University	Bayesian teaching to select examples and features from the training data to explain model inferences to a domain expert Yang et al. 2021	Interactive tool for analyzing a pneumothorax detector for chest x-rays. Targeted user study engaging ˜10 radiologists demonstrated the effectiveness of the explanations. Folke et al. 2021
Team	**Approach**	**Application**
UT Dallas	Tractable probabilistic logic models where local explanations are queries over probabilistic models and global explanations are generated using logic, probability and directed trees and graphs	Activity recognition in videos using TACoS cooking tasks and WetLab scientific lab procedure datasets. Generates explanations about whether activities are present in the video data. Chiradeep et al. 2021
PARC	Reinforcement learning implementing a hierarchical multifactor framework for decision problems.	Simulated drone flight mission planning task where users learned to predict each agent's behavior to choose the best flight plan. User study tested the usefulness of AI-generated local and global explanations in helping users predict AI behavior. Stefik et al. 2021
SRI	Spatial attention VQA (SVQA) and spatial-object attention BERT VQA (SOBERT) Ray et al. 2021 Alipour et al. 2021	Attention-based (gradCAM) explanations for MRI brain-tumor segmentation. Visual salience models for video Q&A.

Raytheon BBN	- CNN based one-shot detector, using network dissection to identify the most salient features Bau et al. 2018 - Explanations produced by heatmaps and text explanations Selvaraju et al. 2017 - Human-machine common ground modeling	-Indoor navigation with a robot (in collaboration with GA Tech) - Video Q&A - Human-assisted one-shot classification system by identifying the most salient features Ferguson et al. 2021
Texas A&M	Mimic learning methodology to detect falsified text. Yuan et al. 2021 Linder et al. 2021	News claim truth classification
UCLA	CX-ToM Framework: A new XAI framework using Theory-of-Mind where we pose explanation as an iterative communication process, i.e. dialog, between the machine and human user. In addition, we replace the standard attention based explanations with novel counterfactual explanations called fault-lines. Akula et al. 2021 Akula et al. 2020	Image Classification, Human body pose estimation.
	A learning framework to acquire interpretable knowledge representation and an Augmented Reality system for explanation interface. Edmonds et al. 2019 Liu et al. 2021	Robot learning to open medicine bottles with locks and allows user interventions to correct wrong behaviors.
	Theory of mind explanation network with multi-level belief updates from learning. Edmonds et al. 2021 (in preparation)	Minesweeper-like game to find optimal path for an agent.
IHMC	Explanation Scorecard	Evaluate the utility of an explanation. Defines seven levels of capability, from the null case of no explanation, to surface features (e.g. heat maps), to AI introspections such as choice logic, to diagnoses of the reasons for failures.
	Cognitive Tutorial	A straightforward way to help users understand complex systems is to provide a tutorial up front but the tutorial should not be restricted to how the system works. Hoffman and Clancey 2021

Stakeholder Playbook	Survey of stakeholder needs, including development team leaders, trainers, system developers and user team leaders in industry and government.
AI Evaluation Guidebook	Identifies methodological shortcomings for evaluating XAI techniques, spanning experimental design, control conditions, experimental tasks and procedures, and statistical methodologies.

Table 1: Technical approaches on DARPA's XAI program

XAI Results, Lessons Learned

Three major evaluations were conducted during the program: one during Phase 1 and two during phase 2. In order to evaluate the effectiveness of XAI techniques, researchers on the program designed and executed user studies. User studies are still the gold standard for assessing explanations. There were approximately 12,700 participants in user studies carried out by XAI researchers, including approximately 1900 supervised participants, where the individual was guided through the experiment by the research team (e.g. in person or on Zoom) and 10800 unsupervised participants, where the individual self-guided through the experiment and was not actively guided by the research team (e.g. Amazon Mechanical Turk). In accordance with policy for all US Department of Defense (DoD) funded human subjects research, each research protocol was reviewed by a local institutional review board (IRB) and then a DoD human research protection office reviewed the protocol and the local IRB findings.

In the course of those user studies, several key takeaways were identified.

- Users prefer systems that provide decisions with explanations over systems that provide only decisions. Tasks where explanations provide the most value are those where a user needs to understand the inner workings of how an AI system makes decisions. [Supported by 11 experiments across performer teams]
- In order for explanations to improve user task performance, the task must be difficult enough that the AI explanation helps. [PARC, UT Dallas]
- User cognitive load to interpret explanations can hinder user performance. Combined with the previous point, explanations and task difficulty need to be calibrated in order to improve user performance. [UCLA, Oregon State]
- Explanations are more helpful when an AI is incorrect and are particularly valuable for edge cases. [UCLA, Rutgers]
- Measures of explanation effectiveness can change over time. [Raytheon, BBN]
- Advisability can improve user trust significantly over explanations alone. [UC Berkeley]
- XAI is useful for measuring and aligning mental models for users and XAI systems. [Rutgers, SRI]
- Lastly, since the last year of XAI took place during the unprecedented times of the COVID-19 pandemic, our performer teams developed best-practices for designing web interfaces to conduct XAI user studies when in-person studies were not possible. [OSU, UCLA] Dikkala 2021

As mentioned earlier, there seemed to be a natural tension between learning performance and explainability. However, throughout the course of the program, we found evidence that explainability can improve performance (Kim et al. 2021, Watkins et al. 2021). From an intuitive perspective, training a system to

produce explanations provides additional supervision, via additional loss functions, training data, or other mechanisms, that encourages a system to learn more effective representations of the world. While this may not be true in all cases and significant work remains to characterize when explainable techniques will be more performant, it provides hope that future XAI systems can be more performant than current systems while meeting user needs for explanations.

State of the world, AI, and XAI after DARPA program in 2021

There currently is no universal solution to XAI. As discussed earlier, different user types require different types of explanations. This is no different from what we face interacting with other humans. Consider, for example, a doctor needing to explain a diagnosis to a fellow doctor, a patient, or a medical review board. Perhaps future XAI systems will be able to automatically calibrate and communicate explanations to a specific user within a large range of user types, but that is still significantly beyond the current state of the art.

One of the challenges in developing XAI is measuring the effectiveness of an explanation. DARPA's XAI effort has helped develop foundational technology in this area, but much more needs to be done, including drawing more from the human factors and psychology communities. Measures of explanation effectiveness need to be well established, well understood, and easily implemented by the developer community in order for effective explanations to become a core capability of ML systems.

UC Berkeley's result (Kim et al. 2021) demonstrating that advisability, the ability for an AI system to take advice from a user, improves user trust beyond explanations is intriguing. Certainly, users will likely prefer systems where they can quickly correct the behavior of a system in the same ways that humans can provide feedback to each other. Such advisable AI systems that can both produce and consume explanations will be key to enabling closer collaborations between humans and AI systems.

Close collaboration is required across multiple disciplines including computer science, machine learning, artificial intelligence, human factors, and psychology, among others, in order to effectively develop XAI techniques. This can be particularly challenging, as researchers tend to focus on a single domain and often need to be pushed to work across domains. Perhaps in the future a XAI-specific research discipline will be created at the intersection of multiple current disciplines. Towards this end, we have worked to create an Explainable AI Toolkit (XAITK), which collects the various program artifacts (e.g. code, papers, reports, etc.) and lessons learned from the four-year DARPA XAI program into a central, publicly accessible location (Hu et al. 2021). We believe the toolkit will be of broad interest to anyone who deploys AI capabilities in operational settings and needs to validate, characterize and trust AI performance across a wide range of real-world conditions and application areas.

Today we have a more nuanced, less dramatic, and, perhaps, more accurate understanding of AI, than we had in 2015. We certainly have a more accurate understanding of the possibilities and the limitations of deep learning. The AI apocalypse has faded from an imminent danger to a distant curiosity. Similarly, The XAI program has produced a more nuanced, less dramatic, and, perhaps, more accurate understanding of XAI. The program certainly acted as a catalyst to stimulate XAI research (both inside and outside of the program). The results have produced a more nuanced understanding of XAI uses and users, the psychology of XAI, the challenges of measuring explanation effectiveness, as well as producing a new portfolio of XAI ML and HCI techniques. There is certainly more work to be done, especially as new AI techniques are developed that will continue to need explanation. XAI will continue as an active research area for some time. The authors believe that the XAI program has made a significant contribution by providing the foundation to launch that endeavor.

References

Akula, Arjun R., et al. "CX-ToM: Counterfactual Explanations with Theory-of-Mind for Enhancing Human Trust in Image Recognition Models." *arXiv preprint arXiv:* 2109.01401 (2021) (accepted to iScience 2021).

Akula, Arjun R., et al. "CoCoX: Generating Conceptual and Counterfactual explanations via Fault-Lines." AAAI, 2020.

Bau, David, et al. "Gan dissection: Visualizing and understanding generative adversarial networks." *arXiv preprint arXiv:1811.10597*(2018).

Cellan-Jones, R. (2014). Stephen Hawking warns artificial intelligence could end mankind. *BBC news* , *2* (10), 2014.

Danesh, Mohamad H., et al. "Re-understanding Finite-State Representations of Recurrent Policy Networks." *International Conference on Machine Learning* . PMLR, 2021.

Dikkala, Rupika, et al. "Doing Remote Controlled Studies with Humans: Tales from the COVID Trenches." ACM-IEEE CHASE. 2021.

Edmonds, Mark, et al. "A tale of two explanations: Enhancing human trust by explaining robot behavior." *Science Robotics* 4.37 (2019).

Folke, Tomas, et al. "Explainable AI for medical imaging: explaining pneumothorax diagnoses with Bayesian teaching." *arXiv preprint arXiv:2106.04684* (2021).

Gibbs, S. Elon Musk leads 116 experts calling for outright ban of killer robots. *The Guardian* , *20* , 2017.

Gunning, David, and David Aha. "DARPA's explainable artificial intelligence (XAI) program." AI Magazine 40.2 (2019): 44-58.

Johnson, W. Lewis. "Agents that Learn to Explain Themselves." *AAAI* . 1994.

Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." *Science* 349.6245 (2015): 255-260.

Khorram, Saeed, Tyler Lawson, and Li Fuxin. "iGOS++ integrated gradient optimized saliency by bilateral perturbations." *Proceedings of the Conference on Health, Inference, and Learning* . 2021.

Anurag Koul, Alan Fern, and Sam Greydanus. "Learning Finite State Representations of Recurrent Policy Networks." *International Conference on Learning Representations.* 2019

Kulesza, Todd, et al. "Principles of explanatory debugging to personalize interactive machine learning." *Proceedings of the 20th international conference on intelligent user interfaces* . 2015

Lacave, Carmen, and Francisco J. Díez. "A review of explanation methods for Bayesian networks." *The Knowledge Engineering Review* 17.2 (2002): 107-127.

Letham, Benjamin, et al. "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model." *The Annals of Applied Statistics* 9.3 (2015): 1350-1371.

Zhengxian Lin, Kim-Ho Lam, and Alan Fern. "Contrastive Explanations for Reinforcement Learning via Embedded Self Predictions."

International Conference on Learning Representations.

Liu, Hangxin, et al. "Patching Interpretable And-Or Graph Knowledge Representation using Augmented Reality." Applied AI Letters 2021 (under review)

Mai, Theresa, et al. "Keeping it" organized and logical" after-action review for AI (AAR/AI)." *Proceedings of the 25th International Conference on Intelligent User Interfaces* . 2020.

Marr, Bernard. "Is Artificial Intelligence dangerous? 6 AI risks everyone should know about." *Forbes* (2018).

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* . 2016

Richmond, Sheldon. "Superintelligence: Paths, Dangers, Strategies. By Nick Bostrom. Oxford University Press, Oxford, 2014

Selvaraju, Ramprasaath R., et al. "Grad-CAM: Why did you say that?." *arXiv preprint arXiv:1611.07450* (2016).

Shortliffe, Edward H., and Bruce G. Buchanan. "A model of inexact reasoning in medicine." *Mathematical biosciences* 23.3-4 (1975): 351-379.

Sun, Mingjie, Siddhant Agarwal, and J. Zico Kolter. "Poisoned classifiers are not only backdoored, they are fundamentally broken." *arXiv preprint arXiv:2010.09080* (2020).

Swartout, William, Cecile Paris, and Johanna Moore. "Explanations in knowledge systems: Design for explainable expert systems." *IEEE Expert* 6.3 (1991): 58-64.

Van Lent, Michael, William Fisher, and Michael Mancuso. "An explainable artificial intelligence system for small-unit tactical behavior." *Proceedings of the national conference on artificial intelligence* . Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." *European conference on computer vision* . Springer, Cham, 2014.