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.

 ISSN 0738-4602

Early AI systems were predominantly logical and symbolic; they performed some form of logical inference and could provide a trace of their inference steps, which became the basis for explanation. There was substantial work on making these systems more explainable, but they fell short of user needs for comprehension (for example, simply summarizing the inner workings of a system does not yield a sufficient explanation) and proved too brittle against real-world complexities.

Recent AI success is due largely to new ML techniques that construct models in their internal representations. These include support vector machines, random forests, probabilistic graphical models, reinforcement learning (RL), and deep learning (DL) neural networks. Although these models exhibit high performance, they are opaque. As their use has increased, so has research on explainability from the perspectives of ML (Chakraborty et al. 2017; Ras et al. 2018) and cognitive psychology (Miller 2017). Similarly, many XAI-related workshops have been held recently on ML (for example, the International Conference on Machine Learning, the Conference on Neural Information Processing Systems), AI (for example, the International Joint Conference on Artificial Intelligence), and HCI (for example, the Conference on Human-Computer Interaction, Intelligent User Interfaces) conferences, as have special topic meetings related to XAI.

There seems to be an inherent tension between ML performance (for example, predictive accuracy) and explainability; often the highest-performing methods (for example, DL) are the least explainable, and the most explainable (for example, decision trees) are the least accurate. Figure 1 illustrates this with a notional graph of the performance-explainability trade-off for various ML techniques.

When DARPA formulated the XAI program, it envisioned three broad strategies to improve explainability, while maintaining a high level of learning performance, based on promising research at the time (figure 2): deep explanation, interpretable models, and model induction.

Deep explanation refers to modified or hybrid DL techniques that learn more explainable features or representations or that include explanation generation facilities. Several design choices might produce more explainable representations (for example, training data selection, architectural layers, loss functions, regularization, optimization techniques, training sequences). Researchers have used deconvolutional networks to visualize convolutional network layers, and techniques existed for associating semantic concepts with deep network nodes. Approaches for generating image captions could be extended to train a second deep network that generates explanations without explicitly identifying the original network's semantic features.

Interpretable models are ML techniques that learn more structured, interpretable, or causal models. Early examples included Bayesian rule lists (Letham et al. 2015), Bayesian program learning, learning models of causal relationships, and use of stochastic grammars to learn more interpretable structure.

Model induction refers to techniques that experiment with any given ML model— such as a black box— to infer an approximate explainable model. For example, the model-agnostic explanation system of Ribeiro et al. (2016) inferred explanations by observing and analyzing the input-output behavior of a black box model.

DARPA used these strategies to categorize a portfolio of new ML techniques and provide future practitioners with a wider range of design options covering the performance-explainability trade space.

XAI Concept and Approach

The XAI program's goal is to create a suite of new or modified ML 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 is 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 vehicle to drive a route needs to understand the system's decision-making model to appropriately use it in future missions. Figure 3 illustrates the XAI concept: 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 or different situations, and perhaps permit users to correct the system's mistakes.

This user-centered concept poses interrelated research challenges: (1) how to produce more explainable models, (2) how to design explanation interfaces, and (3) how to understand the psychologic requirements for effective explanations. The first two challenges are being addressed by the 11 XAI research teams, which are developing new ML techniques to produce explainable models, and new principles, strategies, and HCI techniques (for example, visualization, language understanding, language generation) to generate effective explanations. The third challenge is the focus of another XAI research team that is summarizing, extending, and applying psychologic theories of explanation.

The XAI program addresses two operationally relevant challenge problem areas (figure 4): data analytics (classification of events of interest in heterogeneous multimedia data) and autonomy (decision policies for autonomous systems). These areas represent two important ML problem categories (supervised learning and RL) and Department of Defense interests (intelligence analysis and autonomous systems).

The data analytics challenge was motivated by a common problem: intelligence analysts are presented with decisions and recommendations from big data

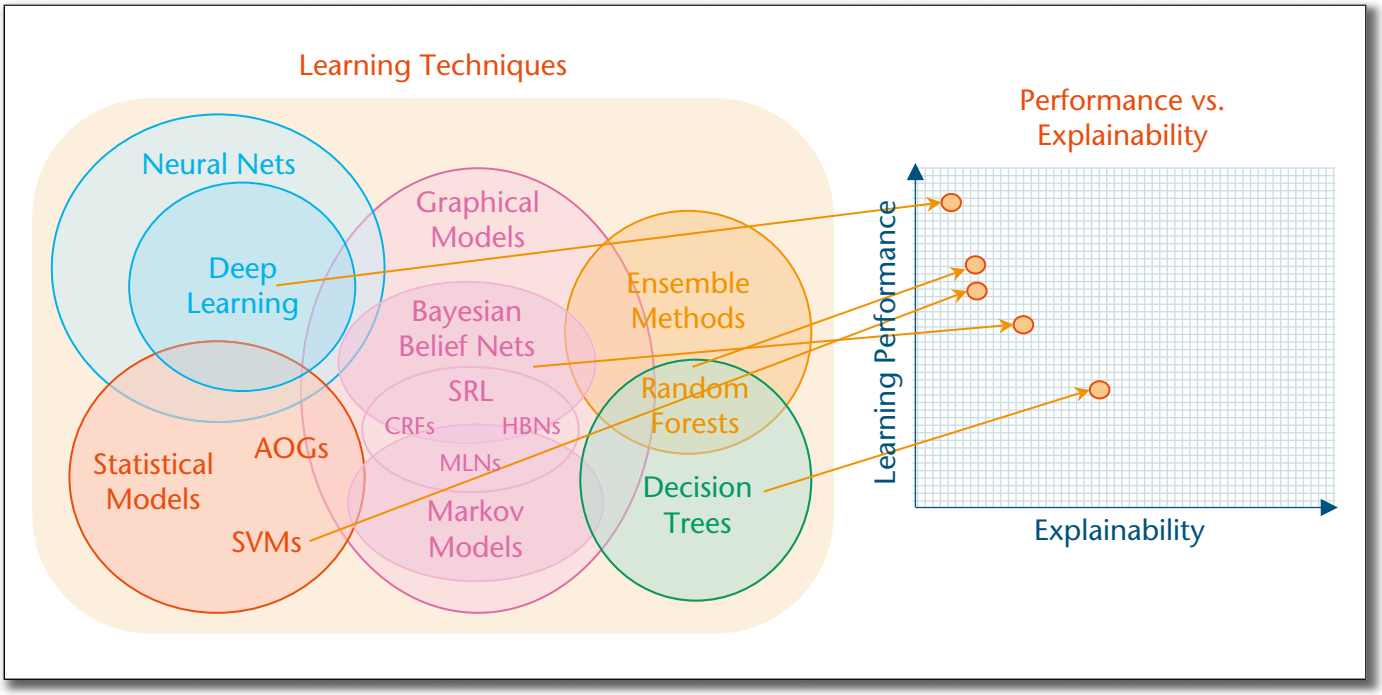


Figure 1. Learning Performance Versus Explainability Trade-Off for Several Categories of Learning Techniques.

analytics algorithms and must decide which to report as supporting evidence in their analyses and which to pursue further. These algorithms often produce false alarms that must be pruned and are subject to concept drift. Furthermore, these algorithms often make recommendations that the analyst must assess to determine whether the evidence supports or contradicts their hypotheses. Effective explanations will help confront these issues.

The autonomy challenge was motivated by the need to effectively manage AI partners. For example, the Department of Defense seeks semiautonomous systems to augment warfighter capabilities. Operators will need to understand how these behave so they can determine how and when to best use them in future missions. Effective explanations will better enable such determinations.

For both challenge problem areas, it is critical to measure explanation effectiveness. While it would be convenient if a learned model's explainability could be measured automatically, an XAI system's explanation effectiveness must be assessed according to how its explanations aid human users. This requires human-in-the-loop psychologic experiments to measure the user's satisfaction, mental model, task performance, and appropriate trust. DARPA formulated an initial explanation evaluation framework that includes potential measures of explanation effectiveness (figure 5). Exploring and refining this framework is an important part of the XAI program's research agenda.

The XAI program's goal, concept, strategies, challenges, and evaluation framework are described in the program's 2016 broad agency announcement. Figure 6 displays the XAI program's schedule, which consists of two phases. Phase 1 (18 months) commenced in May 2017 and includes initial technology demonstrations of XAI systems. Phase 2 (30 months) includes a sequence of evaluations against challenge problems selected by the system developers and the XAI evaluator. The first formal evaluations of XAI systems took place during the fall of 2018. This article describes the developer teams' progress leading up to these evaluations, whose results were presented at an XAI program meeting during the winter of 2019.

XAI Program Development and Progress

Figure 7 summarizes the 11 XAI Technical Area 1 (TA1) developer teams and the TA2 team [from the Florida Institute for Human and Machine Cognition (IHMC)] that is developing the psychologic model of explanation. Three TA1 teams are pursuing both challenge problem areas (autonomy and data analytics), three are working on only the former, and five are working on only the latter. Per the strategies described in figure 2, the TA1 teams are investigating a diverse range of techniques for developing explainable models and explanation interfaces.

Naturalistic Decision-Making Foundations of XAI

The objective of the IHMC team (which includes researchers from MacroCognition and Michigan Technological University) is to develop and evaluate psychologically plausible models of explanation and develop actionable concepts, methods, measures, and metrics for explanatory reasoning. The IHMC team is

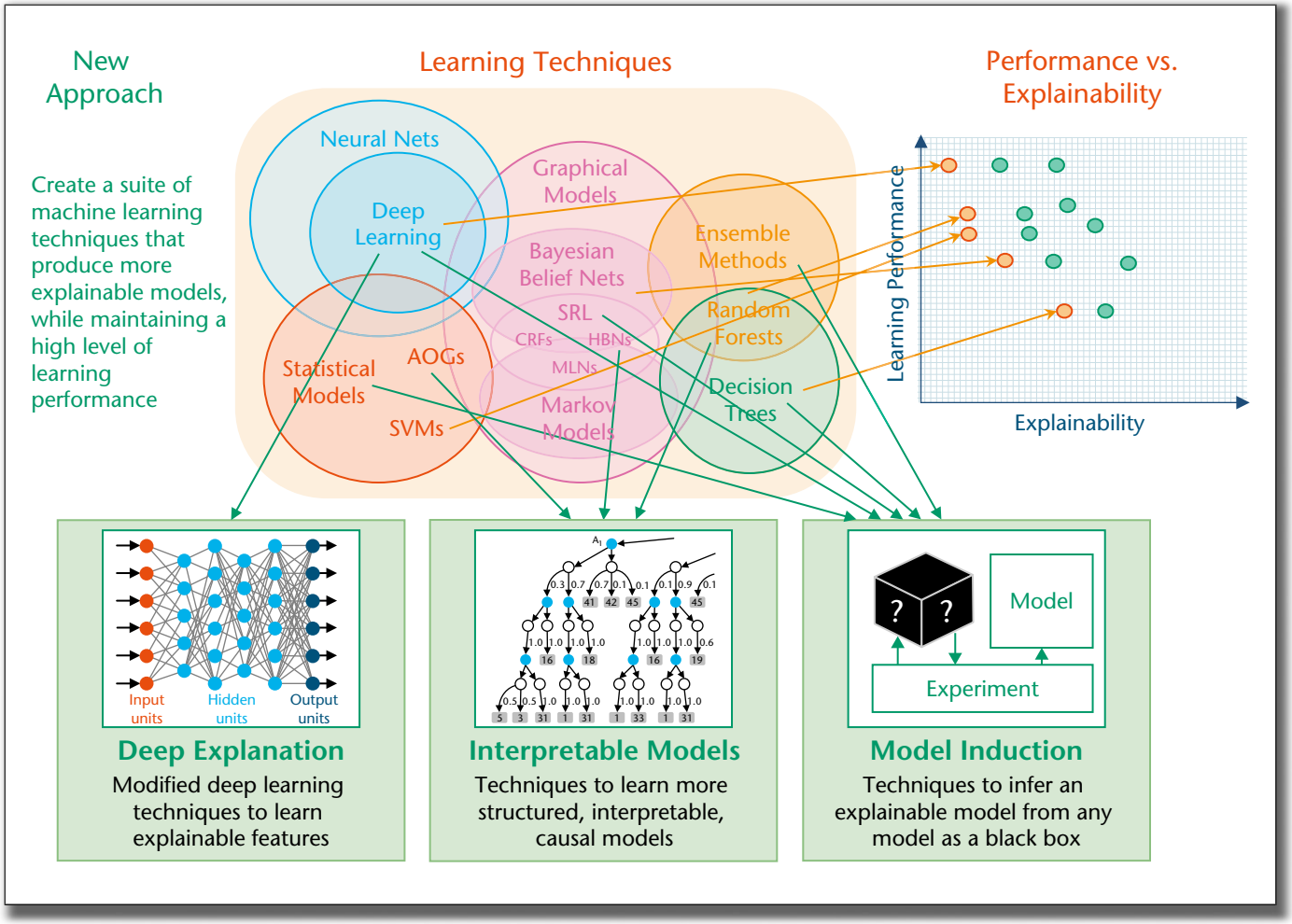


Figure 2. Strategies for Developing Explainable Models.

investigating the nature of explanation itself. What must happen for a person to be satisfied with an attempt to explain (1) the workings of a complex system and (2) why it acted the way it did in a given situation? To address such questions, the team has formulated a naturalistic model of human explanatory reasoning and is providing guidance to the performer teams on methods for evaluating the effectiveness of their XAI systems' explanations. The team reviewed the relevant literatures in philosophy of science and specializations within psychology, from which criteria were synthesized for assessing the "goodness" of explanations. The team is also collecting and analyzing a corpus of cases in which individuals create or receive explanations of the workings of complex systems.

This team developed measures for explanation goodness, a user's mental model (for example, correctness, completeness), and user task performance. From this, the user can make reasonably accurate judgments about when to trust (or doubt) the system. To gain this insight, the user must explore the decision-making processes and performance of an XAI system, especially for boundary cases, including ways in which deep neural networks (DNNs) can be spoofed. This methodology was described in a series of essays (Hoffman and Klein 2017; Hoffman et al. 2017; Klein 2018; Hoffman et al. 2018).

Figure 8 illustrates IHMC's model of the XAI explanation process, highlighting measurement categories for assessing explanation effectiveness. The user receives a recommendation or decision from an XAI system, along with an explanation that could be tested for goodness (versus preestablished criteria) and user satisfaction. The explanation contributes to the user's mental model of the AI system, which could be tested for accuracy and comprehension. The AI system's recommendations and the user's mental model may enable, or decrease, user task performance, which could also be measured. These processes contribute to the user's appropriate, or inappropriate, trust of the AI system. The XAI evaluator is using this model to test the developer teams' XAI systems.

Evaluation

The XAI program's independent government evaluator is the Naval Research Laboratory. For Phase 1, the laboratory (with IHMC's help) prepared an evaluation framework for the TA1 teams to use as a template for

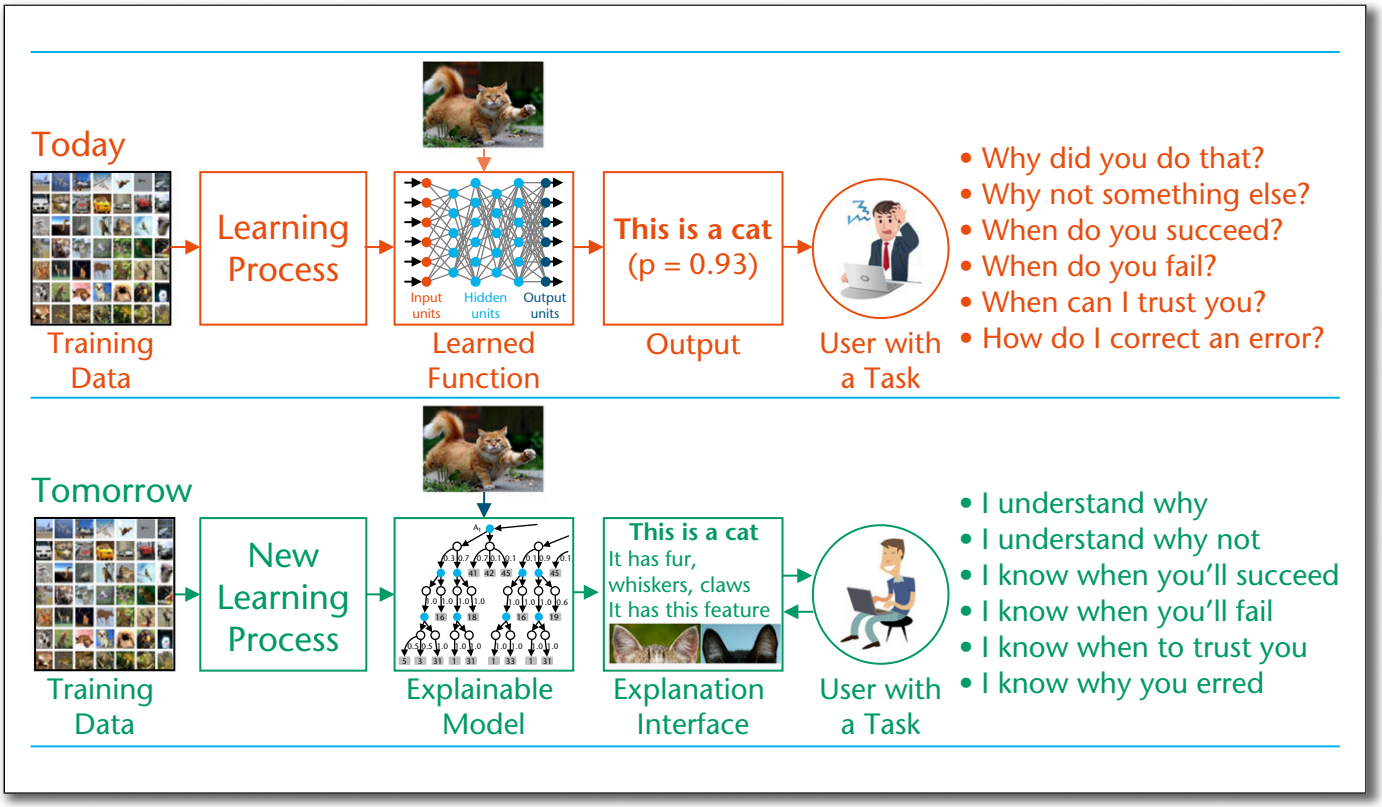


Figure 3. The XAI Concept.

designing and implementing their Phase 1 evaluation experiments, where they will select a test problem or problems in the challenge problem areas of data analytics or autonomy; apply their new ML techniques to learn an explainable model for their problems; evaluate the performance of their learned ML model (table 1); combine their learned model with their explanation interface to create their explainable learning system; conduct experiments in which users perform specified tasks using the explainable learning system; and measure explanation effectiveness by employing IHMC's model of the explanation process (figure 8) and explanation effectiveness measurement categories (table 1).

The evaluations will include the following experimental conditions: (1) without explanation: the XAI system is used to perform a task without providing explanations to the user; (2) with explanation: the XAI system is used to perform a task and generates explanations for every recommendation or decision it makes and every action it takes; (3) partial explanation: the XAI system is used to perform a task and generates only partial or ablated explanations (to assess various explanation features); and (4) control: a baseline state-of-the-art nonexplainable system is used to perform a task.

Explainable Learning Systems

Table 2 summarizes the TA1 teams' technical approaches and Phase 1 test problems.

Deeply Explainable AI

The University of California, Berkeley (UCB) team (including researchers from Boston University, the University of Amsterdam, and Kitware) is developing an AI system that is human understandable by virtue of explicit structural interpretation (Hu et al. 2017), provides post hoc (Park et al. 2018) and introspective (Ramanishka et al. 2017) explanations, has predictive behavior, and allows for appropriate trust (Huang et al. 2018). The key challenges of deeply explainable AI (DEXAI) are to generate accurate explanations of model behavior and select those that are most useful to a user. UCB is addressing the former by creating implicit or explicit explanation models: they can implicitly present complex latent representations in understandable ways or build explicit structures that are inherently understandable. These DEXAI models create a repertoire of possible explanatory actions. Because these actions are generated without any user model, they are called reflexive. For the second challenge, UCB proposes rational explanations that use a model of the user's beliefs when deciding which explanatory actions to select. UCB is also developing an explanation interface based on these innovations informed by iterative design principles.

UCB is addressing both challenge problem areas. For autonomy, DEXAI will be demonstrated in vehicle control (using the Berkeley Deep Drive data set and the CARLA simulator) (Kim and Canny 2017) and strategy game scenarios (StarCraft II). For data

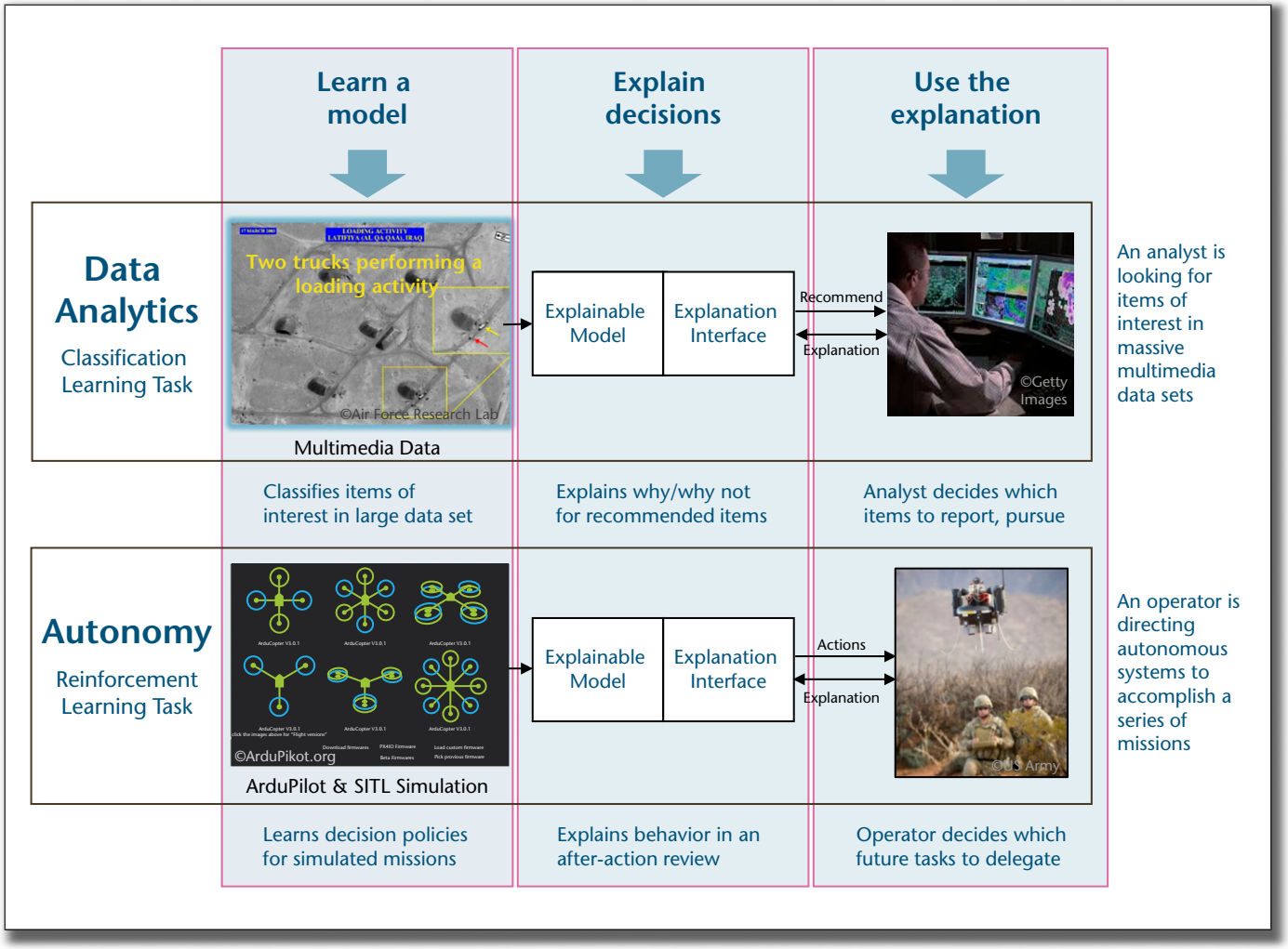


Figure 4. XAI Challenge Problem Areas.

analytics, DEXAI will be demonstrated using visual question answering (VQA) and filtering tasks (for example, using large-scale data sets such as VQA-X and ACT-X for VQA tasks and activity recognition tasks, respectively), xView, and Distinct Describable Moments (Hendricks et al. 2018).

Causal Models to Explain Learning

The goal of the Charles River Analytics (CRA) team (including researchers from the University of Massachusetts and Brown University) is to generate and present causal explanations of ML operation, through its causal models to explain learning (CAMEL) approach. CAMEL explanations are presented to a user as narratives in an interactive, intuitive interface. CAMEL includes a causal probabilistic programming framework that combines representations and learning methods from causal modeling (Marazopoulou et al. 2015) with probabilistic programming languages (Pfeffer 2016) to describe complex and rich phenomena. CAMEL can be used to describe what an ML system did, how specific data characteristics influenced its outcome, and how changing these factors would affect this outcome. Generative probabilistic models, represented in a probabilistic programming language, naturally express causal relationships; they are well suited for this task of explaining ML systems.

CAMEL probes the internal representation of an ML system to discover how it represents user-defined, natural domain concepts. It then builds a causal model of their effect on the ML system's operation by conducting experiments in which the domain concepts are systematically included or removed. CRA has applied this approach to DNNs for classification and RL.

Once learned, it uses causal probabilistic models to infer explanations of the system's predictions or actions. Because inferences can be large and complex and can contain many interacting components, CAMEL composes them into explanatory narratives that walk the user through the interactions of the major concepts and their influence on the ML system's output. The CAMEL explanation interface, based on cognitive systems engineering design principles and established

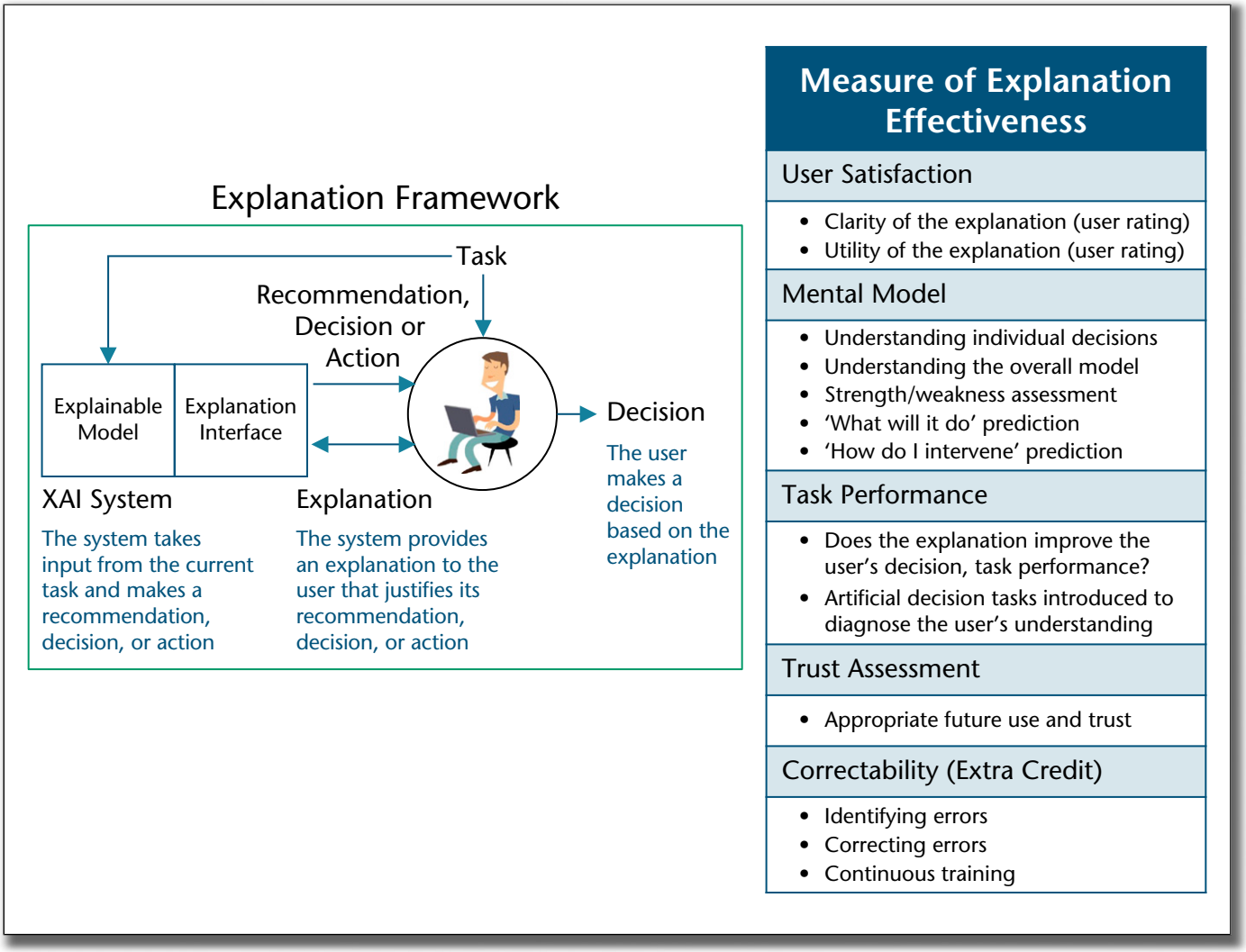


Figure 5. Evaluating Explanation Effectiveness.

HCI techniques, allows users to understand and interact with the explanatory narratives, engendering trust in automation and enabling effective user-system teamwork.

CRA is addressing both challenge problem areas. For data analytics, CAMEL has been demonstrated using pedestrian detection (using the INRIA pedestrian data set) (Harradon et al. 2018), and CRA is working toward activity recognition tasks (using ActivityNet). For autonomy, CAMEL has been demonstrated on the Atari game Amidar, and CRA is working toward demonstrating it on StarCraft II.

Learning and Communicating Explainable Representations for Analytics and Autonomy

The University of California, Los Angeles (UCLA) team (including researchers from Oregon State University and Michigan State University) is developing interpretable models that combine representational paradigms, including interpretable DNNs, compositional graphical models such as and-or graphs, and models that produce explanations at three levels (that is, compositionality, causality, and utility).

UCLA's system includes a performer that executes tasks on multimodal input data and an explainer that explains its perception, cognitive reasoning, and decisions to a user. The performer outputs interpretable representations in a spatial, temporal, and causal parse graph (STC-PG) for three-dimensional scene perception (for analytics) and task planning (for autonomy). STC-PGs are compositional, probabilistic, attributed, interpretable, and grounded on DNN features from images and videos. The explainer outputs an explanatory parse graph in a dialogue process (She and Chai 2017), localizes the relevant subgraph in the STC-PG, and infers the user's intent.

The system represents explanations at three levels: (1) concept compositions, represented by parse graph fragments that depict how information is aggregated from its constituents and contexts, how decisions are made at nodes under uncertainty, and the decision's

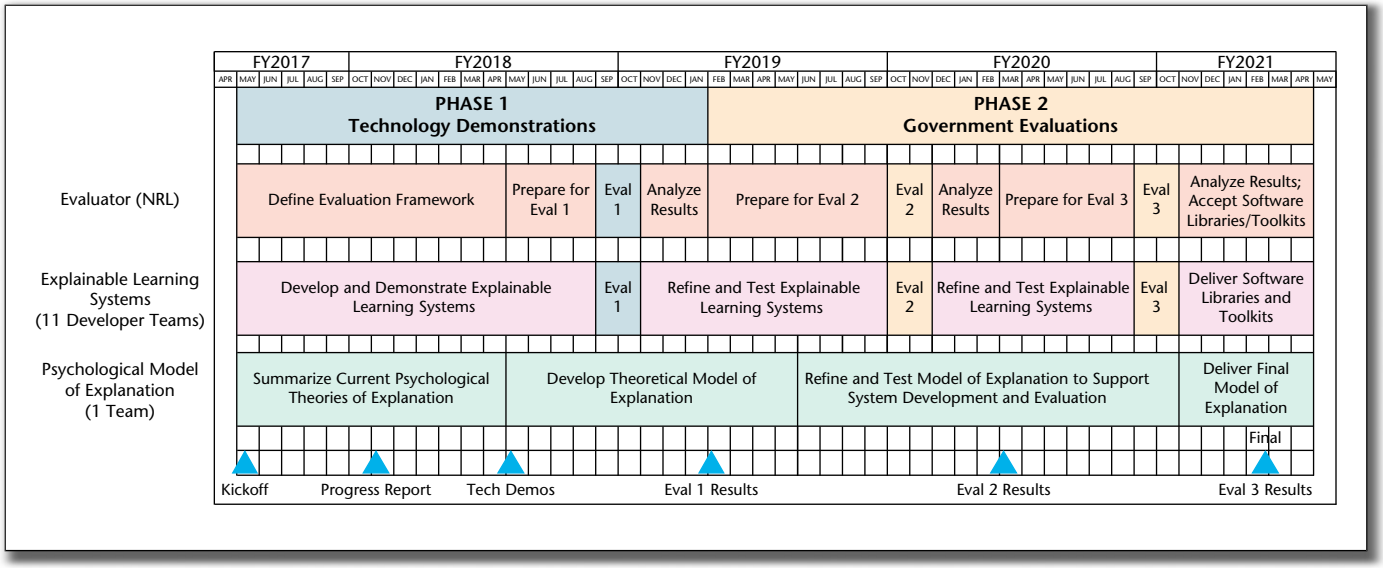


Figure 6. XAI Program Schedule.

confidence levels; (2) causal and counterfactual reasoning, realized by extracting causal diagrams from STC-PGs to predict what will happen if certain alternative actions had been taken; and (3) utility explanations, which explain why the system made certain decisions.

UCLA is addressing both XAI challenge problem areas using a common framework of representation and inference. For data analytics, UCLA demonstrated their system using a network of video cameras for scene understanding and event analysis. For autonomy, UCLA demonstrated it in scenarios using robots executing tasks in physics-realistic virtual reality platforms and autonomous vehicle driving game engines.

Explanation-Informed Acceptance Testing of Deep Adaptive Programs

Oregon State University (OSU) is developing tools for explaining learned agents that perform sequential decision making and is identifying best principles for designing explanation user interfaces. OSU's explainable agent model employs explainable deep adaptive programs (xDAPs), which combine adaptive programs, deep RL, and explainability. With xDAPs, programmers can create agents by writing programs that include choice points, which represent decisions that are automatically optimized via deep RL through simulator interaction. For each choice point, deep RL attaches a trained deep decision neural network (dNN), which can yield high performance but is inherently unexplainable.

After initial xDAP training, xACT trains an explanation neural network (Qi and Li 2017) for each dNN. These provide a sparse set of explanation features (x-features) that encode properties of a dNN's decision logic. Such x-features, which are neural networks, are not initially human interpretable. To address this, xACT enables domain experts to attach interpretable descriptions to x-features, and xDAP programmers to annotate environment reward types and other concepts, which are automatically embedded into the dNNs as "annotation concepts" during learning.

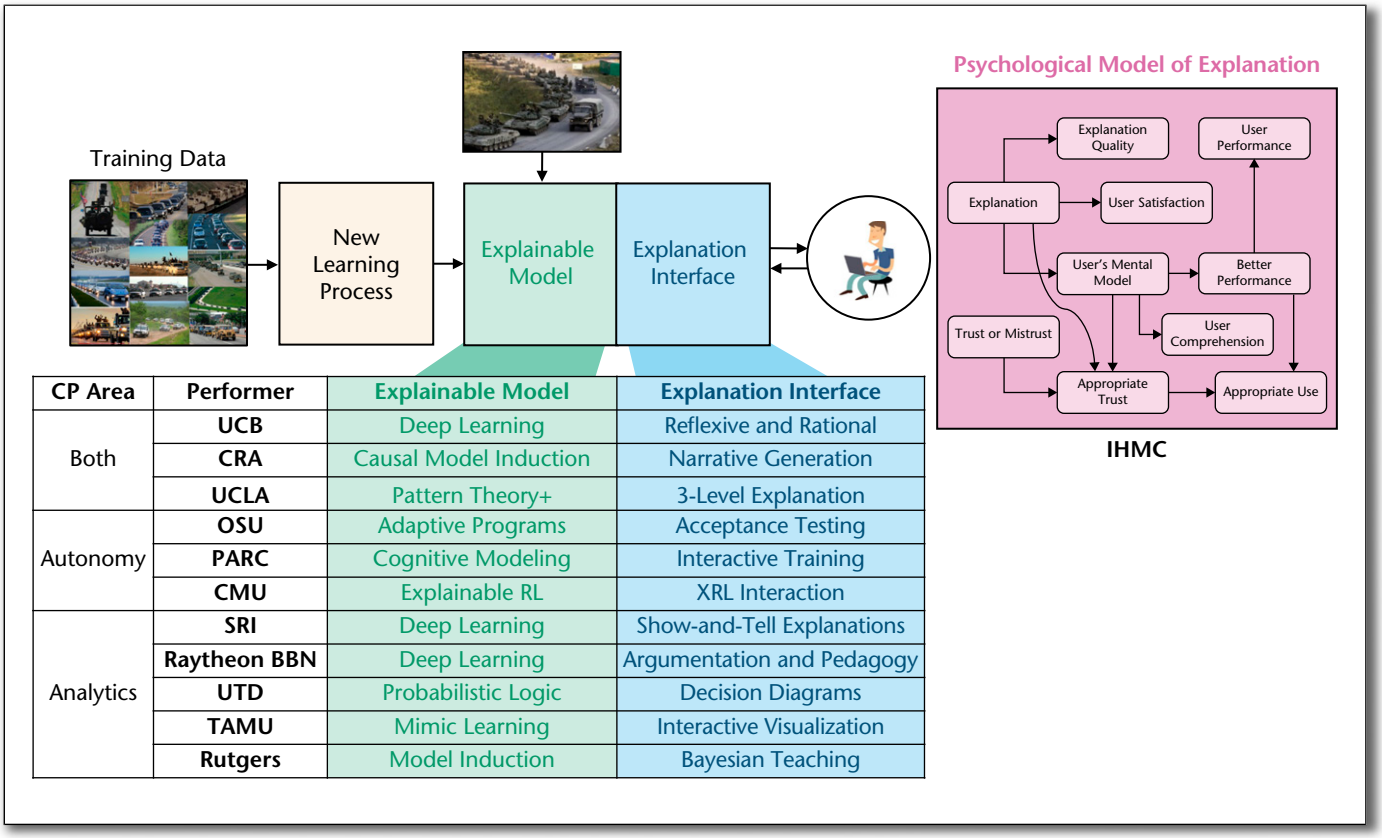
The dNN decisions can be explained via the descriptions of relevant x-features and annotation concepts, which can be further understood via neural network saliency visualization tools. OSU is investigating the utility of saliency computations for explaining sequential decision making.

OSU's explanation user interface allows users to navigate thousands of learned agent decisions and obtain visual and natural language (NL) explanations. Its design is based on information foraging theory (IFT), which allows a user to efficiently drill down to the most useful explanatory information at any moment. The assessment of rationales for learned decisions may more efficiently identify flaws in the agent's decision making and improve user trust.

OSU is addressing the autonomy challenge problem area and has demonstrated xACT in scenarios using a custom-built real-time strategy game engine. Pilot studies have informed the explanation user interface design by characterizing how users navigate AI-agent game play and tend to explain game decisions (Dodge et al. 2018).

Common Ground Learning and Explanation

The Palo Alto Research Center (PARC) team (including researchers from Carnegie Mellon University, the Army Cyber Institute, the University of Edinburgh, and the University of Michigan) is developing an interactive sensemaking system that can explain the learned capabilities of an XAI system that controls a simulated unmanned aerial system.



CP Area	Performer	Explainable Model	Explanation Interface
Both	UCB	Deep Learning	Reflexive and Rational
	CRA	Causal Model Induction	Narrative Generation
	UCLA	Pattern Theory+	3-Level Explanation
Autonomy	OSU	Adaptive Programs	Acceptance Testing
	PARC	Cognitive Modeling	Interactive Training
	CMU	Explainable RL	XRL Interaction
Analytics	SRI	Deep Learning	Show-and-Tell Explanations
	Raytheon BBN	Deep Learning	Argumentation and Pedagogy
	UTD	Probabilistic Logic	Decision Diagrams
	TAMU	Mimic Learning	Interactive Visualization
	Rutgers	Model Induction	Bayesian Teaching

Figure 7. XAI Research Teams.

An XAI system's explanations should communicate what information it uses to make decisions, whether it understands how things work, and its goals. To address this, PARC's common ground learning and explanation (COGLE) and its users establish common ground about what terms to use in explanations and their meaning. This is enabled by PARC's introspective discourse model, which interleaves learning and explaining processes.

Performing tasks in the natural world is challenging for autonomous systems, requiring experience to create enough knowledge for reliable high performance. COGLE employs K-models that encode an AI agent's domain-specific task knowledge. K-models organize this knowledge into levels of elements, where higher (lower) levels model actions with longer range (local) effects. They support a competency-based framework that informs and guides the learning and testing of XAI systems.

COGLE's multilayer architecture partitions its information processing into sensemaking, cognitive modeling, and learning. The learning layer employs capacity constrained recurrent and hierarchical DNNs to produce abstractions and compositions over the states and actions of unmanned aerial systems to support an understanding of generalized patterns. It combines learned abstractions to create hierarchical, transparent policies that match those learned by the system. The cognitive layer bridges human-usable symbolic representations to the abstractions, compositions, and generalized patterns.

COGLE's explanation interfaces support performance review, risk assessment, and training. The first provides a map that traces an unmanned aerial systems' mission actions and divides the action or decision (flight) path into explainable segments. The second interface's tools enable users to examine and assess the system's competencies and make predictions about mission performance.

COGLE will be demonstrated in ArduPilot's Software-in-the-Loop Simulator and a discretized abstract simulation test bed. It will be evaluated by drone operators and analysts. Competency-based evaluation will help PARC to determine how best to develop appropriate domain understandable models.

Explainable Reinforcement Learning

Carnegie Mellon University is creating a new discipline of explainable RL to enable dynamic human-machine interaction and adaptation for maximum team performance. This effort has two goals: to develop new methods for learning inherently explainable RL policies and to develop strategies that can explain existing black-box policies. For the former, Carnegie Mellon is developing methods to improve model learning for RL agents to capture the benefits

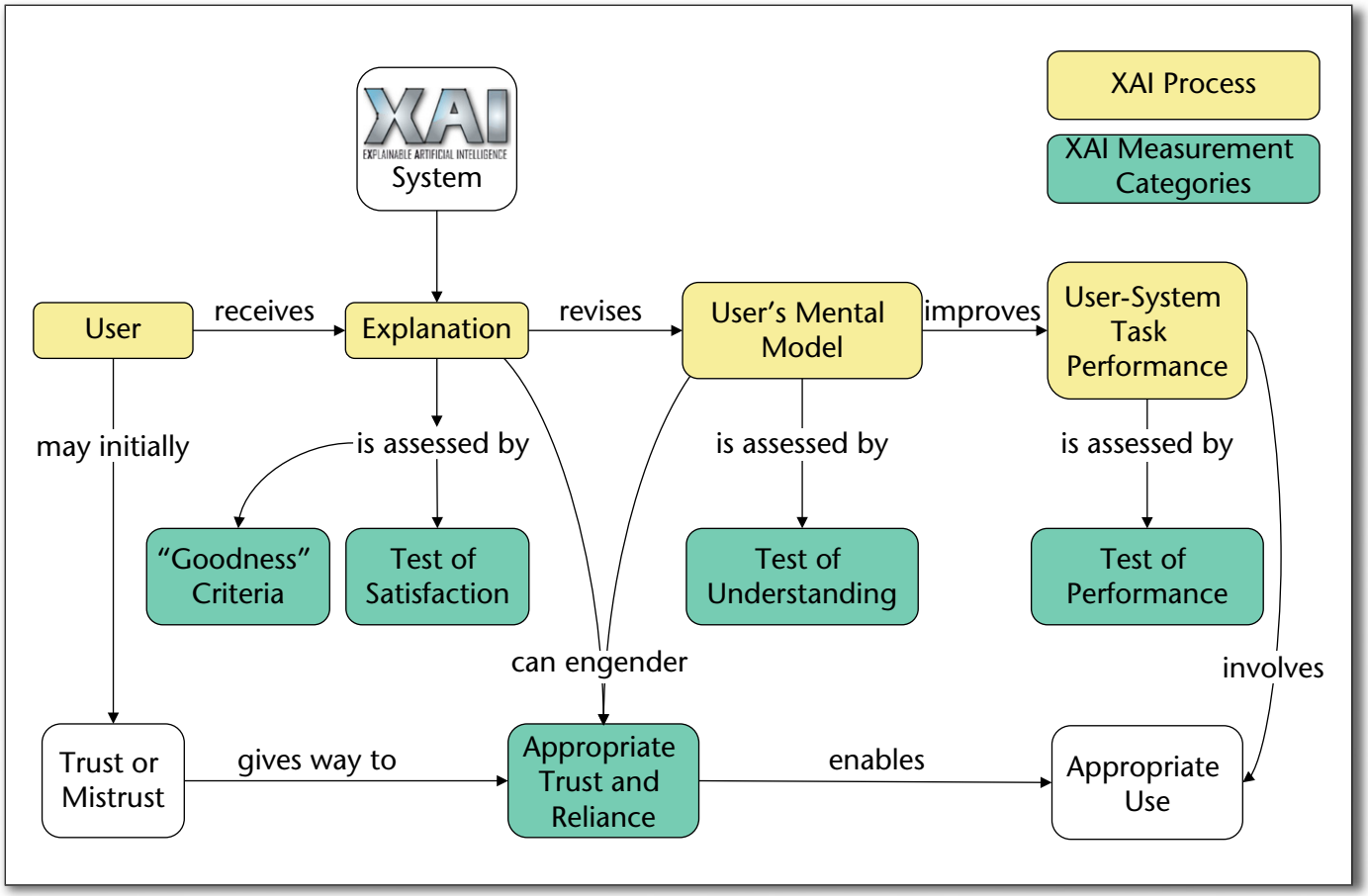


Figure 8. Initial Model of the Explanation Process and Explanation Effectiveness Measurement Categories.

of model-based approaches (ability to visualize plans in the internal model space), while integrating the benefits of model-free approaches (simplicity and higher ultimate performance). These include methods that incrementally add states and actions to world models after discovering relevant latent information, learn models via end-to-end training of complex model-based optimal control policies, learn general DL models that directly integrate and exploit rigid body physics (Belbute-Peres and Kolter 2017), and learn understandable predictive state representations using recurrent architectures (Hefny et al. 2018).

Carnegie Mellon is also developing methods that can explain the actions and plans of black-box RL agents, observed either online or from system logs. This involves answering questions such as, why did an agent choose a particular action? or, what training data were most responsible for this choice? To achieve this, Carnegie Mellon developed techniques that generate NL descriptions of agents from behavior logs and detect outliers or anomalies. Carnegie Mellon also developed improvements over traditional influence function methods in DL, allowing its XRL system to precisely identify the portions of a training set that most influence a policy's outcome.

Carnegie Mellon is addressing the autonomy challenge problem area and has demonstrated XRL in several scenarios, including OpenAI Gym, Atari games, autonomous vehicle simulation, mobile service robots, and self-improving educational software and games.

Deep Attention-Based Representations for Explanation/ Explainable Generative Adversarial Networks

SRI International's team (including researchers from the University of Toronto, the University of Guelph, and the University of California, San Diego) is developing an explainable ML framework for multimodal data analytics that generates show-and-tell explanations with justifications of decisions accompanied by visualizations of input data used to generate inferences.

The deep attention-based representations for explanationxplainable generative adversarial networks (DARE/X-GANS) system employs DNN architectures inspired by attentional models in visual neuroscience. It identifies, retrieves, and presents evidence to a user as part of an explanation. The attentional mechanisms provide a user with a means for system probing and collaboration.

Measure	Description
ML Model performance	
Various measures (on a per-challenge problem area basis)	Accuracy/performance of the ML model in its given domain (to understand whether performance improved or degraded relative to state-of-the-art nonexplainable baselines)
Explanation Effectiveness	
Explanation goodness	Features of explanations assessed against criteria for explanation goodness
Explanation satisfaction	User's subjective rating of explanation completeness, usefulness, accuracy, and satisfaction
Mental model understanding	User's understanding of the system and the ability to predict the system's decisions/behavior in new situations
User task performance	Success of the user performing the tasks for which the system is designed to support
Appropriate Trust and Reliance	User's ability to know when to, and when not to, trust the system's recommendations and decisions

Table 1. Measurement Categories.

DARE/X-GANS uses generative adversarial networks (GANs), which learn to understand data by creating it, while learning representations with explanatory power. GANs are made explainable by using interpretable decoders that map unsupervised clusters onto parts-based representations. This involves generating visual evidence, given text queries, using text-to-parts generation (Vicol et al. 2018), the parts being interpretable features such as human poses or bounding boxes. This evidence is then used to search the queried visual data.

The system presents explanations of its answers based on visual concepts extracted from the multimodal data inputs and knowledge base queries. Given explanatory questions, it provides justifications, visual evidence used for decisions, and visualizations of the system's inner workings. This show-and-tell explanation interface ensures highly intuitive explanations, made possible by attentional modules that localize evidence used for each visual task. Initial studies demonstrate that such explanations substantially improve user task performance.

SRI is addressing the data analytics challenge problem area and has demonstrated DARE/X-GANs using VQA and multimodal QA tasks with image and video data sets.

Explainable Question Answering System

The Raytheon BBN Technologies team (including researchers from the Georgia Institute of Technology, MIT, and the University of Texas, Austin) is developing a system that answers unrestricted NL questions posed by users about multimedia data and provides interactive, explorable explanations of why it derived an answer.

The explainable question answering system (EQUAS) learns explainable DNN models in which internal structures (for example, individual neurons) have been aligned to semantic concepts (for example, wheels and handlebars) (Zhou et al. 2015). This allows neural activations within the network, during a decision process, to be translated to NL explanations (for example, "this object is a bicycle because it has two wheels and handlebars"). EQUAS also uses neural visualization techniques to highlight input regions associated with neurons that most influenced its decisions. To express case-based explanations, EQUAS retains indexes and retrieves cases from its training data that support its choices. Rejected alternatives are recognized and ruled out using contrastive language, visualization, and examples. Four explanation modalities map to key elements of argument construction and interactive pedagogy: didactic statements, visualizations, cases, and rejections of alternative choices.

The EQUAS explanation interface allows users to explore the explanation space populated by these explanation modes. It enables iterative and guided collaborative interaction, allowing users to drill down into the supporting evidence from each explanation category.

Raytheon BBN is addressing the analytics challenge problem area and has demonstrated initial EQUAS capabilities on VQA tasks for images, exploring how different explanation modalities enable users to understand and predict the behavior of the underlying VQA system.

Team	Explainable Model	Explanation Interface	Challenge Problem
UC Berkeley	• Post hoc explanations by training additional DL models • Explicit introspective explanations (NMNs) • Reinforcement learning (informative rollouts, explicit modular agent)	• Reflexive explanations (arise from the model) • Rational explanations (come from reasoning about user's beliefs)	• **Autonomy:** vehicle control (BDD-X, CARLA), strategy games (StarCraft II) • **Analytics:** visual QA and filtering tasks (VQA-X, ACT-X, xView, DiDeMo, etc.)
Charles River Analytics	Experiment with the learned model to team an explainable, causal, probabilistic programming model	Interactive visualization based on the generation of temporal, spatial narratives from the causal, probabilistic models	• **Autonomy:** Atari, StarCraft II • **Analytics:** pedestrian detection (INRIA), activity recognition (ActivityNet)
UCLA	• Interpretable representations: STC-AOG (spatial, temporal, causal models), STC-PG (scene and event interpretations in analytics), STC-PG+ (task plans in autonomy) • Theory of mind representations (user's beliefs, user's mental model of agent)	• Three-level explanation concept compositions, causal and counterfactual reasoning, utility explanation • Explanation representations: X-AOG (explanation model), X-PG (explanatory parse graph as dialogue), X-Utility (priority and loss for explanations)	• **Autonomy:** robot executing daily tasks in physics-realistic VR platform autonomous vehicle driving (GTA5 game engine) • **Analytics:** network of video cameras for scene understanding and event analysis
Oregon State	xDAPs, combination of adaptive programs, deep learning, and explainabilty	Provides a visual and NL explanation interlace for acceptance testing by test pilots based on IFT	**Autonomy:** real-time strategy sames based on custom-designed game engine designed to support explanation; StarCraft II
PARC	Three-layer architecture: learning layer (DNNs), cognitive layer (ACT-R cognitive model), explanation layer (HCI)	• Interactive visualization of states, actions, policies, values • Module for test pilots to refine and train the system	**Autonomy:** MAVSim wrapper over ArduPilot simulation environment
Carnegie Mellon University	A new scientific discipline for XRL XRL with work on new algorithms and representations	• Interactive explanations of dynamic systems • Human-machine interaction to improve performance	**Autonomy:** OpenAI Gym, autonomy in the electrical grid, mobile service robots, self-improving educational software
SRI	Multiple DL techniques: attention-based mechanisms, compositional NMNs, GANs	• DNN visualization • Query evidence that explains DNN decisions • Generate NL justifications	**Analytics:** VQA (Visual Gnome, Flickr30), MovieQA
Raytheon BBN	• Semantic labeling of DNN neurons • DNN audit trail construction • Gradient-weighted class activation mapping	• Comprehensive strategy based on argumentation theory • NL generation • DNN visualization	**Analytics:** VQA for images and video
UTD	TPLMs	Enables users to explore and correct the underlying model as well as add background knowledge	**Analytics:** infer activities in multimodal data (video and text), wet lab (biology), and Textually Annotated Cooking Scenes data sets

(continued on following page)

Team	Explainable Model	Explanation Interface	Challenge Problem
Texas A&M	• Mimic learning framework combines DL models for prediction and shallow models for explanations. • Interpretable learning algorithms extract knowledge from DNNs for relevant explanations	Interactive visualization over multiple news, using heat maps and topic modeling clusters to show predictive features	**Analytics:** multiple tasks using data from Twitter, Facebook, ImageNet, and news websites
Rutgers	Select the optimal trading examples to explain model decisions based on Bayesian Teaching	Example-based explanation of the full model, user-selected substructure, user submitted examples	**Analytics:** image processing, text corpora, VQA, movie events

Table 2. Summary of Explainable Learning System Developer Approaches and Selected Phase 1 Test Problems.

Tractable Probabilistic Logic Models: A New, Deep Explainable Representation

The University of Texas at Dallas (UTD) team (including researchers from UCLA, Texas A&M, and the Indian Institute of Technology, Delhi) is developing a unified approach to XAI using tractable probabilistic logic models (TPLMs).

TPLMs are a family of representations that include (for example) decision trees, binary decision diagrams, cutset networks, sentential decision diagrams, first-order arithmetic circuits, and tractable Markov logic (Gogate and Domingos 2016). The UTD system extends TPLMs to generate explanations of query results; handle continuous variables, complex constraints, and unseen entities; compactly represent complex objects such as parse trees, lists, and shapes; and enable efficient representation and reasoning about time.

For scalable inference, the system uses novel algorithms to answer complex explanation queries using techniques including lifted inference, variational inference, and their combination. For fast and increased learning accuracy, it uses discriminative techniques, deriving algorithms that compose NNs and support vector machines with TPLMs, using interpretability as a bias to learn more interpretable models. These approaches are then extended to handle real-world situations.

The UTD explanation interface displays interpretable representations with multiple related explanations. Its interactive component allows users to debug a model and suggest alternative explanations.

UTD is addressing the analytics challenge problem area and has demonstrated its system for recognizing human activities in multimodal data (video and text), such as the Textually Annotated Cooking Scenes data set.

Transforming Deep Learning to Harness the Interpretability of Shallow Models: An Interactive End-to-End System

The Texas A&M University (TAMU) team (including researchers from Washington State University) is developing an interpretable DL framework that uses mimic learning to leverage explainable shallow models and facilitates domain interpretation with visualization and interaction. Mimic learning bridges the gap between deep and shallow models and enables interpretability. The system also mines informative patterns from raw data to enhance interpretability and learning performance.

The system's interpretable learning algorithms extract knowledge from DNNs for relevant explanations. Its DL module connects to a pattern-generation module by leveraging the interpretability of the shallow models. The learning system's output is displayed to users with visualization including coordinated and integrated views.

The TAMU system handles image (Du et al. 2018) and text (Gao et al. 2017) data and is being applied to the XAI analytics challenge problem area. It provides effective interpretations of detected inaccuracies from diverse sources while maintaining a competitive detection performance. The TAMU system combines model-level (that is, model transparency) and instance-level (that is, instance explanation) interpretability to generate explanations that are more easily comprehended by users. This system has been deployed on multiple tasks using data from Twitter, Facebook, ImageNet, CIFAR-10, online health care forums, and news websites.

Model Explanation by Optimal Selection of Teaching Examples

Rutgers University is extending Bayesian teaching to enable automatic explanation by selecting the data

subset that is most representative of a model's inference. Rutgers' approach allows for explanation of the inferences of any probabilistic generative and discriminative model, as well as influential DL models (Yang and Shafto 2017).

Rutgers is also developing a formal theory of human-machine cooperation and supporting interactive guided explanation of complex compositional models. Common among these is a core approach of building from models of human learning to foster explainability and carefully controlled behavioral experiments to quantify explainability.

Explanation by Bayesian teaching inputs a data set, a probabilistic model, and an inference method and returns a small subset of examples that best explains the model's inference. Experiments with unfamiliar images show that explanations of inferences about categories of (and specific) images increase the accuracy of people's reasoning about a model (Vong et al. 2018). Experiments with familiar image categories show that explanations allow users to accurately calibrate trust in model predictions.

Explanation of complex models is facilitated by interactive guided explanations. By exploiting compositionality and cooperative modifications of ML models, Rutgers provides a generic approach to fostering understanding via guided exploration. Interaction occurs through an interface that exposes model structure and explains each component with aspects of the data. The Rutgers approach has been demonstrated to facilitate understanding of large text corpora, as assessed by a human's ability to accurately summarize the corpus after short, guided explanations.

Rutgers is addressing the data analytics challenge problem area and has demonstrated its approach on images, text, combinations of these (for example, VQA), and structured simulations involving temporal causal structure.

Conclusions and Future Work

DARPA's XAI program is developing and evaluating a wide variety of new ML techniques: modified DL techniques that learn explainable features; methods that learn more structured, interpretable, causal models; and model induction techniques that infer an explainable model from any black-box model. One year into the XAI program, initial technology demonstrations and results indicate that these three broad strategies merit further investigation and will provide future developers with design options covering the performance versus explainability trade space. The developer teams' XAI systems are being evaluated to assess the value of explanations that they provide, localizing the contributions of specific techniques within this trade space.

Acknowledgments

The authors thank the XAI development teams, specifically their principle investigators, for their innovative research and contributions to this article: Trevor Darrell (UCB), Brian Ruttenberg and Avi Pfeffer (CRA), Song-Chun Zhu (UCLA), Alan Fern (OSU), Mark Stefik (PARC), Zico Kolter (Carnegie Mellon), Mohamed Amer and Giedrius Burachas (SRI International), Bill Ferguson (Raytheon BBN), Vibhav Gogate (UTD), Xia (Ben) Hu (TAMU), Patrick Shafto (Rutgers), and Robert Hoffman (IHMC). The authors owe a special thanks to Marisa Carrera for her exceptional technical support to the XAI program and her extensive editing skills.

References

Belbute-Peres, F., and Kolter, J. Z. 2017. A Modular Differentiable Rigid Body Physics Engine. Paper presented at the Neural Information Processing Systems Deep Reinforcement Learning Symposium. Long Beach, CA, December 7.

Chakraborty, S.; Tomsett, R.; Raghavendra, R.; Harborne, D.; Alzantot, M.; Cerutti, F.; and Srivastava, M.; et al. 2017. Interpretability of Deep Learning Models: A Survey of Results. Presented at the IEEE Smart World Congress 2017 Workshop: DAIS 2017 — Workshop on Distributed Analytics Infrastructure and Algorithms for Multi-Organization Federations, San Francisco, CA, August 4–8. doi.org/10.1109/UIC-ATC.2017.8397411

Dodge, J.; Penney, S.; Hilderbrand, C.; Anderson, A.; and Burnett, M. 2018. How the Experts Do It: Assessing and Explaining Agent Behaviors in Real-Time Strategy Games. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery. doi.org/10.1145/3173574.3174136

Du, M.; Liu, N.; Song, Q.; and Hu, X. 2018. Towards Explanation of DNN-Based Prediction and Guided Feature Inversion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1358–67. New York: Association for Computing Machinery. doi.org/10.1145/3219819.3220099

Gao, J.; Liu, N.; Lawley, M.; and Hu, X. 2017. An Interpretable Classification Framework for Information Extraction from Online Healthcare Forums. *Journal of Healthcare Engineering*: 2460174. doi.org/10.1155/2017/2460174

Gogate, V., and Domingos, P. 2016. Probabilistic Theorem Proving. *Communications of the ACM* 59(7): 107–15. doi.org/10.1145/2936726

Harradon, M.; Druce, J.; and Ruttenberg, B. 2018. Causal Learning and Explanation of Deep Neural Networks via Autoencoded Activations. arXiv preprint. arXiv:1802.00541v1 [cs.AI]. Ithaca, NY: Cornell University Library.

Hefny, A.; Marinho, Z.; Sun, W.; Srinivasa, S.; and Gordon, G. 2018. Recurrent Predictive State Policy Networks. In *Proceedings of the 35th International Conference on Machine Learning*, 1954–63. International Machine Learning Society.

Hendricks, L. A.; Hu, R.; Darrell, T.; and Akata, Z. 2018. Grounding Visual Explanations. Presented at the European Conference of Computer Vision (ECCV). Munich, Germany; September 8–14. doi.org/10.1007/978-3-030-01216-8_17

Hoffman, R.; Miller, T.; Mueller, S. T.; Klein, G.; and Clancey, W. J. 2018. Explaining Explanation, Part 4: A Deep Dive on Deep Nets. *IEEE Intelligent Systems* 33(3): 87–95. doi.org/10.1109/MIS.2018.033001421

Hoffman, R. R.; and Klein, G. 2017. Explaining Explanation, Part 1: Theoretical Foundations. *IEEE Intelligent Systems* 32(3): 68–73. doi.org/10.1109/MIS.2017.54

Hoffman, R. R., Mueller, S. T.; and Klein, G. 2017. Explaining Explanation, Part 2: Empirical Foundations. *IEEE Intelligent Systems* 32(4): 78–86. doi.org/10.1109/MIS.2017.3121544

Hu, R.; Andreas, J.; Rohrbach, M.; Darrell, T.; and Saenko, K. 2017. Learning to Reason: End-to-End Module Networks for Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 804–13. New York: IEEE. doi.org/10.1109/ICCV.2017.93

Huang, S. H.; Bhatia, K.; Abbeel, P.; and Dragan, A. 2018. Establishing Appropriate Trust via Critical States. Presented at the 13th Annual ACM/IEEE International Conference on Human-Robot Interaction Workshop on Explainable Robot Behavior. Madrid, Spain; October 1–5. doi.org/10.1109/IROS.2018.8593649

Kim, J., and Canny, J. 2017. Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention. In *Proceedings of the International Conference on Computer Vision*, 2942–50. New York: IEEE. doi.org/10.1109/ICCV.2017.320

Klein, G. 2018. Explaining Explanation, Part 3: The Causal Landscape. *IEEE Intelligent Systems* 33(2): 83–88. doi.org/10.1109/MIS.2018.022441353

Letham, B.; Rudin, C.; McCormick, T. H.; and Madigan, D. 2015. Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model. *Annals of Applied Statistics* 9(3): 1350–71. doi.org/10.1214/15-AOAS848

Marazopoulou, K.; Maier, M.; and Jensen, D. 2015. Learning the Structure of Causal Models with Relational and Temporal Dependence. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, 572–81. Association for Uncertainty in Artificial Intelligence.

Miller, T. 2017. Explanation in Artificial Intelligence: Insights from the Social Sciences. arXiv preprint. arXiv:1706.07269v1 [cs.AI]. Ithaca, NY: Cornell University Library.

Park, D. H.; Hendricks, L. A.; Akata, Z.; Rohrbach, A.; Schiele, B.; Darrell, T.; and Rohrbach, M. 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. New York: IEEE. doi.org/10.1109/CVPR.2018.00915

Pfeffer, A. 2016. *Practical Probabilistic Programming*. Greenwich, CT: Manning Publications.

Qi, Z., and Li, F. 2017. Learning Explainable Embeddings for Deep Networks. Paper presented at the NIPS Workshop on Interpreting, Explaining and Visualizing Deep Learning. Long Beach, CA, December 9.

Ramanishka, V.; Das, A.; Zhang, J.; and Saenko, K. 2017. Top-Down Visual Saliency Guided by Captions. In *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition*, 7206–15. New York, IEEE.

Ras, G.; van Gerven, M.; and Haselager, P. 2018 Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. arXiv preprint. arXiv:1803.07517v2 [cs.AI]. Ithaca, NY: Cornell University Library.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44. New York: Association for Computing Machinery. doi.org/10.1145/2939672.2939778

She, L., and Chai, J. Y. 2017. Interactive Learning for Acquisition of Grounded Verb Semantics towards Human-Robot Communication. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 1, 1634–44. Stroudsburg, PA: Association for Computation Linguistics. doi.org/10.18653/v1/P17-1150

Vicol, P.; Tapaswi, M.; Castrejon, L.; and Fidler, S. 2018.MovieGraphs: Towards Understanding Human-Centric Situations from Videos. In *IEEE Conference on Computer Vision and Pattern Recognition*. New York: IEEE. doi.org/10.1109/CVPR.2018.00895

Vong, W.-K.; Sojitra, R.; Reyes, A.; Yang, S. C.-H.; and Shafto, P. 2018. Bayesian Teaching of Image Categories. Paper presented at the 40th Annual Meeting of the Cognitive Science Society (CogSci). Madison, WI, July 25–28.

Yang, S. C.-H., and Shafto, P. 2017. Explainable Artificial Intelligence via Bayesian Teaching. Paper presented at the 31st Conference on Neural Information Processing Systems Workshop on Teaching Machines, Robots and Humans. Long Beach, CA, December 9.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2015. Object Detectors Emerge in Deep Scene CNNs. Paper presented at the International Conference on Learning Representations. San Diego, CA, May 7–9.

David Gunning is a program manager in DARPA's Information Innovation Office, as an Intergovernmental Personnel Act from the Pacific Northwest National Labs. Gunning has more than 30 years of experience in developing AI technology. In prior DARPA tours he managed the PAL project that produced Siri and the CPOF project that the US Army adopted as its C2 system for use in Iraq and Afghanistan. Gunning was also a program director at PARC, a senior research manager at Vulcan, senior vice president at SET Corporation, vice president of Cycorp, and a senior scientist at the Air Force Research Labs. Gunning holds an MS in computer science from Stanford University and an MS in cognitive psychology from the University of Dayton.

David W. Aha is acting director of NRL's Navy Center for Applied Research in AI in Washington, D.C. His interests include goal reasoning, XAI, case-based reasoning, and machine learning, among other topics. He has coorganized many events on these topics (for example, the IJCAI-18 XAI Workshop), launched the UCI Repository for ML Databases, served as an AAAI Councilor, and leads DARPA XAI's evaluation team.