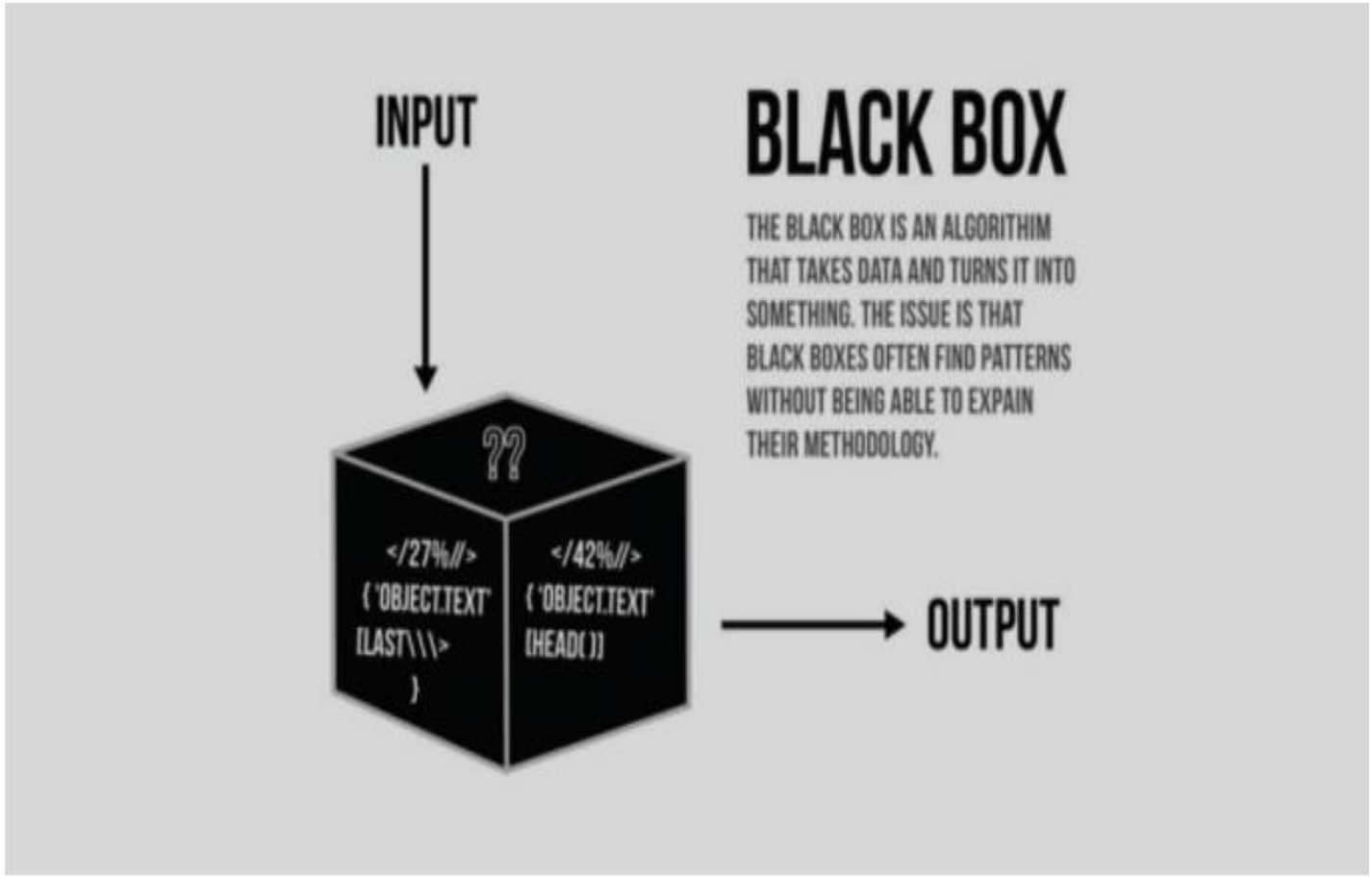
# Outline

- Interpretability = Transparency ≠ Explainability?
- DARPA XAI Grant program
- Results of the DARPA XAI Grant program
- Key takeaways and discussions

https://www.topbots.com/interpretable-machine-learning/




> *If you can't explain it simply, you don't understand it well enough.*
>
> — ***Albert Einstein***

**Table 2**
Overall picture of the classification of ML models attending to their level of explainability.

| Model | Transparent ML Models | | | Post-hoc analysis |
|---|---|---|---|---|
| | Simulatability | Decomposability | Algorithmic Transparency | |
| Linear/Logistic Regression | Predictors are human readable and interactions among them are kept to a minimum | Variables are still readable, but the number of interactions and predictors involved in them have grown to force decomposition | Variables and interactions are too complex to be analyzed without mathematical tools | Not needed |
| Decision Trees | A human can simulate and obtain the prediction of a decision tree on his/her own, without requiring any mathematical background | The model comprises rules that do not alter data whatsoever, and preserves their readability | Human-readable rules that explain the knowledge learned from data and allows for a direct understanding of the prediction process | Not needed |
| K-Nearest Neighbors | The complexity of the model (number of variables, their understandability and the similarity measure under use) matches human naive capabilities for simulation | The amount of variables is too high and/or the similarity measure is too complex to be able to simulate the model completely, but the similarity measure and the set of variables can be decomposed and analyzed separately | The similarity measure cannot be decomposed and/or the number of variables is so high that the user has to rely on mathematical and statistical tools to analyze the model | Not needed |
| Rule Based Learners | Variables included in rules are readable, and the size of the rule set is manageable by a human user without external help | The size of the rule set becomes too large to be analyzed without decomposing it into small rule chunks | Rules have become so complicated (and the rule set size has grown so much) that mathematical tools are needed for inspecting the model behaviour | Not needed |
| General Additive Models | Variables and the interaction among them as per the smooth functions involved in the model must be constrained within human capabilities for understanding | Interactions become too complex to be simulated, so decomposition techniques are required for analyzing the model | Due to their complexity, variables and interactions cannot be analyzed without the application of mathematical and statistical tools | Not needed |
| Bayesian Models | Statistical relationships modeled among variables and the variables themselves should be directly understandable by the target audience | Statistical relationships involve so many variables that they must be decomposed in marginals so as to ease their analysis | Statistical relationships cannot be interpreted even if already decomposed, and predictors are so complex that model can be only analyzed with mathematical tools | Not needed |
| Tree Ensembles | ✗ | ✗ | ✗ | Needed: Usually *Model simplification* or *Feature relevance* techniques |
| Support Vector Machines | ✗ | ✗ | ✗ | Needed: Usually *Model simplification* or *Local explanations* techniques |
| Multi–layer Neural Network | ✗ | ✗ | ✗ | Needed: Usually *Model simplification*, *Feature relevance* or *Visualization* techniques |
| Convolutional Neural Network | ✗ | ✗ | ✗ | Needed: Usually *Feature relevance* or *Visualization* techniques |
| Recurrent Neural Network | ✗ | ✗ | ✗ | Needed: Usually *Feature relevance* techniques |

# Why need explanation?

- **”Right to an explanation”,** suggested e.g. in the GDPR art. 13, 14, and 22, regarding automated decision making? (Goodman & Flaxman, 2017)

- Social issues in explainability (Felten, 2017):
  - **Confidentiality** (e.g. algorithm is trade secret or security risk)
  - **Complexity** (e.g. well understood but highly complex to give
  - **Unreasonablenss** (e.g. the use of rationally justified information to make decision that are not reasonable, discriminative or unfair)
  - **Injustice** (e.g. inconsistent with legal or moral code)

Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI (Arrieta, et.al., 2020)

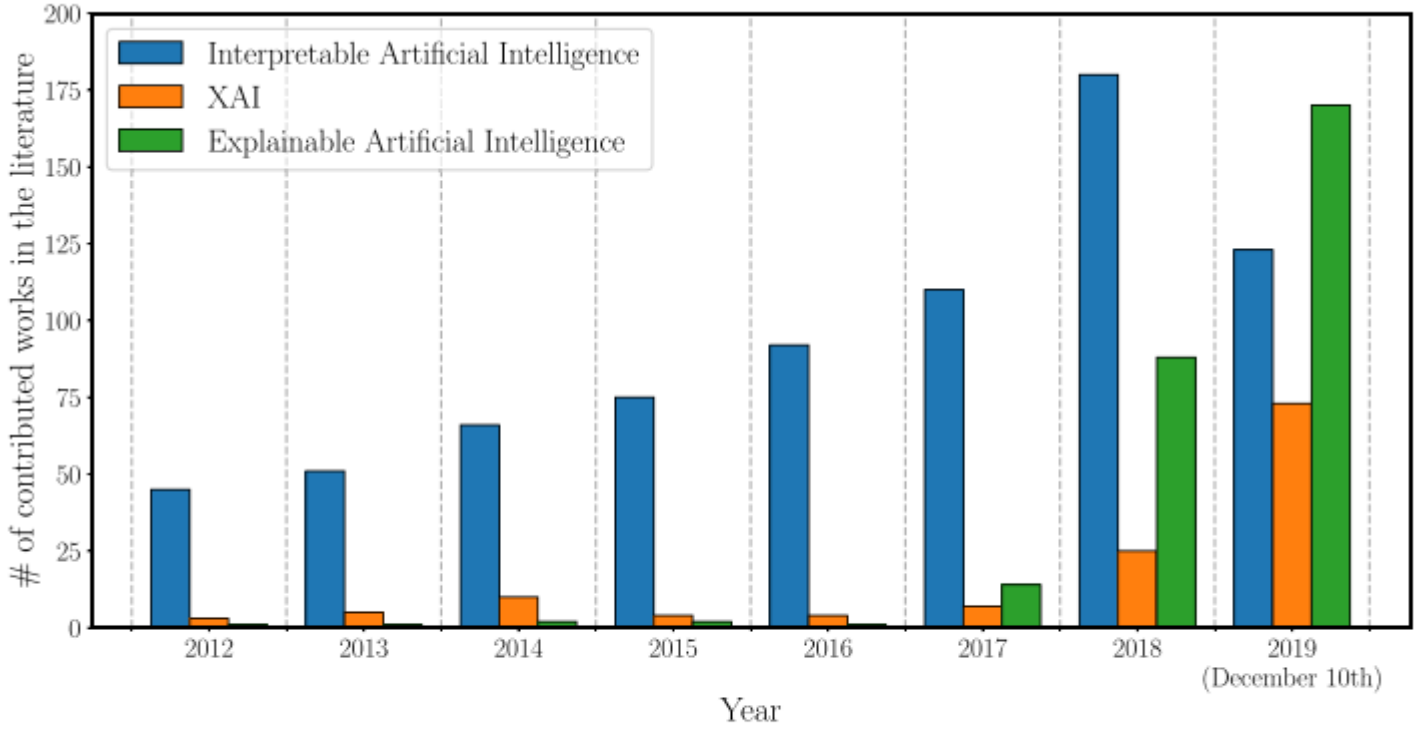



**Fig. 1.** Evolution of the number of total publications whose title, abstract and/or keywords refer to the field of XAI during the last years. Data retrieved from Scopus® (December 10th, 2019) by using the search terms indicated in the legend when querying this database. It is interesting to note the latent need for interpretable AI models over time (which conforms to intuition, as interpretability is a requirement in many scenarios), yet it has not been until 2017 when the interest in techniques to explain AI models has permeated throughout the research community.

Explanation in artificial intelligence: Insights from the social sciences - ScienceDirect

# Interpretability = Explainability

ELSEVIER

Artificial Intelligence
Volume 267, February 2019, Pages 1-38

## Explanation in artificial intelligence: Insights from the social sciences

Tim Miller ✉

### 2.1.5. *Interpretability and justification*

Here, we briefly address the distinction between *interpretability*, *explainability*, *justification*, and *explanation*, as used in this article; and as they seem to be used in artificial intelligence.

Lipton [103] provides a taxonomy of the desiderata and methods for interpretable AI. This paper adopts Lipton's assertion that explanation is post-hoc interpretability. I use Biran and Cotton [9]'s definition of *interpretability* of a model as: the degree to which an observer can understand the cause of a decision. Explanation is thus one mode in which an observer may obtain understanding, but clearly, there are additional modes that one can adopt, such as making decisions that are inherently easier to understand or via introspection. I equate 'interpretability' with 'explainability'.

A *justification* explains why a decision is good, but does not necessarily aim to give an explanation of the actual decision-making process [9].

https://www.sciencedirect.com/science/article/pii/S0004370218305988

# What makes a good explanation?

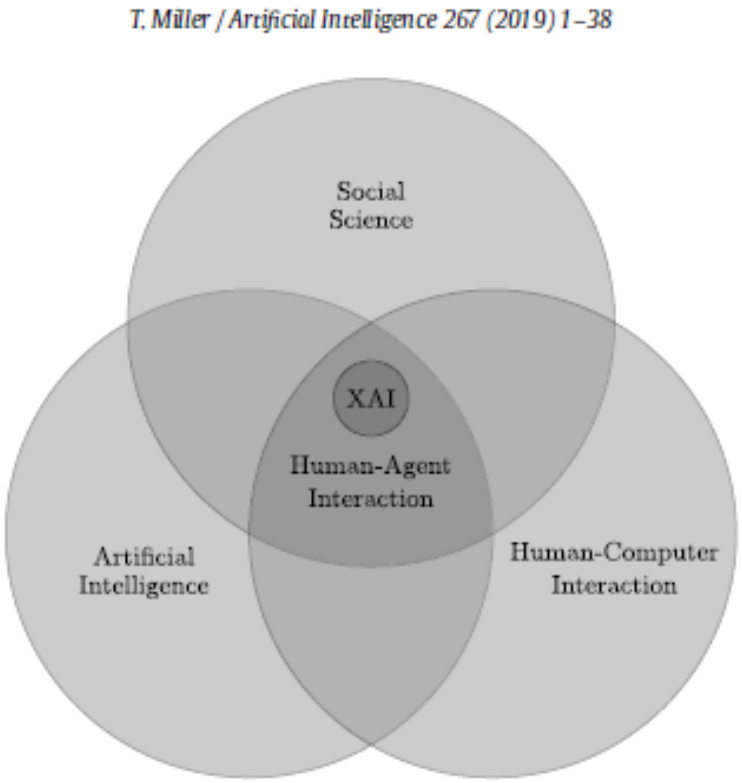



**Fig. 1.** Scope of explainable artificial intelligence.

"……models of how humans explain decisions and behaviour to each other are a good way to start analysing the problem …."

https://arxiv.org/pdf/1806.00069.pdf

# Interpretability ≠ Explainability

## Explaining Explanations: An Overview of Interpretability of Machine Learning

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter and Lalana Kagal
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
{lgilpin, davidbau, bzy, abajwa, specter, lkagal}@ mit.edu

We take the stance that interpretability alone is insufficient. In order for humans to trust black-box methods, we need *explainability* – models that are able to summarize the reasons for neural network behavior, gain the trust of users, or produce insights about the causes of their decisions. While interpetability is a substantial first step, these mechanisms need to *also* be complete, with the capacity to defend their actions, provide relevant responses to questions, and be audited. Although interpretability and explainability have been used interchangeably, we argue there are important reasons to distinguish between them. Explainable models are interpretable by default, but the reverse is not always true.

https://ojs.aaai.org//index.php/aimagazine/article/view/2850

# 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.

# Whos'who in XAI?

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8466590

Funding agency and academia

DARPA DEFENSE ADVANCED RESEARCH PROJECTS

Funding agency, interest on iIntelligent and autonomous product


"FAT" academia (fairness accountability, and transparency): promoting and enabling explainability and **fairness** in algorithmic decision-making system with social and commercial impact

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8466590

Company/business community, focusing on the demand by industry regulation


"Our mission is to build Explainable AI products and solutions that help to optimize human cognitive performance. A cornerstone of that mission is never to operate as a 'black box,'" said Welsh. "Explainable AI means that the system can justify it's reasoning. Kyndi's product exists because Deep Learning is a 'black box' and cannot be used in regulated industries where organizations are required to explain the reasons for any advice on any decision.



# Outline

- Interpretability = Transparency ≠ Explainability?
- DARPA XAI Grant program
- Results of DARPA XAI Grant program
- Key takeaways and discussions

https://www.darpa.mil/

DARPA DEFENSE ADVANCED RESEARCH PROJECTS AGENCY ABOUT US / OUR RESEARCH /

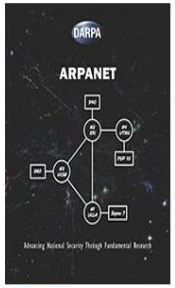
› Defense Advanced Research Projects Agency › Advancing National Security Through Fundamental Research

# Advancing National Security Through Fundamental Research

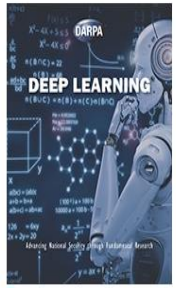
For sixty years, DARPA has held to a singular and enduring mission: to make pivotal investments in breakthrough technologies for national security.

The genesis of that mission and of DARPA itself dates to the launch of Sputnik in 1957, and a commitment by the United States that, from that time forward, it would be the initiator and not the victim of strategic technological surprises. Working with innovators inside and outside of government, DARPA has repeatedly delivered on that mission, transforming revolutionary concepts and even seeming impossibilities into practical capabilities. The ultimate results have included not only game-changing military capabilities such as precision weapons and stealth technology, but also such icons of modern civilian society such as the Internet, automated voice recognition and language translation, and Global Positioning System receivers small enough to embed in myriad consumer devices.

Preventing Pandemics



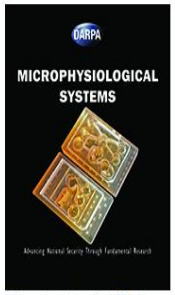
ARPANET



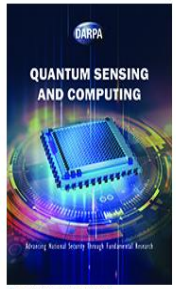
Deep Learning

EPIC

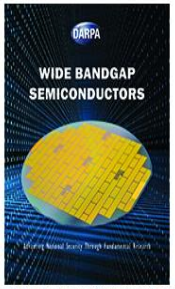
MAKE-IT



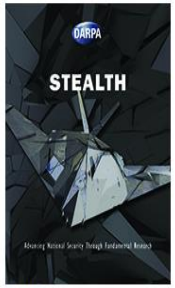
Microphysiological Systems



Quantum Sensing



Wide Bandgap Semiconductor



Stealth

Sigma

Very Large Scale Integrated Circuits

https://www.darpa.mil/program/explainable-artificial-intelligence

DARPA DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

EXPLORE BY TAG

ABOUT US / OUR RESEARCH / NEWS / EVENTS / WORK WITH US /

Defense Advanced Research Projects Agency > Our Research > Explainable Artificial Intelligence

# Explainable Artificial Intelligence (XAI)

Dr. Matt Turek

RESOURCES

DARPA-BAA-16-53

DARPA-BAA-16-53: Proposers Day Slides

XAI Program Portfolio

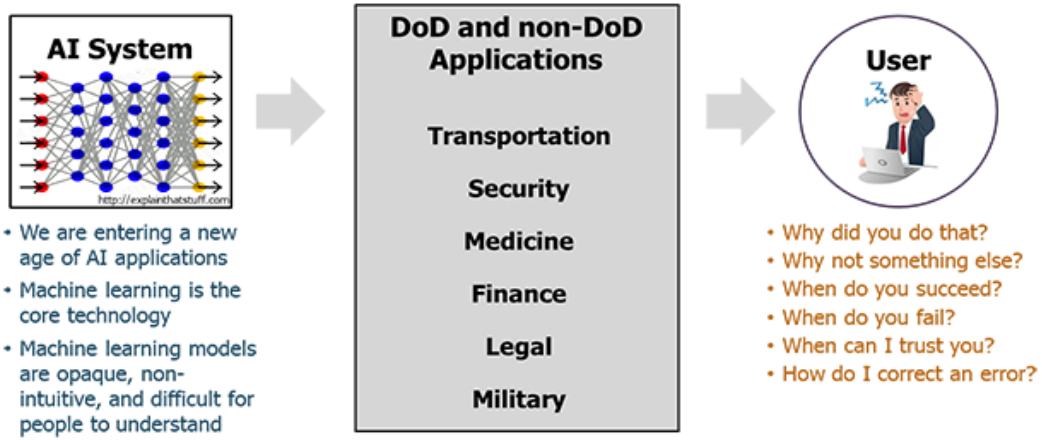



Figure 1. The Need for Explainable AI

## PART I: OVERVIEW INFORMATION

- **Federal Agency Name:** Defense Advanced Research Projects Agency (DARPA), Information Innovation Office (I2O)

- **Funding Opportunity Title:** Explainable Artificial Intelligence (XAI)

- **Announcement Type:** Initial Announcement

- **Funding Opportunity Number:** DARPA-BAA-16-53

- **Catalog of Federal Domestic Assistance Numbers (CFDA):** 12.910 Research and Technology Development

- **Dates**
  - Posting Date: August 10, 2016
  - Proposers Day: August 11, 2016
  - Abstract Due Date: September 1, 2016, 12:00 noon (ET)
  - Proposal Due Date: November 1, 2016, 12:00 noon (ET)
  - BAA Closing Date: November 1, 2016, 12:00 noon (ET)

- **Anticipated Individual Awards:** Multiple awards are anticipated.

- **Types of Instruments that May be Awarded:** Procurement Contracts, Cooperative Agreements, or Other Transactions (OTs). No grants will be awarded under this solicitation.

- **Agency Contacts**

  - **Technical POC**: David Gunning, Program Manager, DARPA/I2O

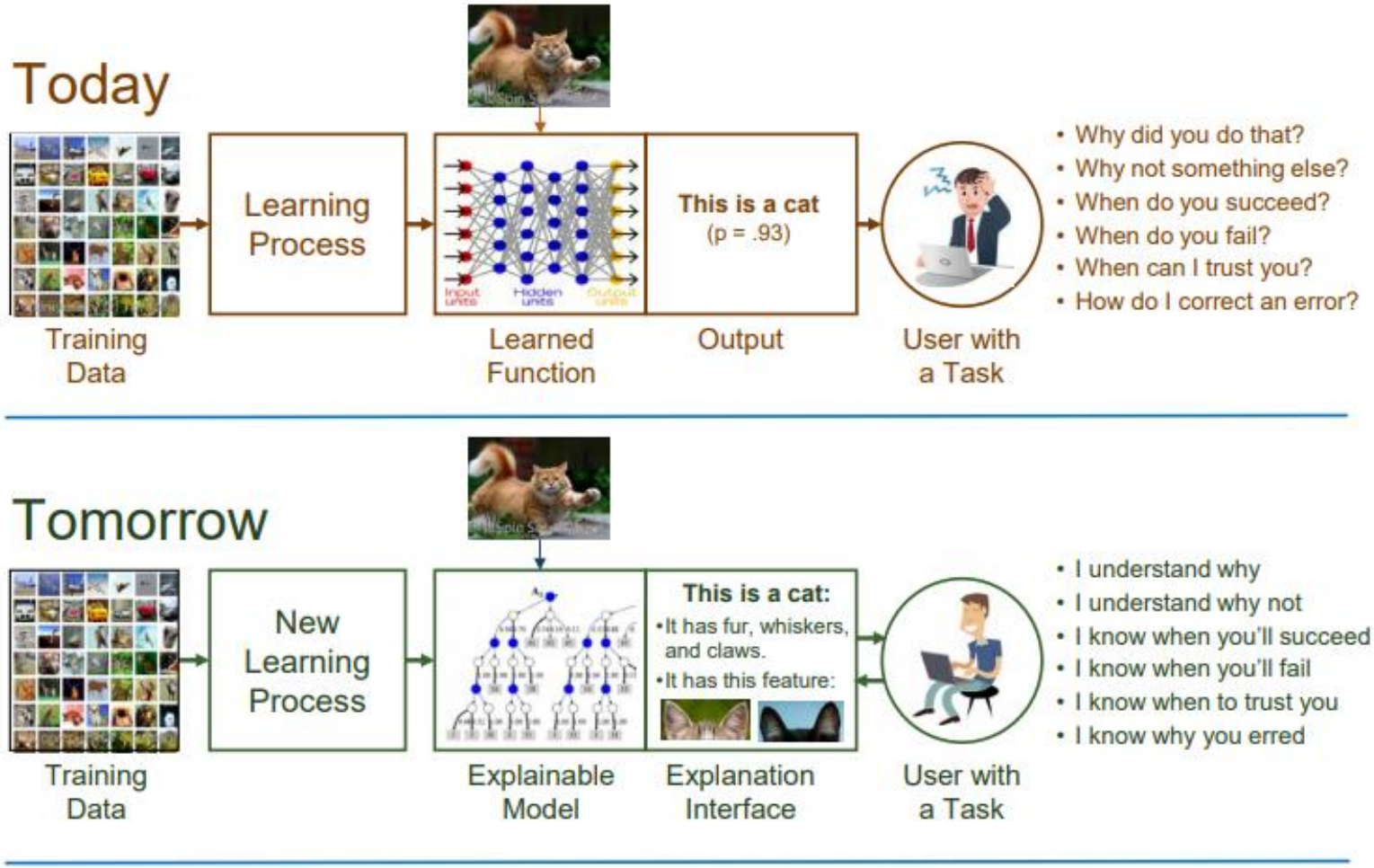
  - **BAA Email**: XAI@darpa.mil

  - **BAA Mailing Address**:
    DARPA/I2O
    ATTN: DARPA-BAA-16-53
    675 North Randolph Street
    Arlington, VA 22203-2114

  - **I2O Solicitation Website:** http://www.darpa.mil/work-with-us/opportunities

https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

What are they trying to do……



# Program Scope

https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

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

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.

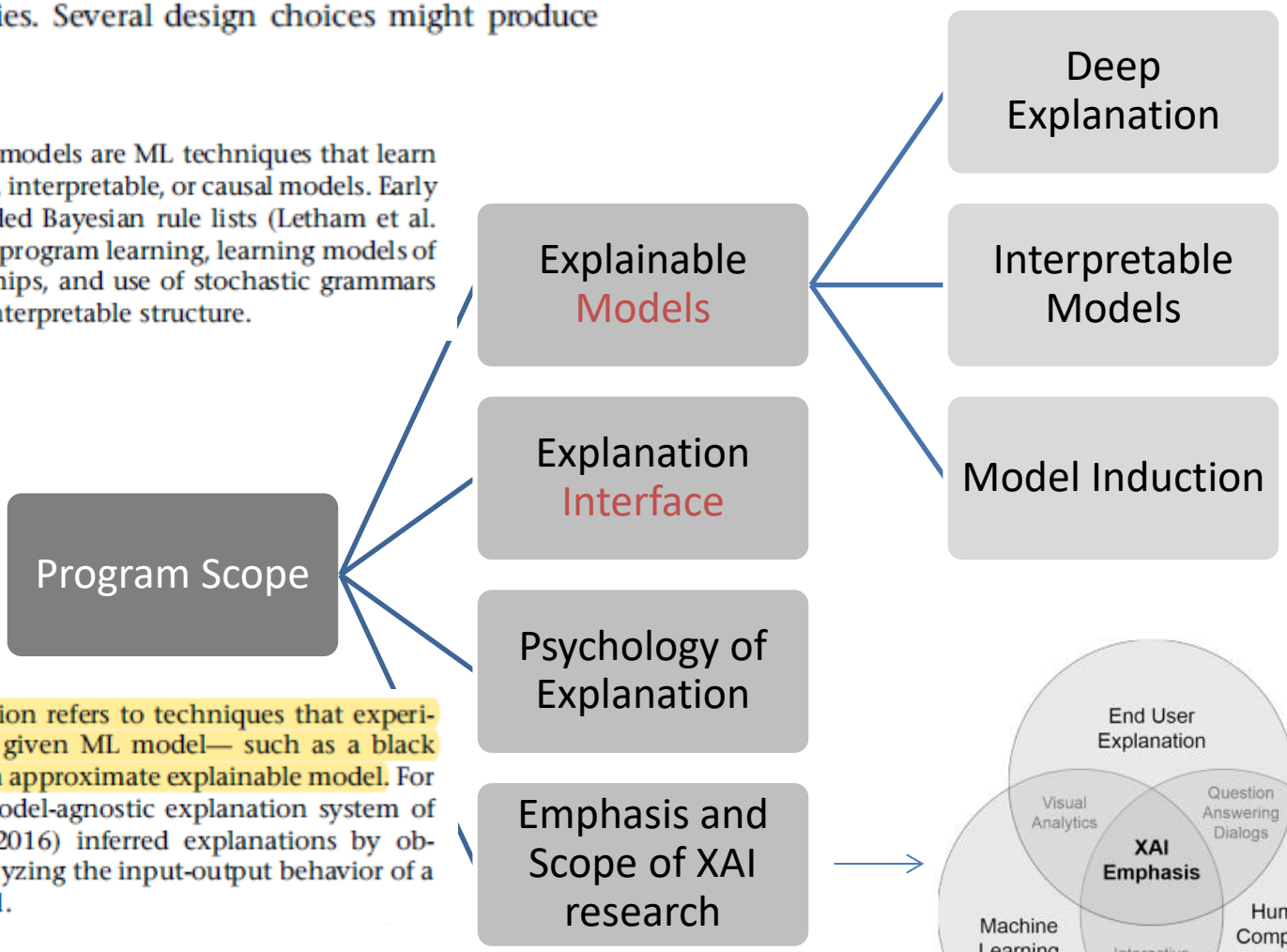



Figure 2: XAI Emphasis

https://ojs.aaai.org//index.php/aimagazine/article/view/2850

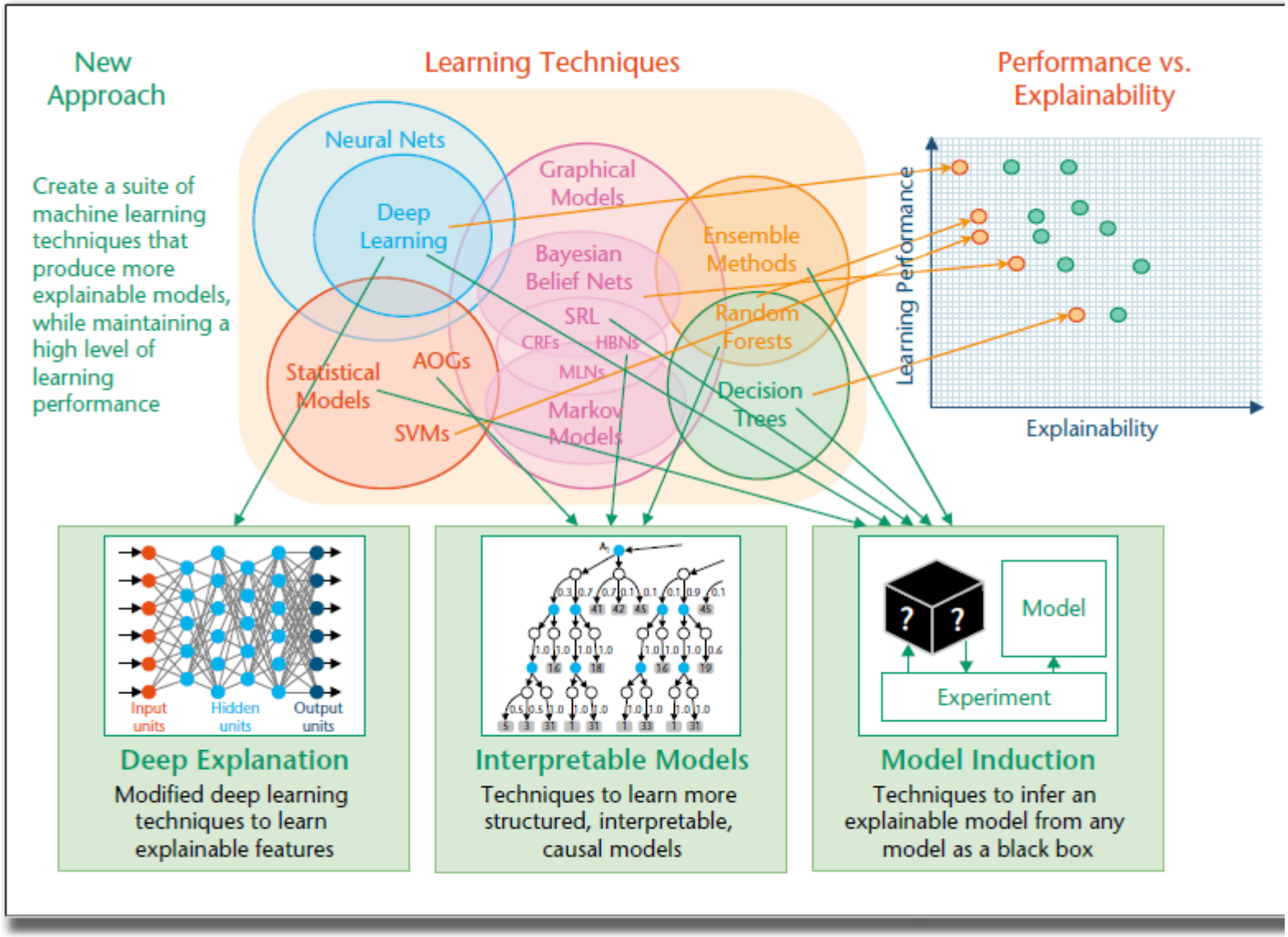
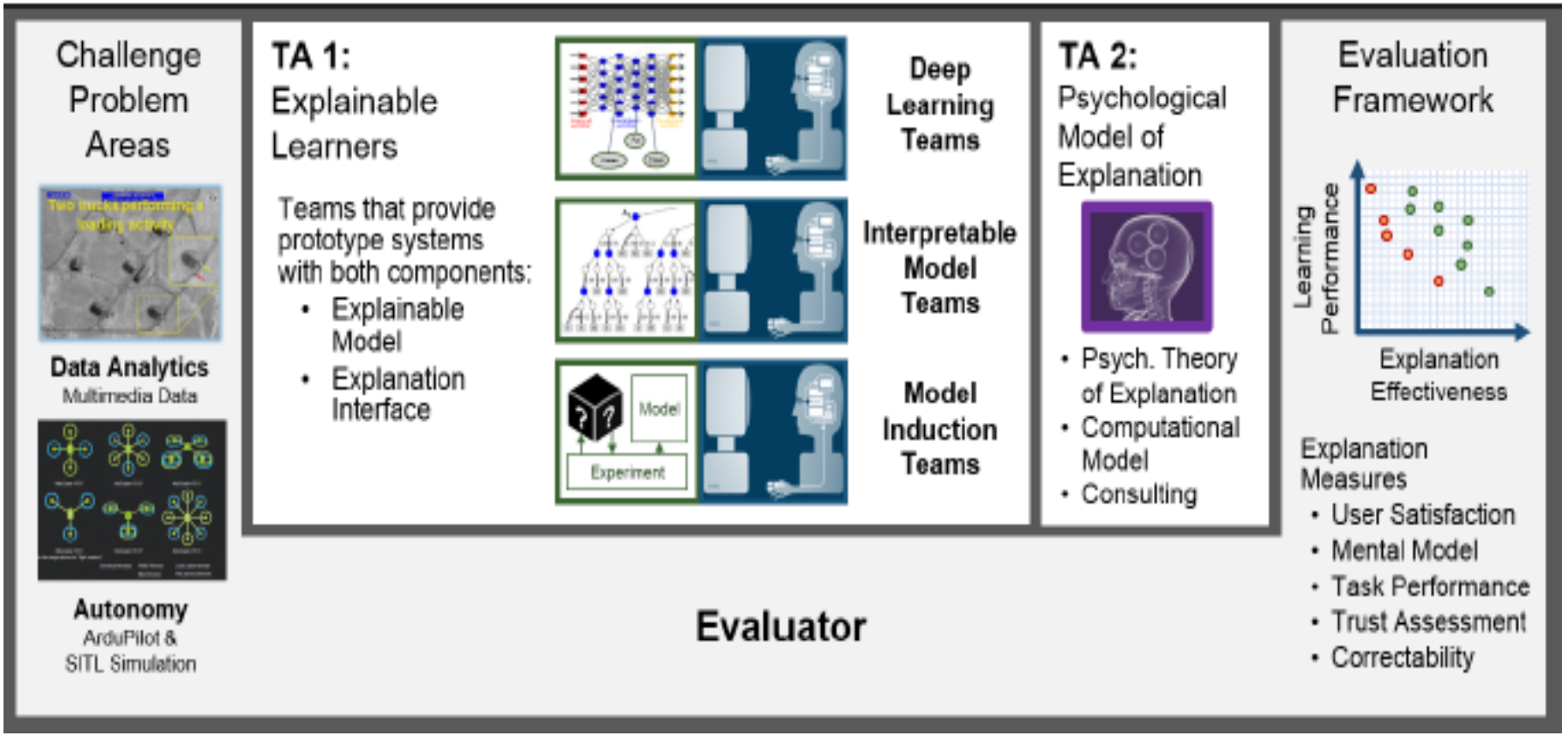



*Figure 2. Strategies for Developing Explainable Models.*

https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

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? --NDM

Program Structure

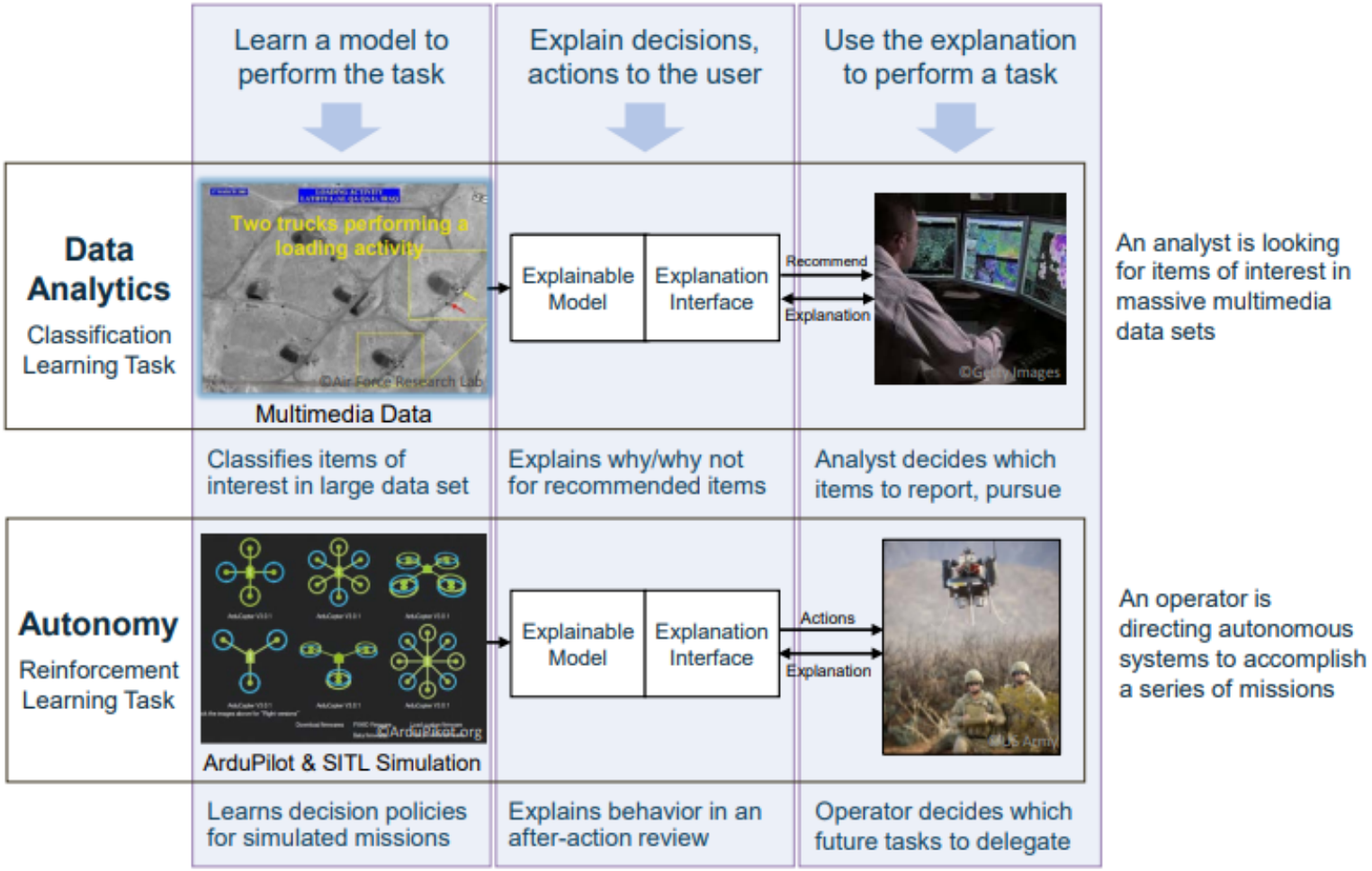



Data analytics: classification of events of interest in heterogeneous multimedia data

Autonomy: decision policies for autonomous systems

https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

# Explainable AI – Challenge Problem Areas



https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

# TA 1: Explainable Learners

## Challenge Problem Area

- Data analytics or Autonomy?
- Proposed test problem(s)
- Datasets, simulation, training data, functions, types of explaination to be provided, user decisions to be supported

## Explainable Model

- Proposed ML approach for learning more explainable models
- Scope of learning problems
- Learning process; required training data, human involvement etc.
- Measuring learning performance and explainaibility
- Strategies used

## Explaination Interface

- Approach for building the explanation interface; initial design ideas, principles or design strategies etc.
- Example of the explanation
- Range of ML models explained

## Development Progression

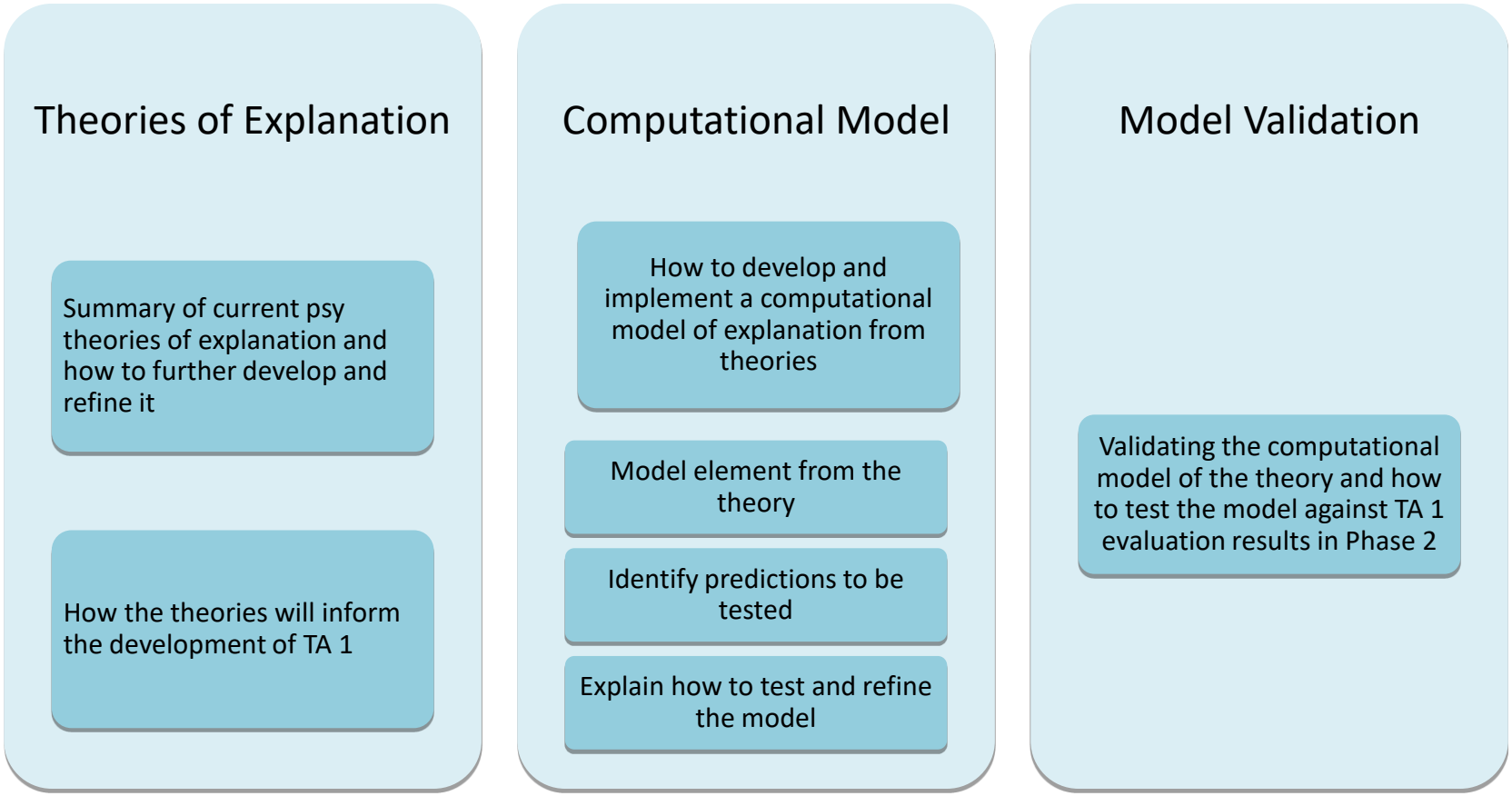
- Proposed development sequence
- Capabilities expected to develop during each phase

## Test and Evaluation Plan

- Proposed how the work will be tested and evaluated in the first phase
- Proposed learning performance measurement and explaination effectiveness

https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

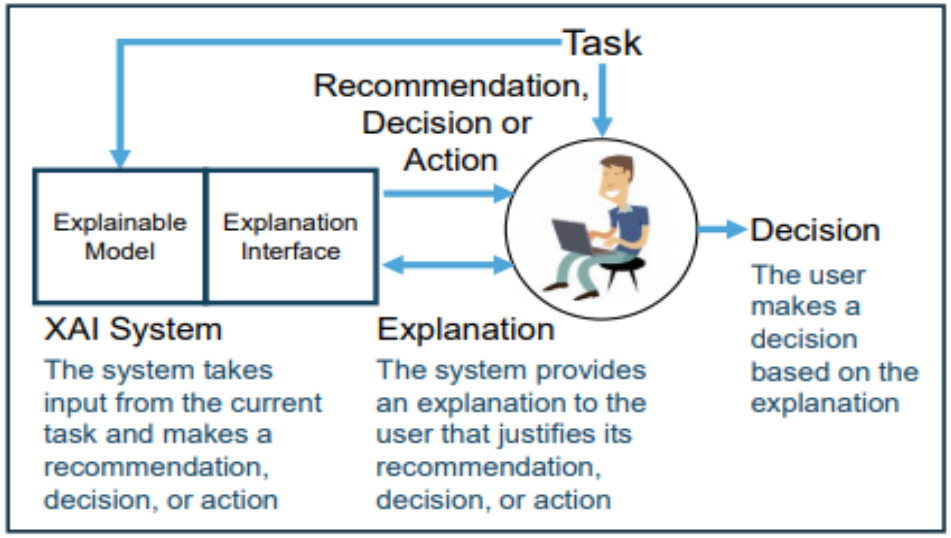
# TA 2: Psychological Model of Explanation



https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

# DARPA Explainable AI – Measuring Evaluation Effectiveness

## Explanation Framework




| Measure of Explanation Effectiveness |
|---|
| User Satisfaction |
| • Clarity of the explanation (user rating)<br>• Utility of the explanation (user rating) |
| Mental Model |
| • Understanding individual decisions<br>• Understanding the overall model<br>• Strength/weakness assessment<br>• 'What will it do' prediction<br>• 'How do I intervene' prediction |
| Task Performance |
| • Does the explanation improve the user's decision, task performance?<br>• Artificial decision tasks introduced to diagnose the user's understanding |
| Trust Assessment |
| • Appropriate future use and trust |
| Correctablity |
| • Identifying errors<br>• Correcting errors<br>• Continuous training |

https://ojs.aaai.org//index.php/aimagazine/article/view/2850

| 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.*

https://ojs.aaai.org//index.php/aimagazine/article/view/2850

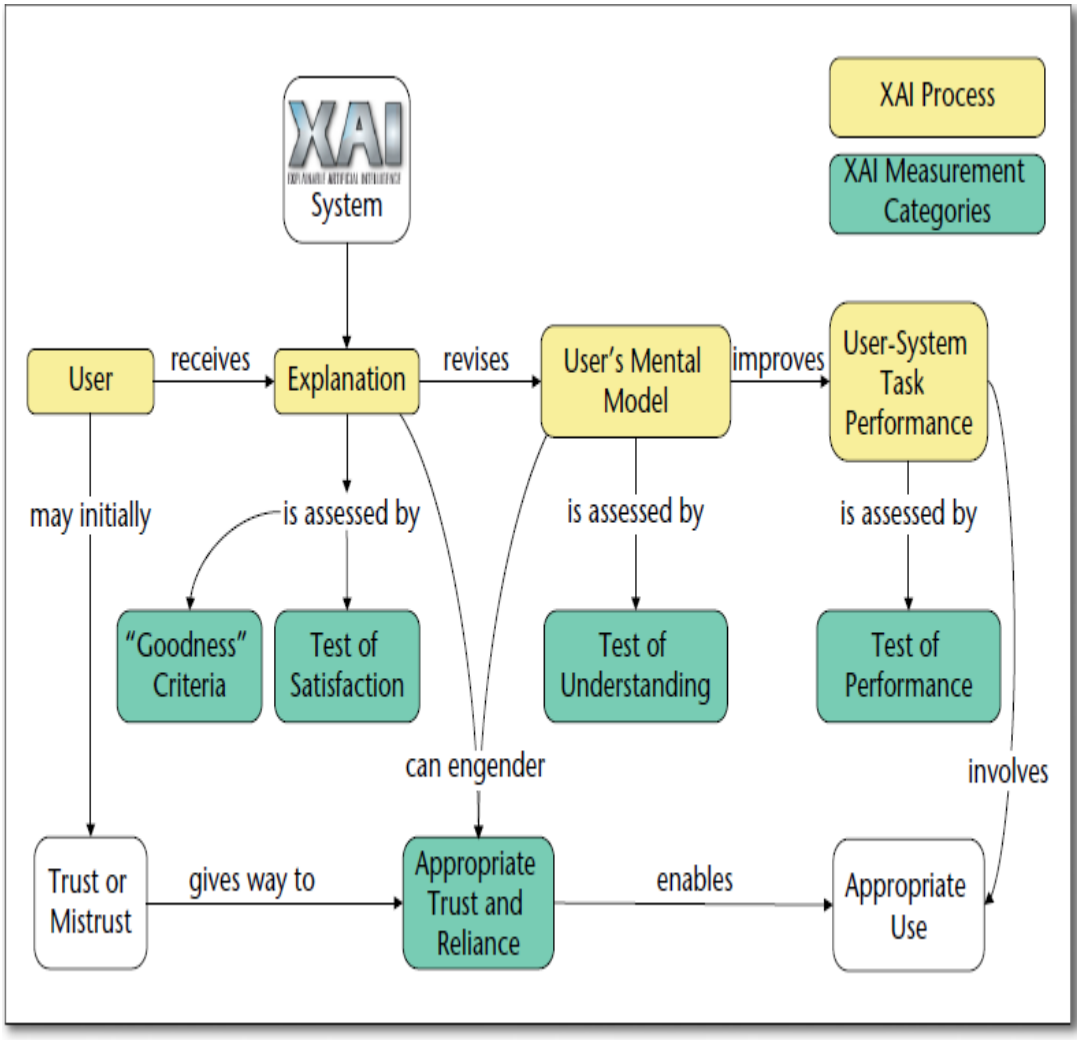



*Figure 8. Initial Model of the Explanation Process and Explanation Effectiveness Measurement Categories.*

## Experimental Conditions

***Without Explanation*** - The explainable learning system is used to perform a task without providing an explanation to the user

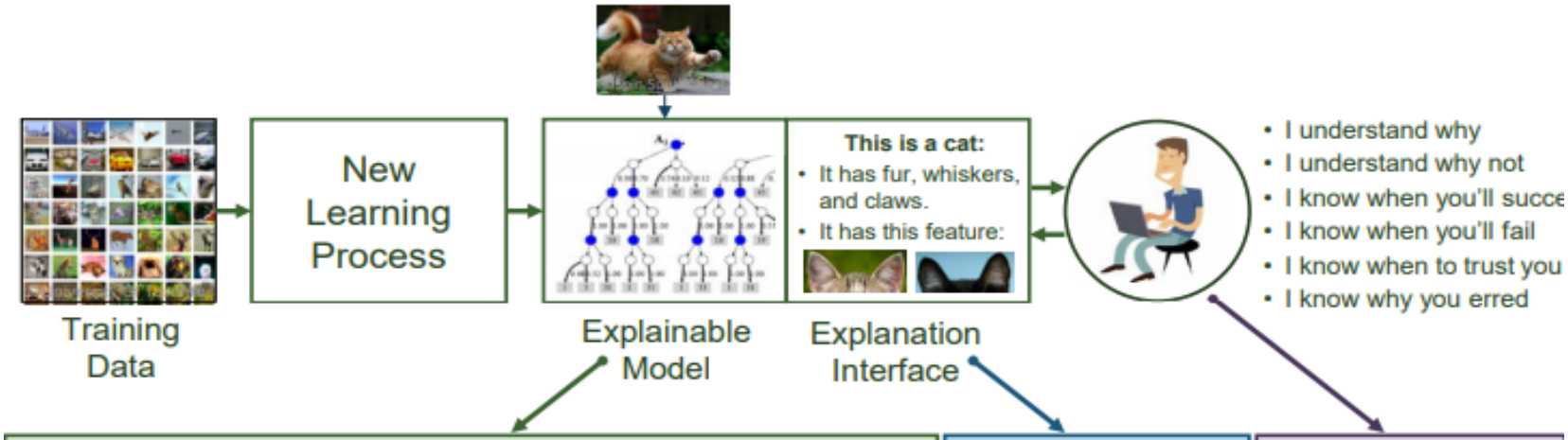
***With Explanation*** - The explainable learning system is used to perform a task and generates explanations for every recommendation or decision it makes, and every action it takes

***Partial Explanation*** - The explainable learning system is used to perform a task and generates only partial or ablated explanations (to assess various explanation features)

***Control*** - A baseline state-of-the-art non-explainable system is used to perform a task

https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

# DARPA Explainable AI – Why Do You Think It Will Be Successful?




| Deep Explanation | Interpretable Models | Model Induction | HCI | Psychology |
|---|---|---|---|---|
| **Learning Semantic Associations**<br>H. Sawhney (SRI Sarnoff) | **Stochastic And-Or-Graphs (AOG)**<br>Song-Chun Zhu (UCLA ) | **Local Interpretable Model-agnostic Explanations (LIME)**<br>C. Guestrin (UW) | **Prototype Explanation Interface**<br>T. Kulesza (OSU/MSR) | **Principles of Explanatory Machine Learning**<br>M. Burnett (OSU) |
| **Learning to Generate Explanations**<br>T. Darrell, P. Abeel (UCB) | **Bayesian Program Learning**<br>J. Tenenbaum (MIT) | **Bayesian Rule Lists**<br>C. Rudin (MIT) | **UX Design, Language Dialog, Visualization**<br>ENGINEERING PRACTICE | **Psychological Theories of Explanation**<br>T. Lombrozo (UCB) |

https://ojs.aaai.org/index.php/aimagazine/article/view/2850

## Program Schedule

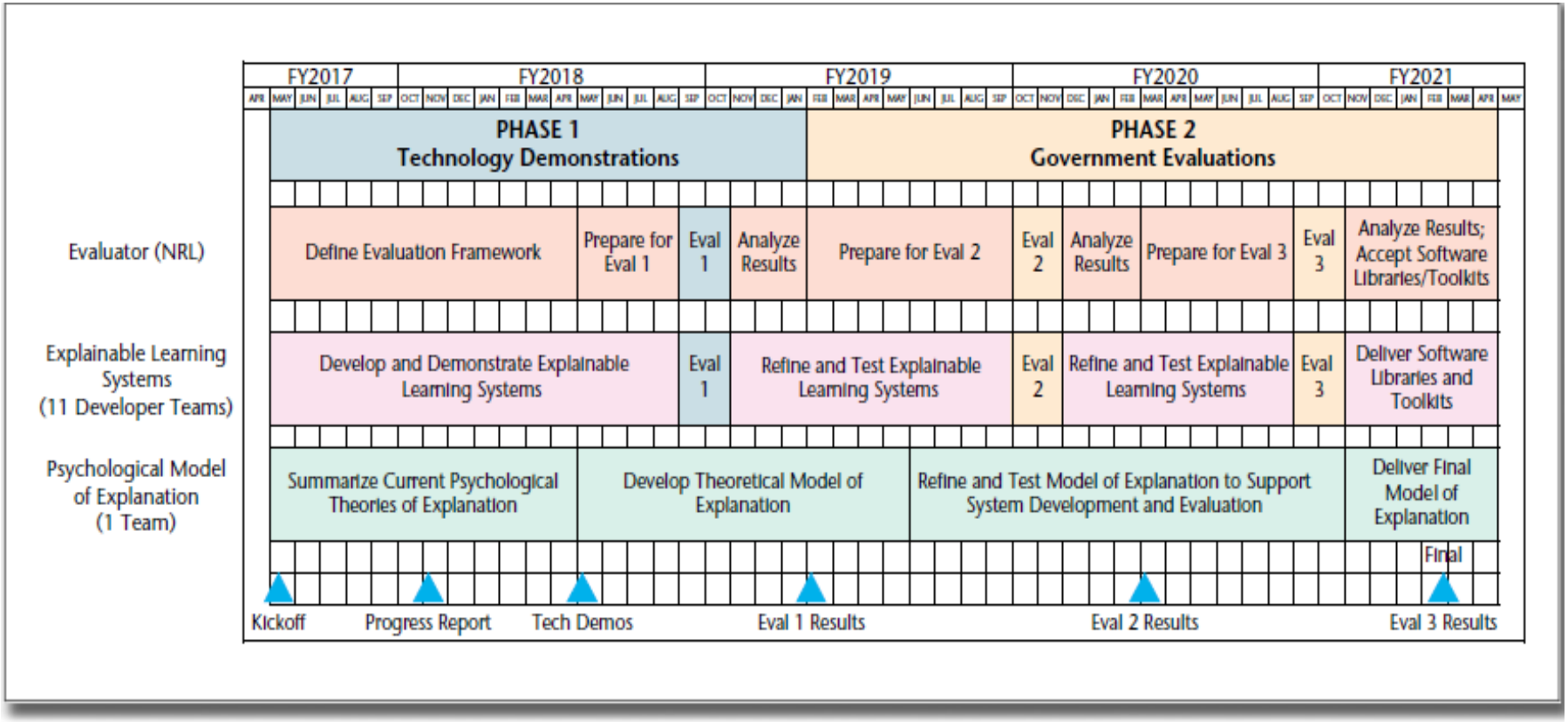



*Figure 6. XAI Program Schedule.*

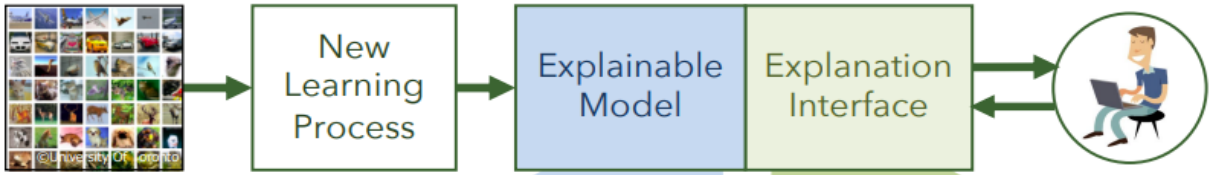
# Outline

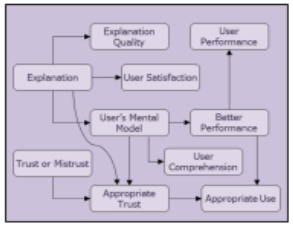
- Interpretability = Transparency ≠ Explainability?
- DARPA XAI Grant program
- Results: Grantees and reports
- Key takeaways and discussions

https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

**The Grantees!**

## DARPA XAI Developers and Technical Approaches




IHMC Psychological Models of Explanation

| CP | Performer | Explainable Model | Explanation Interface |
|---|---|---|---|
| Both | UC Berkeley | Deep Learning | Reflexive and Rational |
| | Charles River | Causal Modeling | Narrative Generation |
| | UCLA | Pattern Theory+ | 3-level Explanation |
| Autonomy | Oregon State | Adaptive Programs | Acceptance Testing |
| | PARC | Cognitive Modeling | Interactive Training |
| | CMU | Explainable RL (XRL) | XRL Interaction |
| Analytics | SRI International | Deep Learning | Show and Tell Explanation |
| | Raytheon BBN | Deep Learning | Argumentation and Pedagogy |
| | UT Dallas | Probabilistic Logic | Decision Diagrams |
| | Texas A&M | Mimic Learning | Interactive Visualization |
| | Rutgers | Model Induction | Bayesian Teaching |

https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

**UC Berkeley**


# Deeply Explainable Artificial Intelligence


**UC Berkeley, Boston U., U. Amsterdam, Kitware**

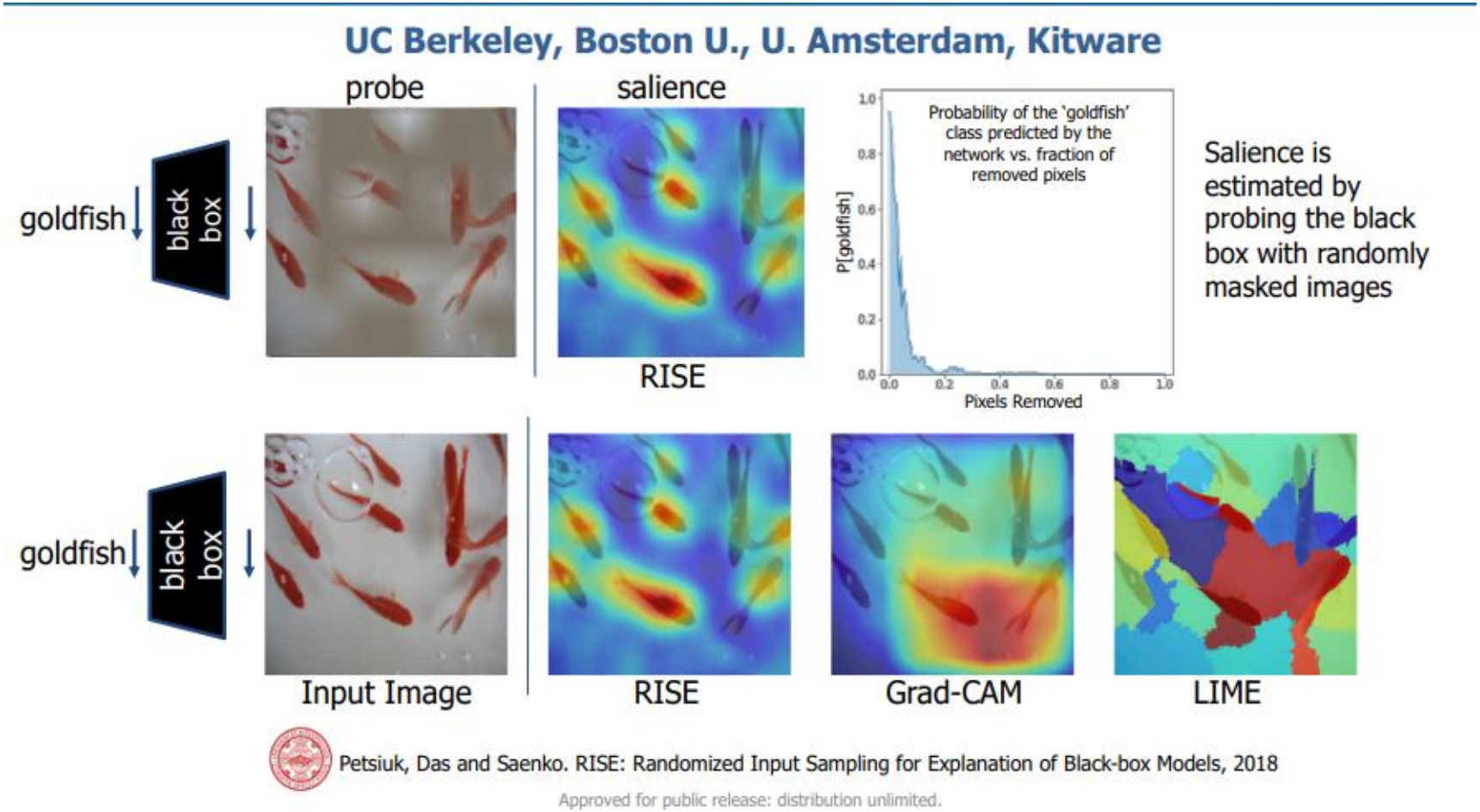
| Explainable Model | Explanation Interface | Challenge Problem |
|---|---|---|
| **Deep Learning**<br>• Post-hoc explanations by training additional DL models<br>• Explicit introspective explanations (Neural Module Networks)<br>• Reinforcement Learning<br>  • Informative rollouts<br>  • Explicit modular agent | **Reflexive and Rational**<br>• Reflexive explanations (arise from the model)<br>• Rational explanations (come from reasoning about user's beliefs)<br>• Evaluation criteria<br>  • Human interpretability<br>  • Predictive behavior<br>  • Appropriate trust | **Autonomy**<br>• Vehicle control (BDD-X, CARLA)<br>• Strategy games (StarCraft II)<br><br>**Data Analytics**<br>• Visual QA and filtering tasks (VQA-X, ACT-X, xView, DiDeMo, etc.) |

- **PI**: Trevor Darrell (UC Berkeley)

- Pieter Abbeel (UC Berkeley)
- Tom Griffiths (UC Berkeley)
- Kate Saenko (Boston U.)
- Zeynep Akata (U. Amsterdam)
- Dan Klein (UC Berkeley)
- John Canny (UC Berkeley)
- Anca Dragan (UC Berkeley)
- Anthony Hoogs (Kitware)

https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

**UC Berkeley**



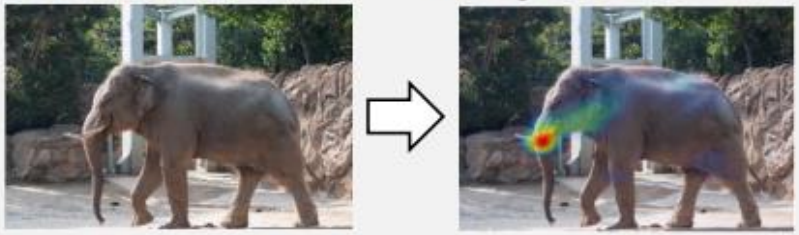
https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

## UC Berkeley

Given the multi-modal explanation generated by the model, do you think the system will answer correctly?

Question: *Does this elephant have tusks?*

*"because there are no bones sticking out from its mouth"*



Yes No

Incorrect! The system answered *"no"* when the ground-truth answer is *"yes"*

Question: *Is this a professional sporting event?*

*"because the players are wearing official jerseys"*

Yes No

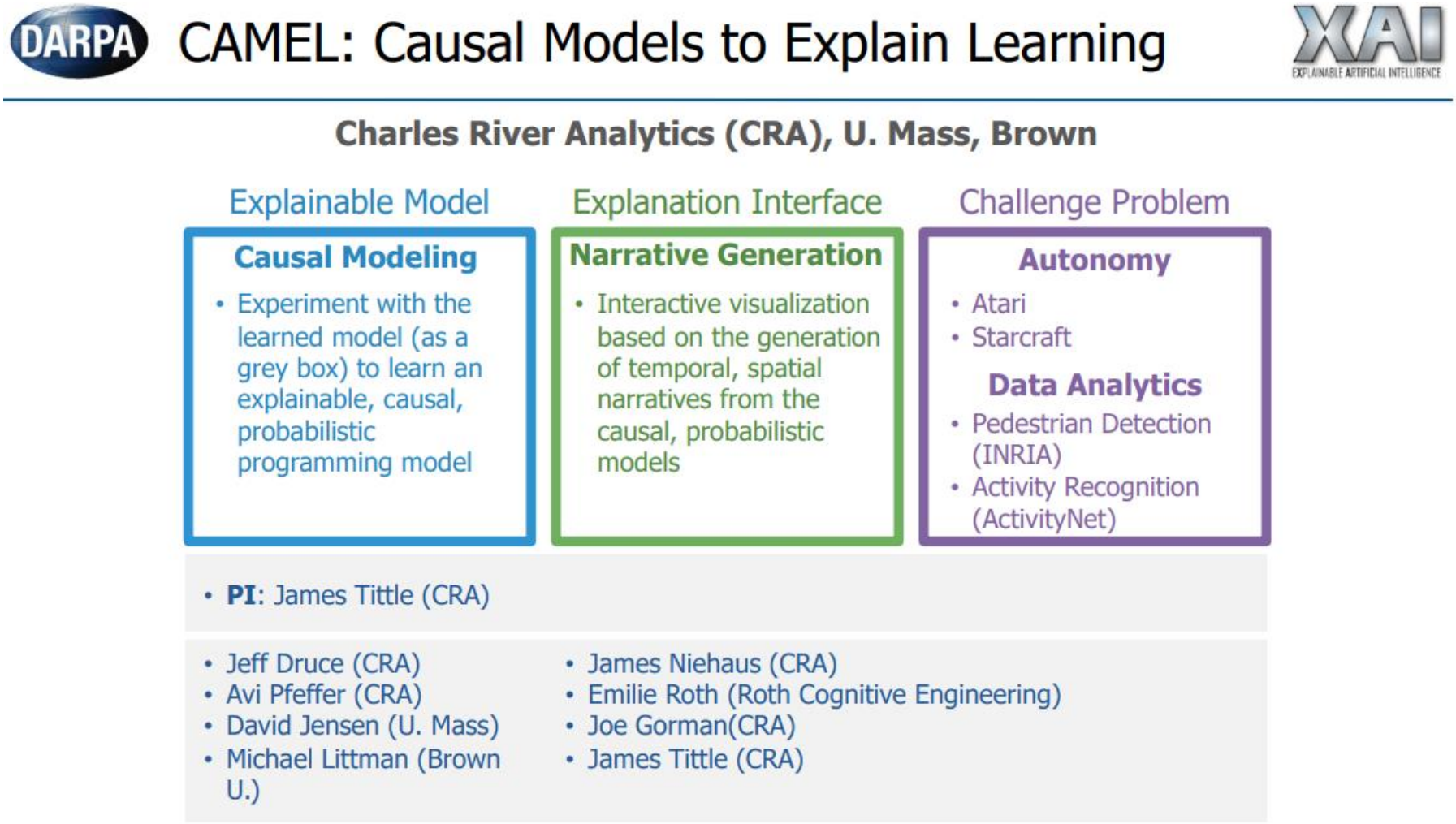
Correct! The system answered *"yes"* when the ground-truth answer is *"yes"*

| Explanation Effectiveness | Attention for Explanation Used? | Accuracy of Users Judgement |
|---|---|---|
| **Without explanation (existing SOTA)** | No | 57.5% |
| **UCB Model on descriptions** | Yes | 66.5% |
| **UCB Model without attention** | No | 61.5% |
| **UCB Model** | Yes | **70.0%** |

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence, 2018

https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

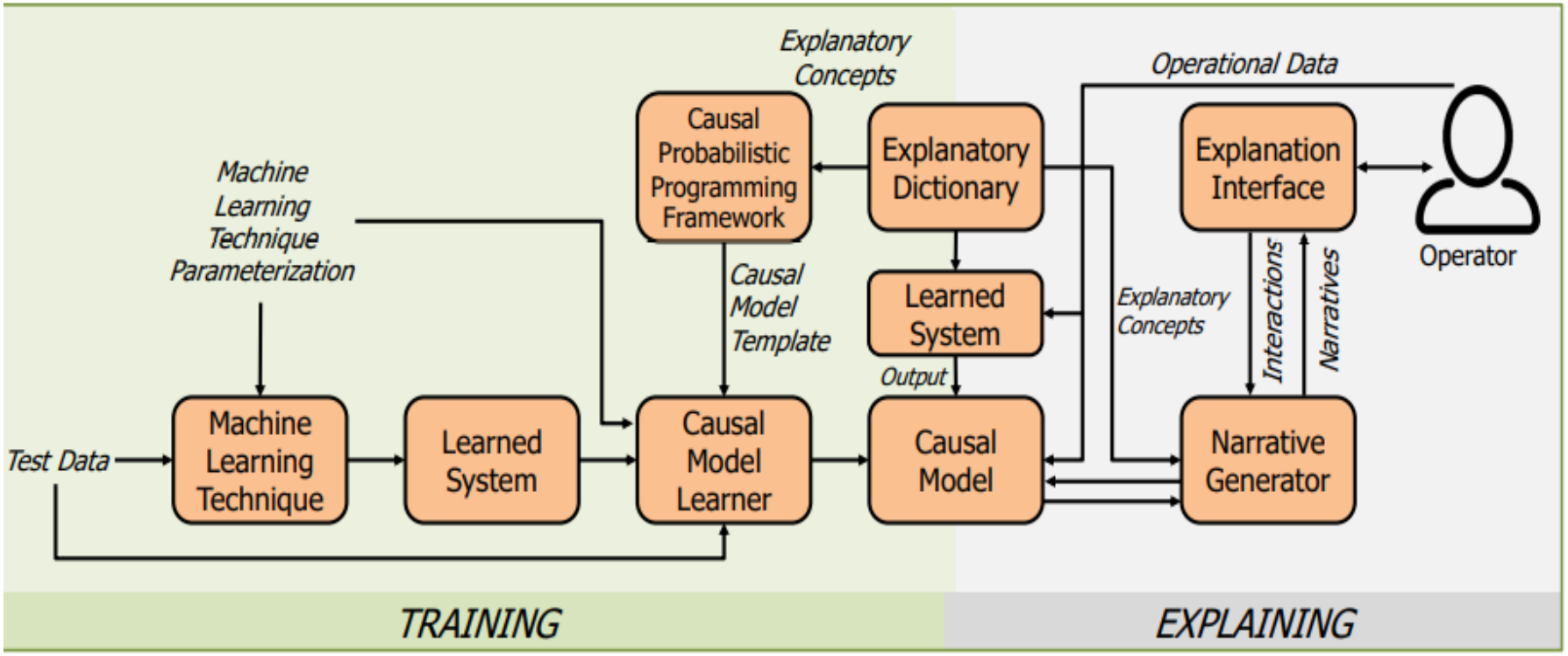
**Charles River Analytics**



https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

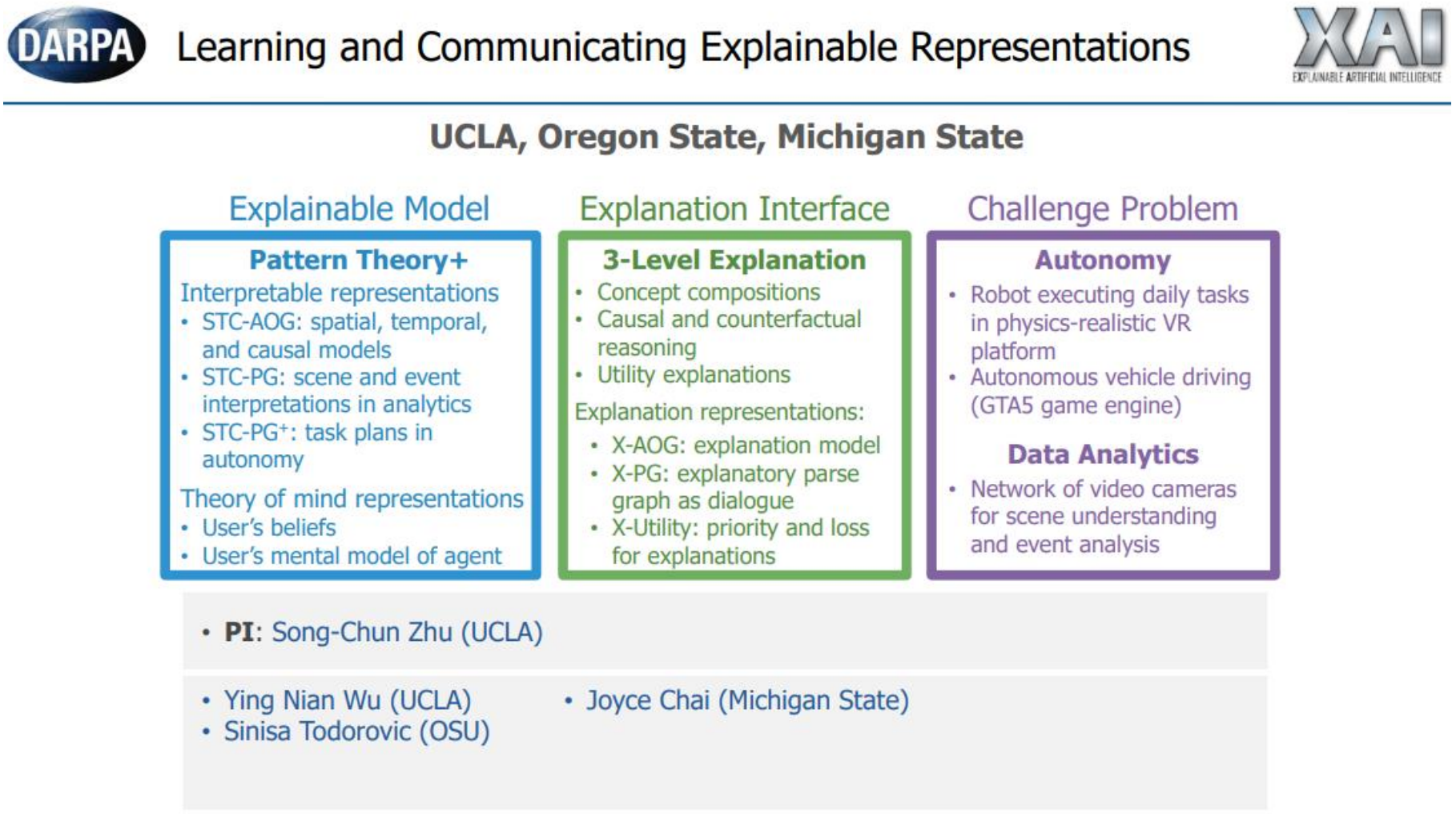
**Charles River Analytics**

Generate causal explanations of ML operation and present them to the user as intuitive narratives in an interactive, easy-to-use interface grounded in cognitive engineering theories



https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

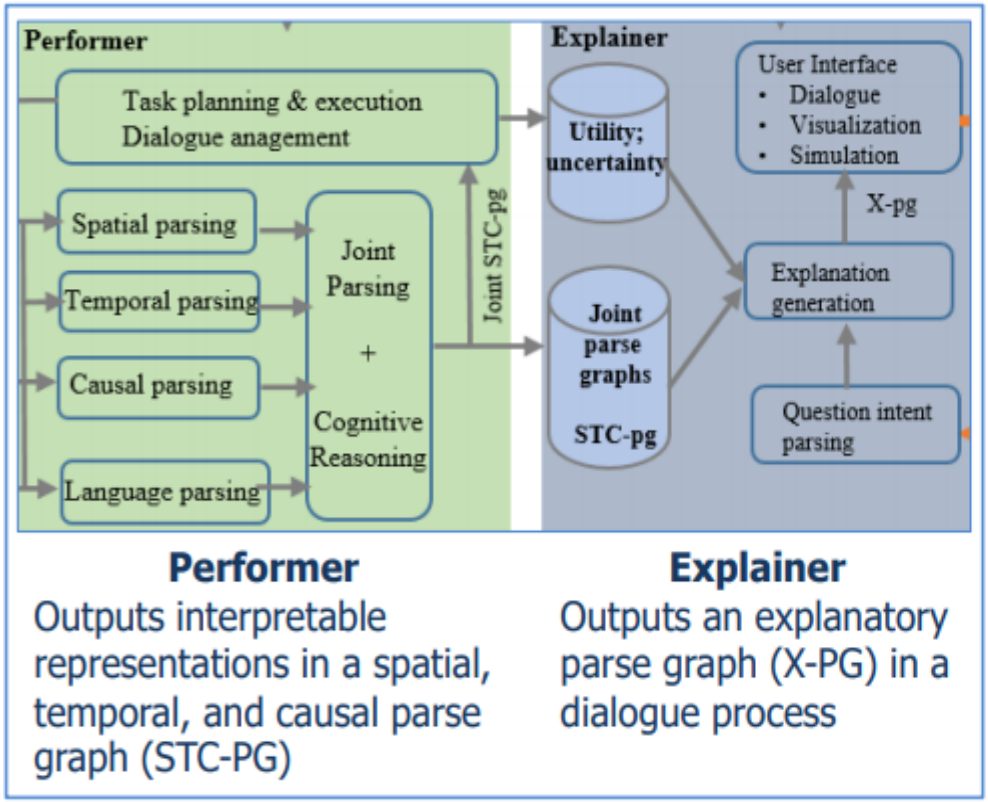
**UCLA**



https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

**UCLA**

## System Architecture



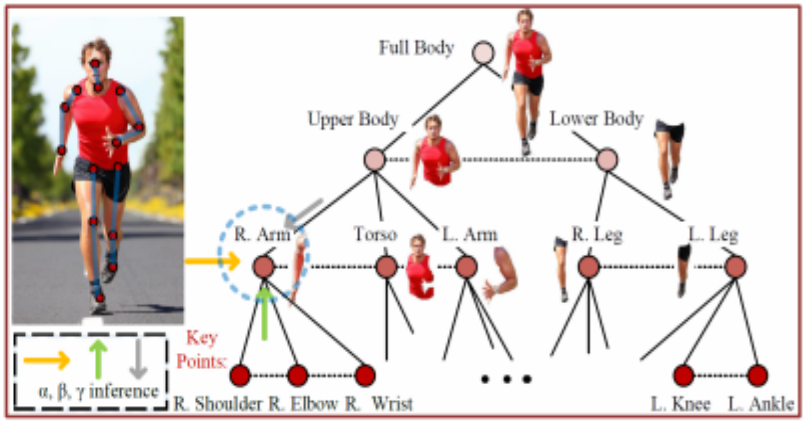

**Performer**

Outputs interpretable representations in a spatial, temporal, and causal parse graph (STC-PG)

**Explainer**

Outputs an explanatory parse graph (X-PG) in a dialogue process

## STC Parse Graph



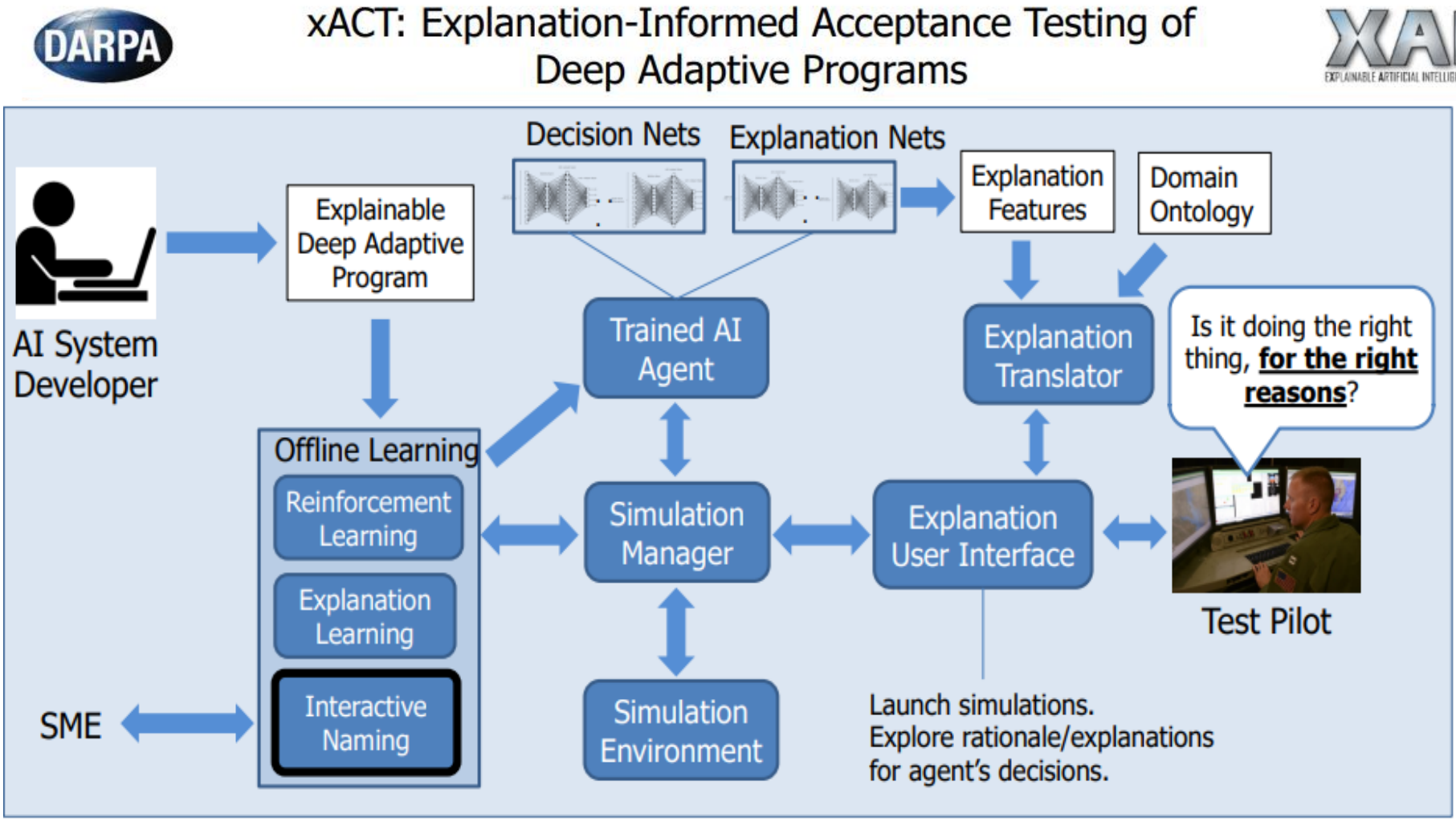

An attributed parse graph for a running person. Each node has 3 computing channels:

- $\alpha$: grounding the node on DNN features;
- $\beta$: bottom-up;
- $\gamma$: top-down.

An explanation is represented as parse graph X-pg

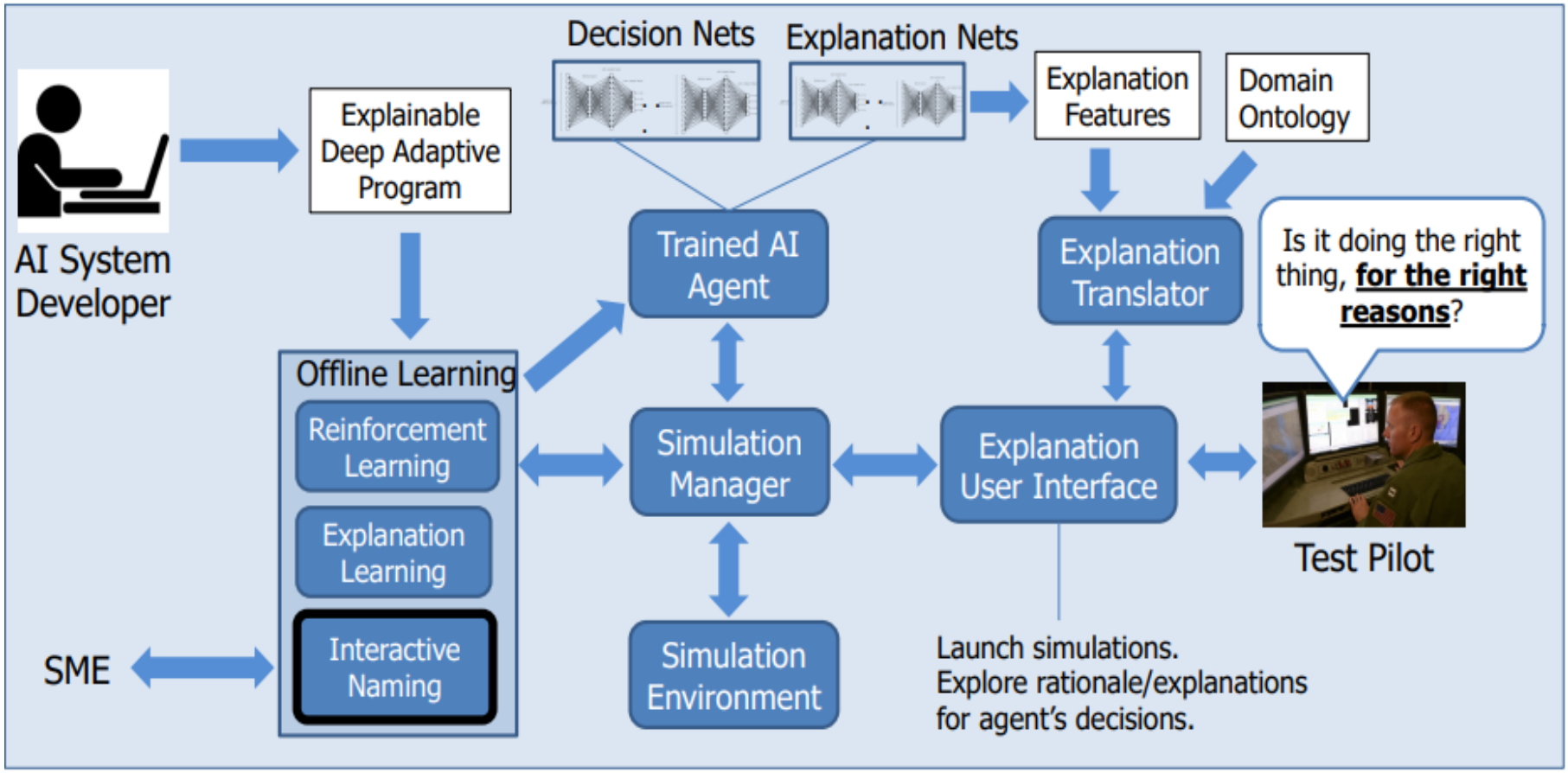
https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

**Oregon State**



https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

**Oregon State**



https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

**PARC**


# COGLE: Common Ground Learning and Explanation


**PARC, CMU, U. Edinburgh, U. Michigan, USMA, IHMC**

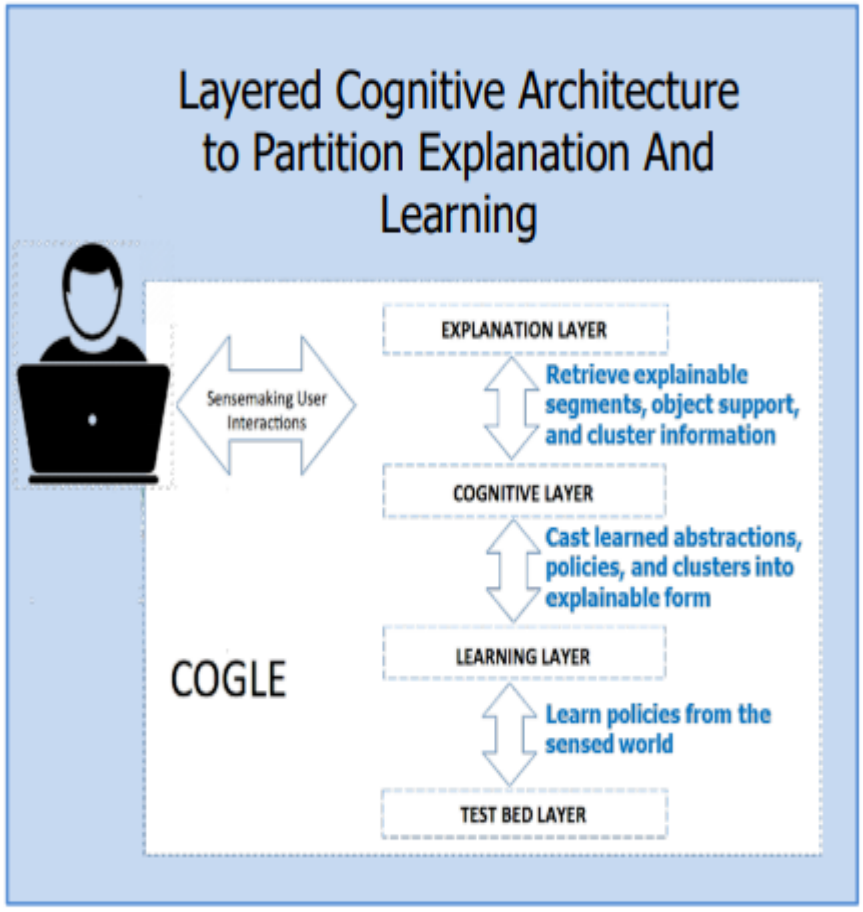
| Explainable Model | Explanation Interface | Challenge Problem |
|---|---|---|
| **Cognitive Model**<br>3-layer architecture<br>• Learning Layer (DNNs)<br>• Cognitive Layer (ACT-R Cognitive Model)<br>• Explanation Layer (HCI) | **Interactive Training**<br>• Interactive visualization of states, actions, policies, and values<br>• Module for test pilots to refine and train the system | **Autonomy**<br>• MAVSim wrapper over ArduPilot simulation environment<br>• Value of Explanation framework for measuring explanation effectiveness |

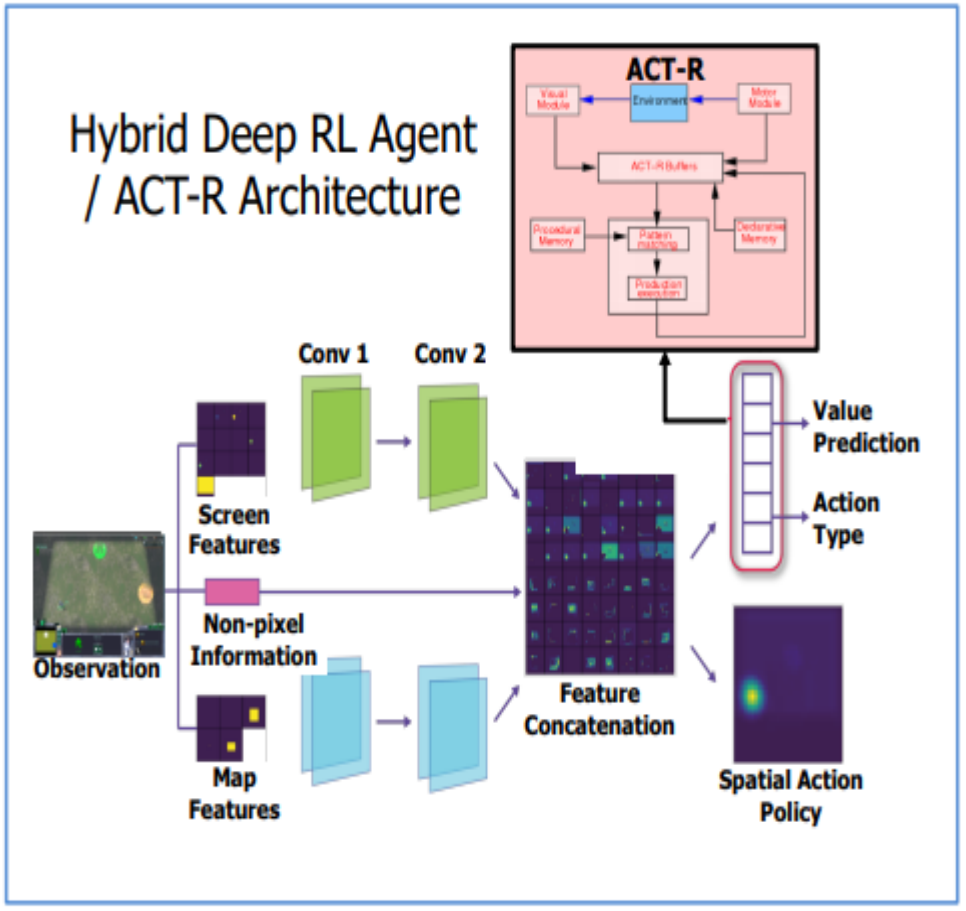
- **PI**: Mark Stefik (PARC)

- Honglak Lee (U. Michigan)
- Subramanian Ramamoorthy (U. Edinburgh)
- Christian Lebiere (CMU)
- John Anderson (CMU)
- Robert Thomson (USMA)
- Michael Youngblood (PARC)

https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

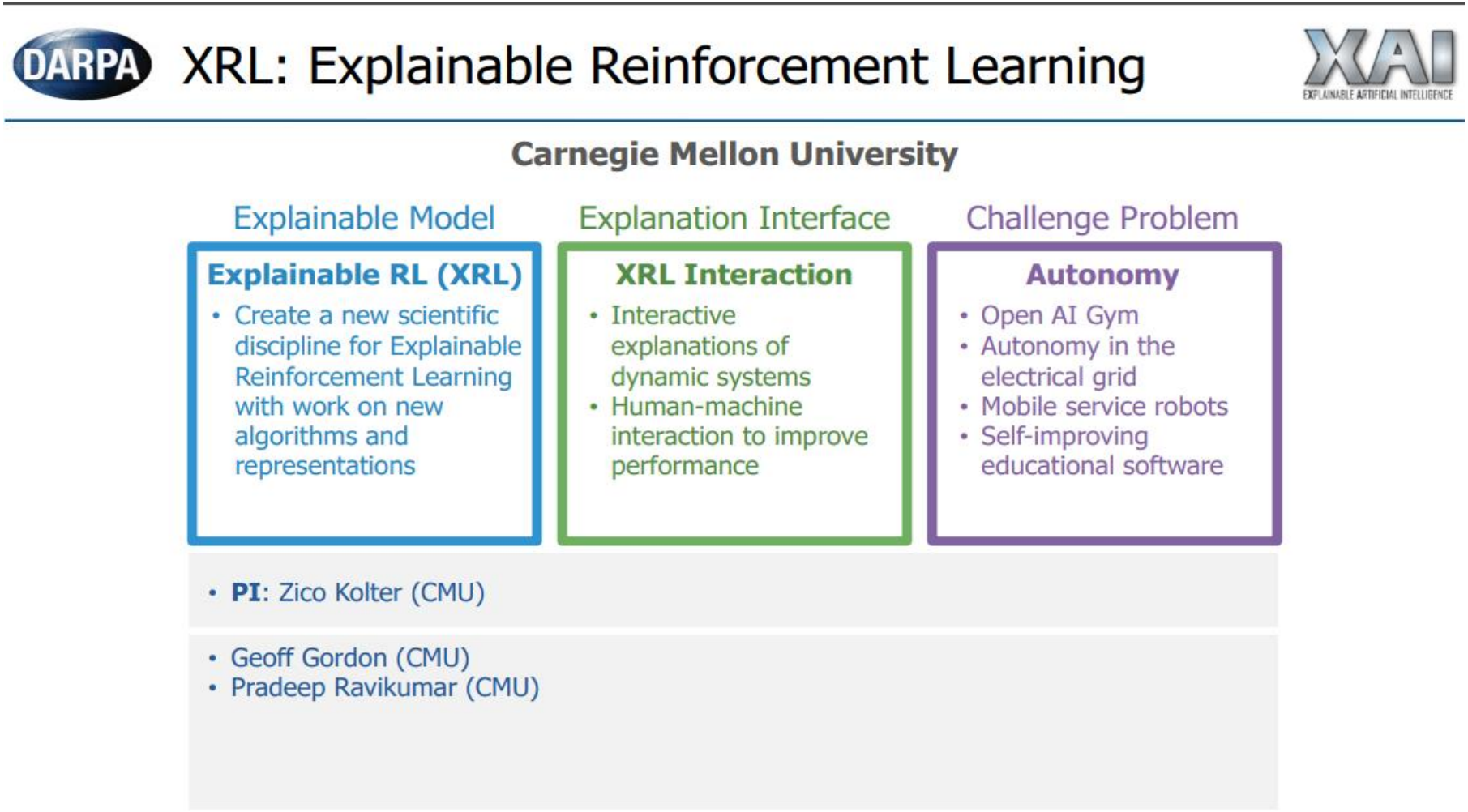
**PARC**



https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

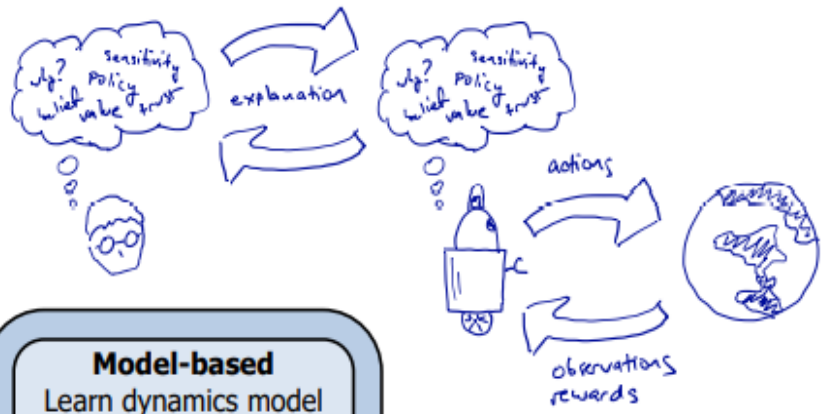
**CMU**



https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

## CMU

Create a new discipline of explainable RL to enable dynamic human-machine interaction and adaptation for maximum team performance

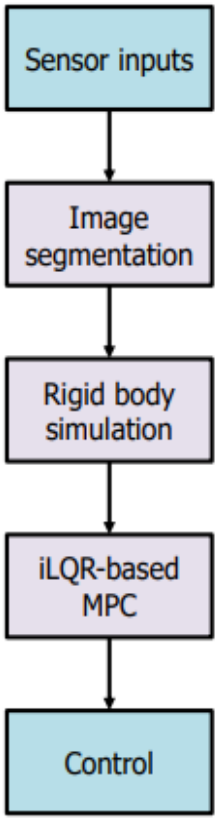



**Model-based**
Learn dynamics model of environment, plan actions in model, and execute in real system

→ Improve model learning/ creation for RL agents to capture benefits of model-based approach

**Model-free**
Directly learn value and/or policy for the environment

→ For any type of RL approach, provide an explanation of why an agent acted in a certain way




**Differentiable Physics -** Applies implicit differentiation to solutions of LCP to analytically derive a backpropagation update of next state with respect to previous state, control, and model parameters

https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

**SRI International**


## DARE: Deep Attention-based Representations for Explanation


### SRI International, U. Toronto, UCSD, U. Guelph

| Explainable Model | Explanation Interface | Challenge Problem |
|---|---|---|
| **Deep Learning**<br>Multiple deep learning techniques<br>• Attention-based mechanisms<br>• Compositional NMNs<br>• GANs | **Show & Tell Explanation**<br>• DNN visualization<br>• Query evidence that explains DNN decisions<br>• Generate natural language justifications | **Data Analytics**<br>• VQA<br>  • Visual Gnome<br>  • Flickr30<br>• MovieQA |

- **PIs**: Giedrius Burachas (SRI), Mohamed Amer (SRI)

- Xiao Lin (SRI)
- Ryan Villamil (SRI)
- Dejan Jovanovic (SRI)
- Avi Ziskind (SRI)
- Michael Wessel (SRI)

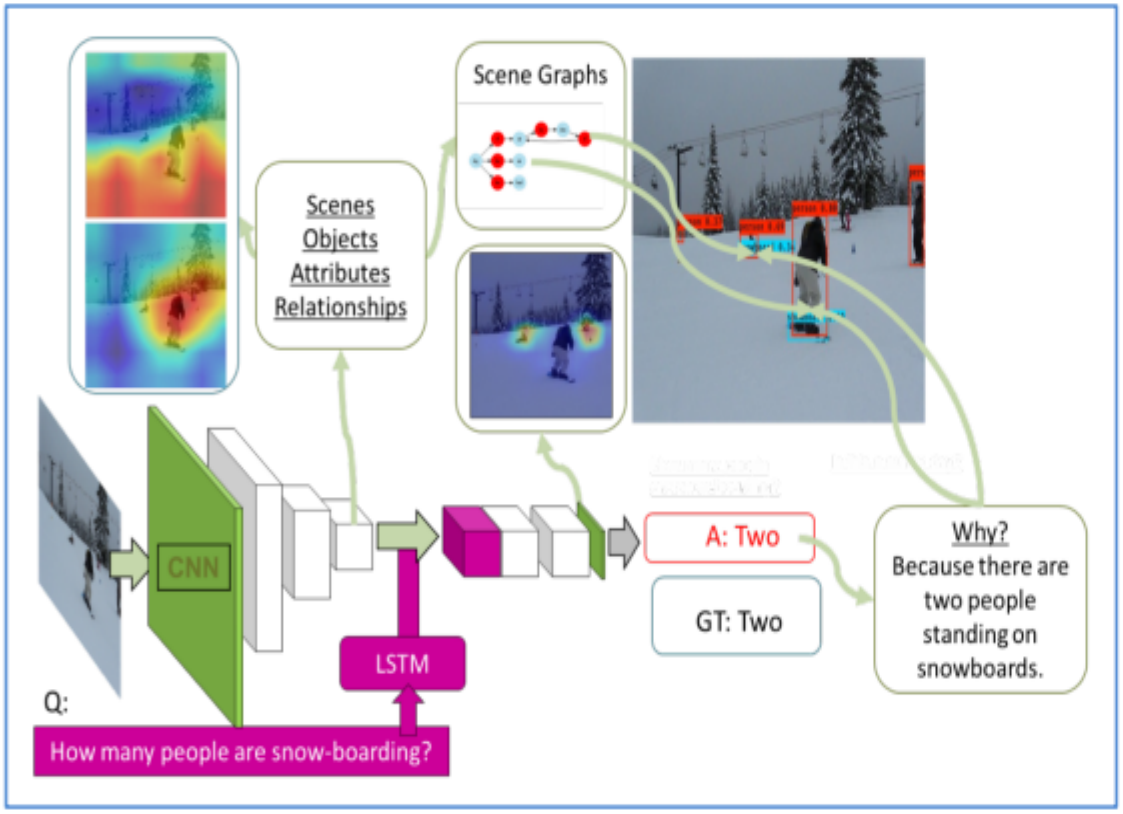
Richard R. Zemel (U. Toronto)
Sanja Fidler (U. Toronto)
David Duvenaud (U. Toronto)
Graham Taylor (U. Guelph)

- Jürgen Schulze (UCSD)

https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

**SRI International**

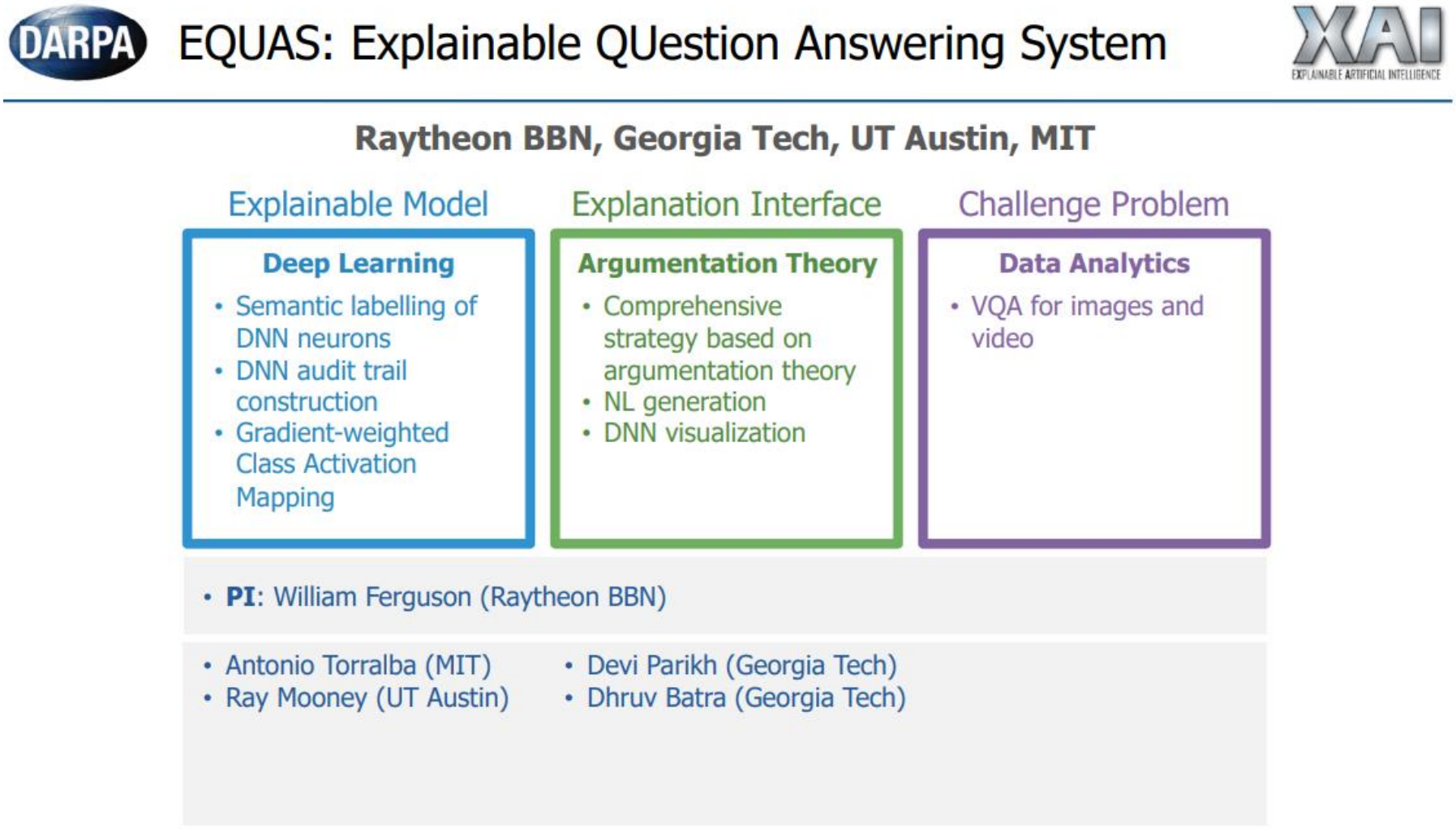
## Interpretable, Scene Graph-based VQA System with Active Attention




- Generate "show-and-tell" explanations with justifications of decisions accompanied by visualizations of input data used to generate inferences
- Scene and Situation Graphs, inferred from images and videos, support rich multimodal data analytics and explanations
- Scene Graphs guide attentional scanning for interpretable analytics

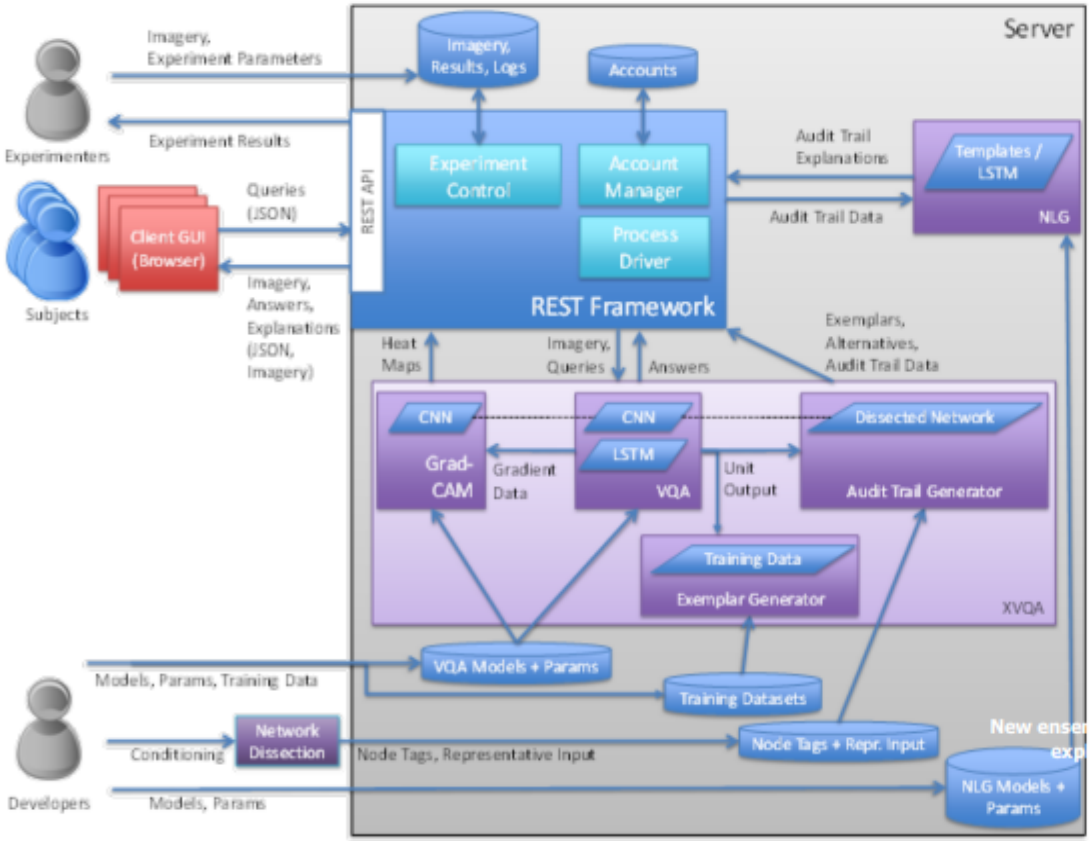
https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

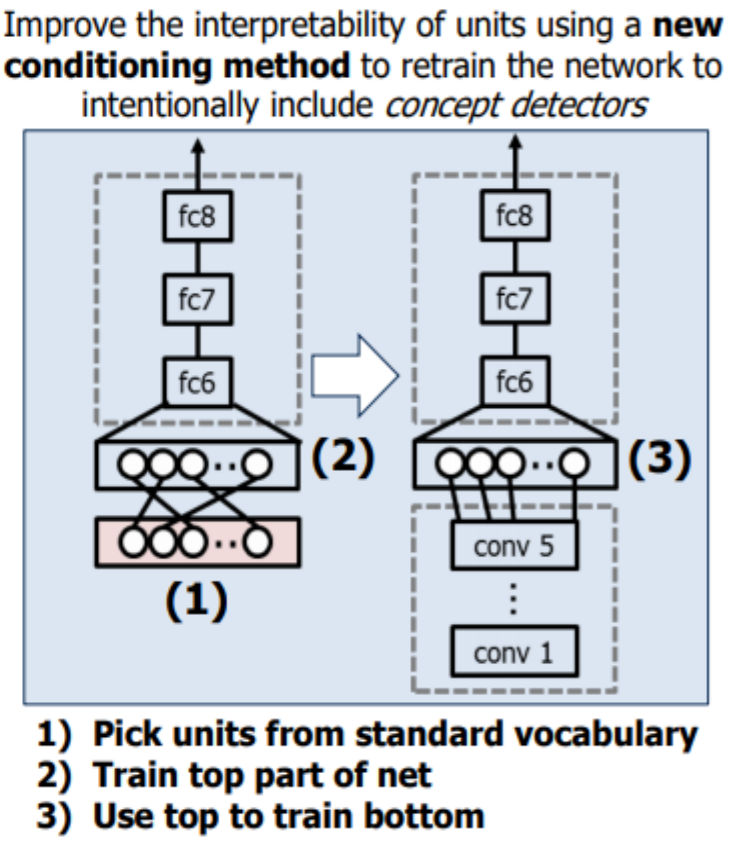
**Raytheon BBN**



https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

## Raytheon BBN



https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

**UT Dallas**


# Tractable Probabilistic Logic Models


**UT Dallas, UCLA, Texas A&M, Indian Institute of Technology**

| Explainable Model | Explanation Interface | Challenge Problem |
|---|---|---|
| **Probabilistic Logic**<br>• Tractable Probabilistic Logic Models (TPLMs) – an important class of (non-deep learning) interpretable models | **Decision Diagrams**<br>• Enables users to explore and correct the underlying model as well as add background knowledge | **Data Analytics**<br>• Infer activities in multimodal data (video and text)<br>• Wetlab (biology) and TACoS (cooking) datasets |

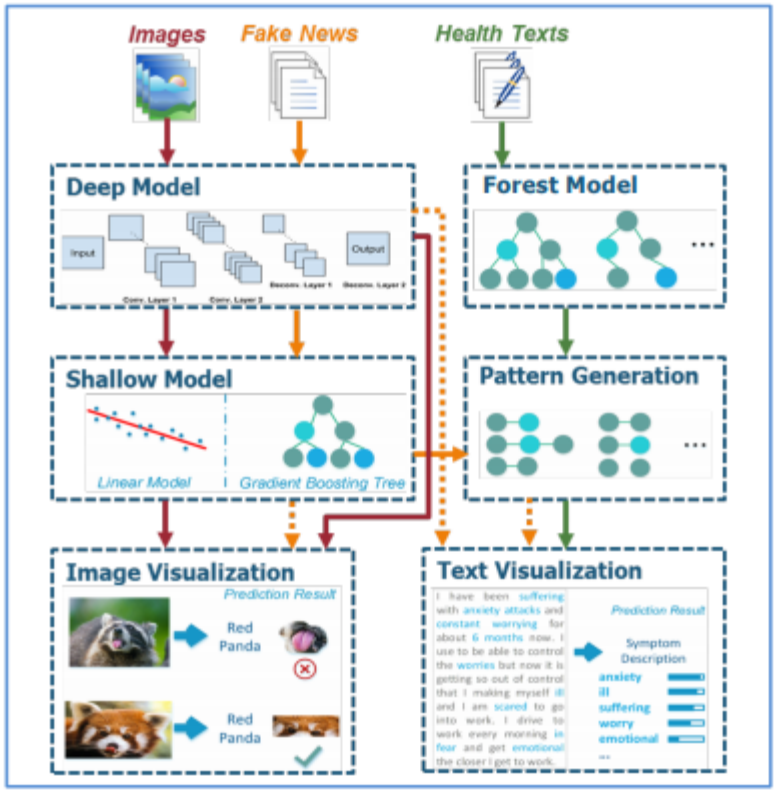
- **PI**: Vibhav Gogate (UT Dallas)

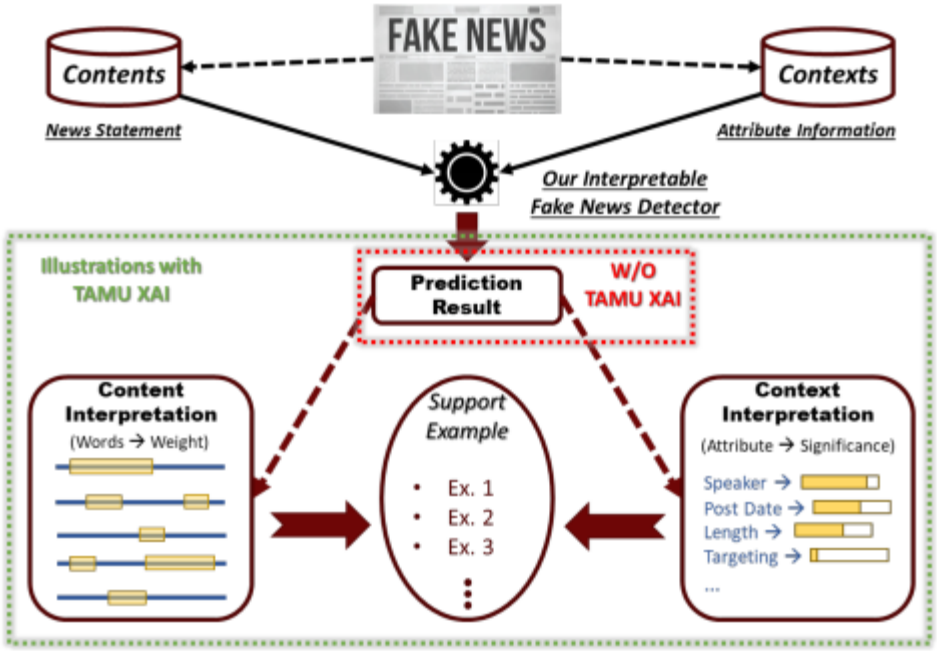
- Adnan Darwiche (UCLA)
- Guy Van Den Broeck (UCLA)
- Nicholas Ruozzi (UT Dallas)
- Eric Ragan (Texas A&M)
- Parag Singla (IIT-Delhi)

https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

**UT Dallas**

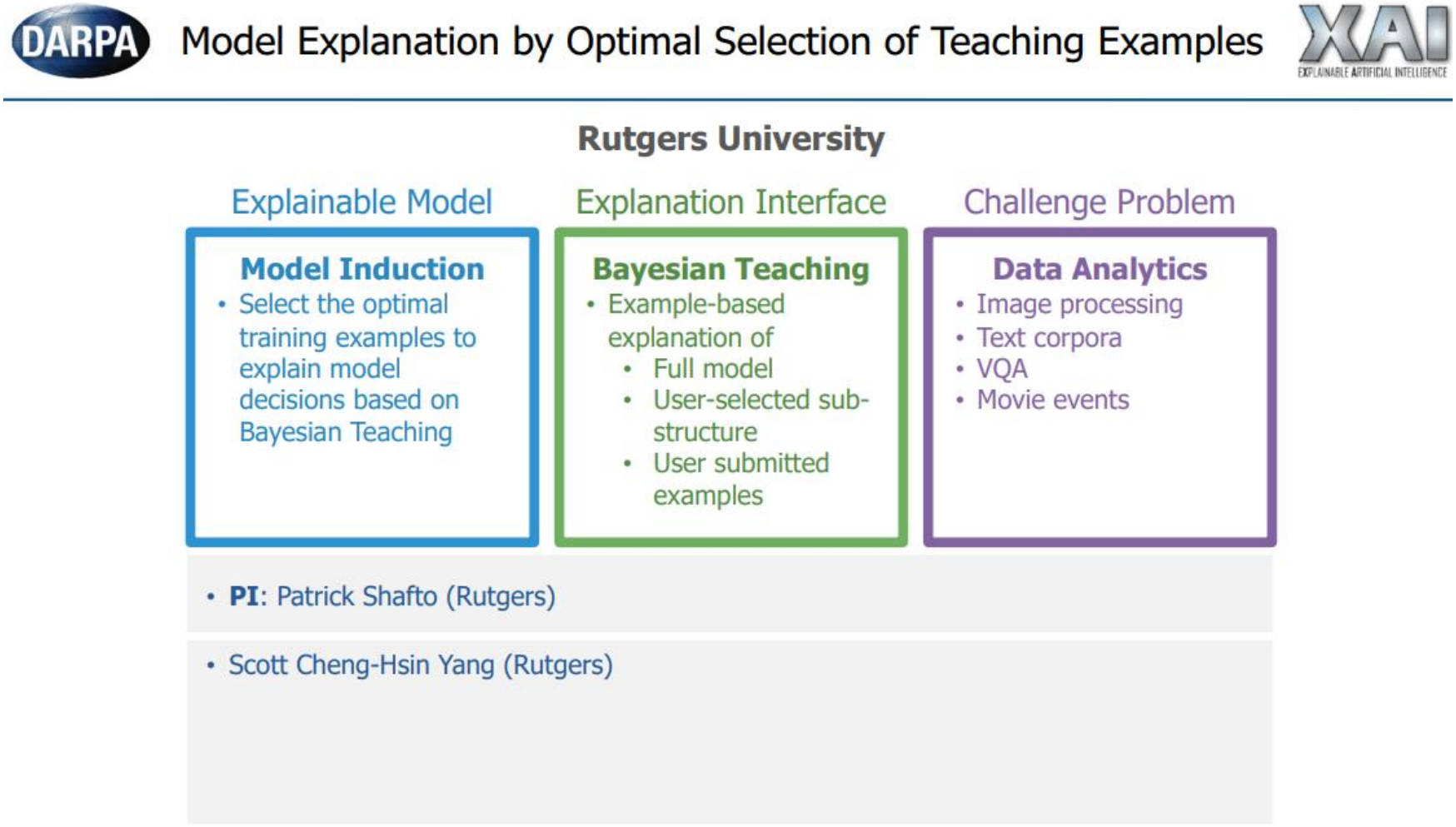
Develop an end-to-end interpretable deep learning infrastructure with image and text datasets



https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

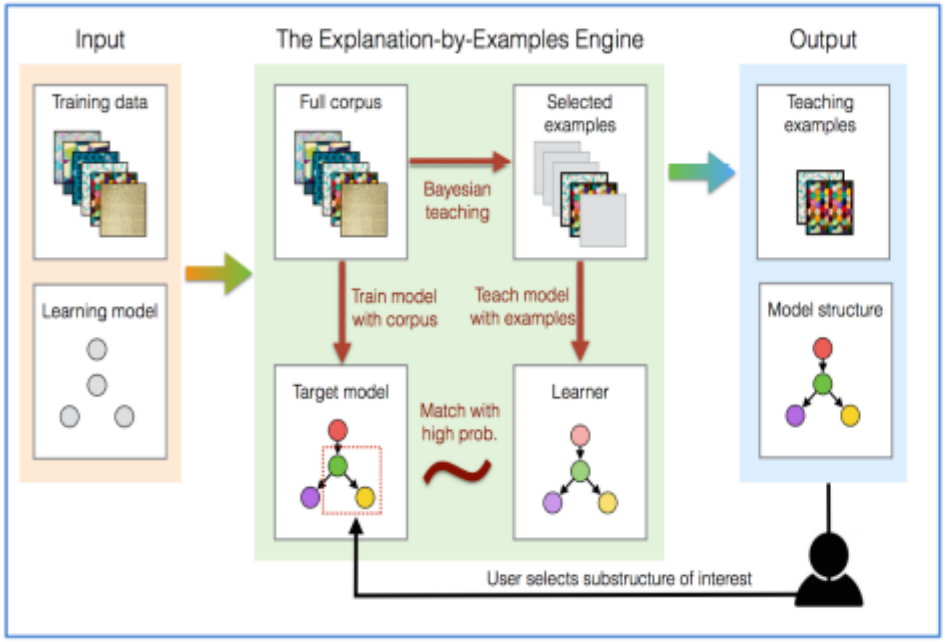
**Rutgers**



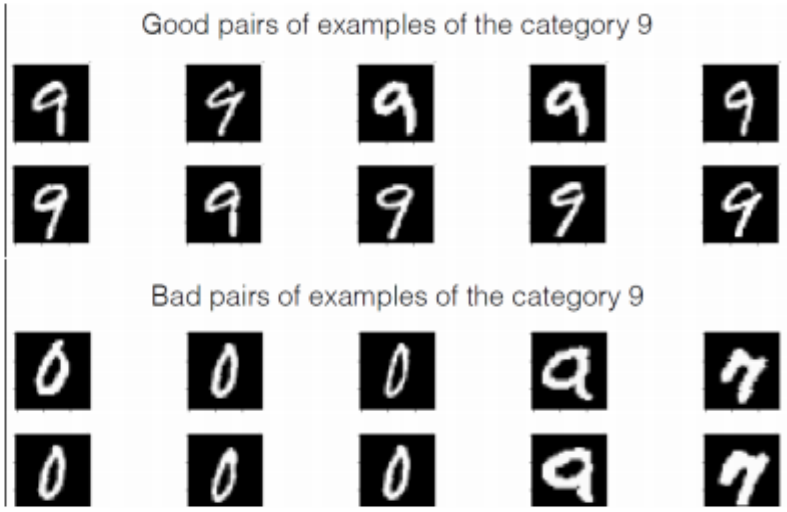
https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf

**Rutgers**

Extend Bayesian teaching to enable automatic explanation by selecting the subset of data that are most representative of the model's generative process




Good and bad examples for teaching a category (illustrates model strengths and weaknesses)



ihmc
FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

https://apps.dtic.mil/sti/pdfs/AD1073994.pdf

Report from TA 2: Psychological Explanation Theories

## What Makes a Good Explanation (Literature Review)

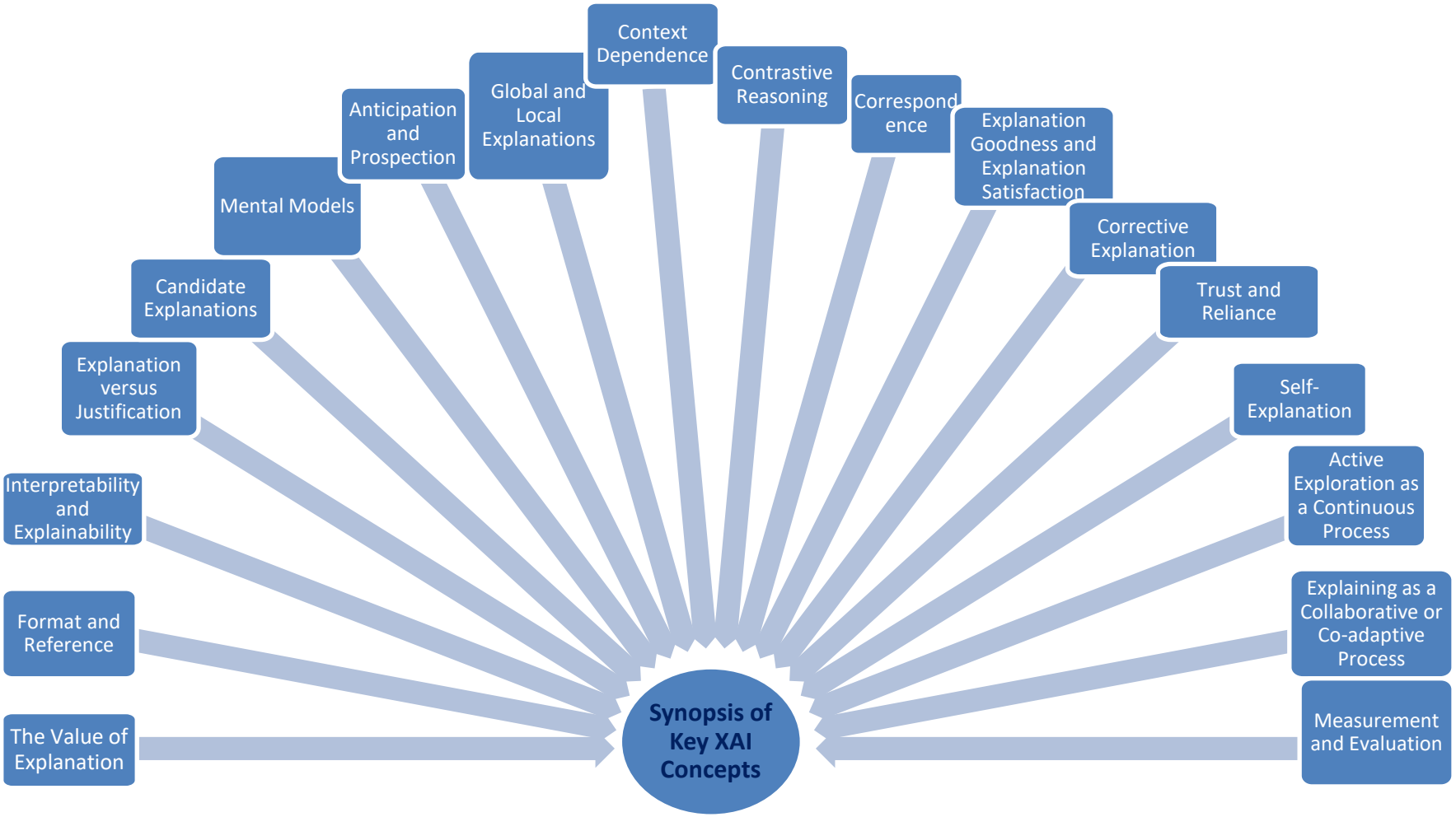
**Recommendation**

Insert detail of empirical/eksperimental methods as in psychological reserach: detail participants, instructions, procedures, tasks, DV, IV and Control

ihmc
FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

https://apps.dtic.mil/sti/pdfs/AD1073994.pdf

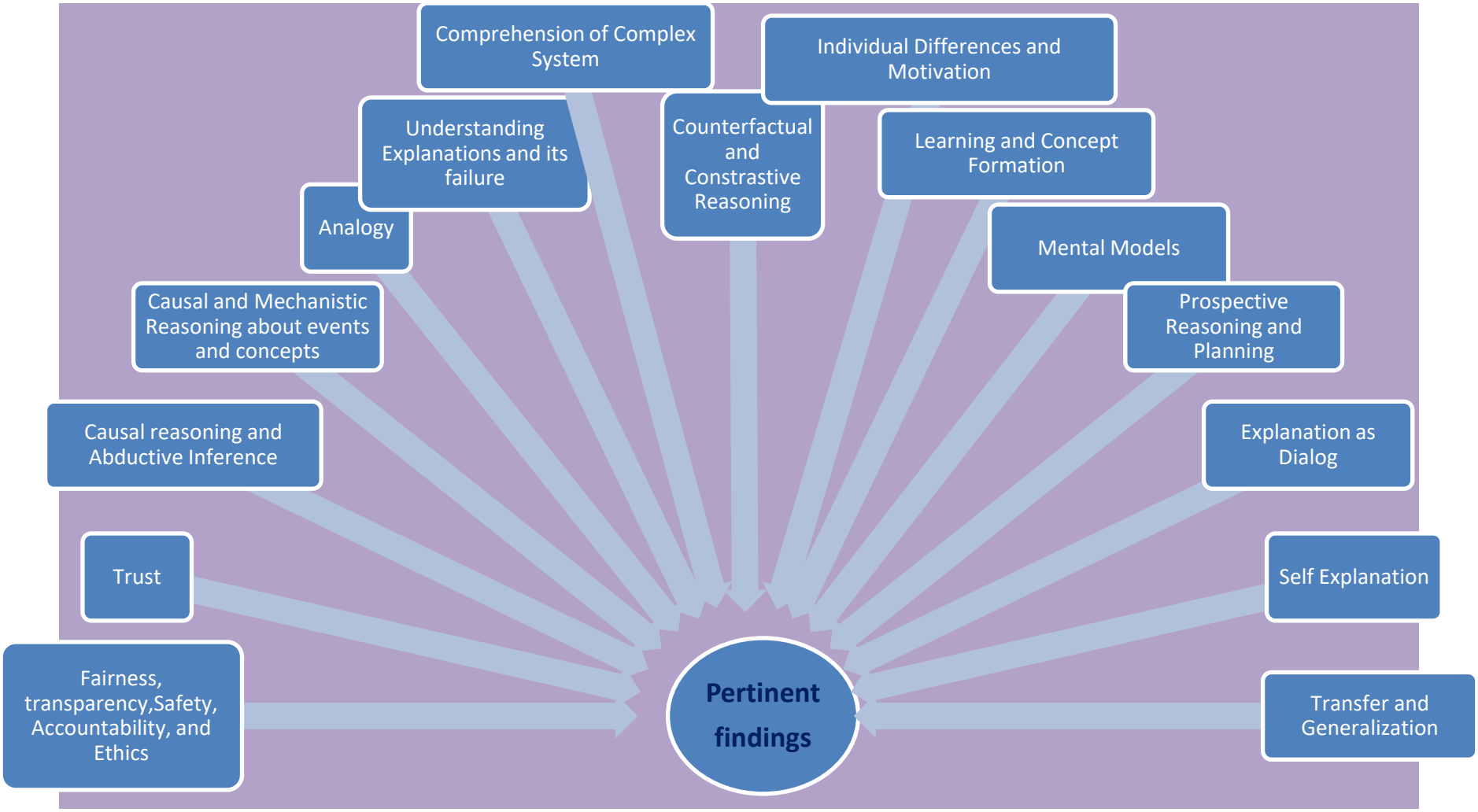
# Report from TA 2: Psychological Explanation Theories



ihmc
FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

https://apps.dtic.mil/sti/pdfs/AD1073994.pdf

# Report from TA 2: Psychological Explanation Theories



# Outline

- Interpretability = Transparency ≠ Explainability?
- DARPA XAI Grant program
- Results: Grantees and reports
- Key takeaways and discussions

- DARPA XAI ”challenge” program has been working in a greater extent in explaining explainable AI compare to other initiatives e.g. EU AI guidelines
- Emphasis more on trusted machine rather than fairness → some regulation issues need more attention?
- Much left to do in psychological explanation areas in the interaction with non end-user  (e.g. tackling the issue of fairness and inequality)

# discussions