Explainable Artificial Intelligence (XAI)
The Need for Explainable AI

- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

- The current generation of AI systems offer tremendous benefits, but their effectiveness will be limited by the machine’s inability to explain its decisions and actions to users
- Explainable AI will be essential if users are to understand, appropriately trust, and effectively manage this incoming generation of artificially intelligent partners

Approved for public release: distribution unlimited.
XAI In the News

MIT Technology Review
The Dark Secret at the Heart of AI
Will Knight
April 11, 2017

The Wall Street Journal
Inside DARPA’s Push to Make Artificial Intelligence Explain Itself
Sara Castellanos and Steven Norton
August 10, 2017

The Register
You better explain yourself, mister: DARPA’s mission to make an accountable AI
Dan Robinson
September 29, 2017

Executive Biz
Charles River Analytics-Led Team Gets DARPA Contract to Support Artificial Intelligence Program
Ramona Adams
June 13, 2017

Entrepreneur
Elon Musk and Mark Zuckerberg Are Arguing About AI -- But They’re Both Missing the Point
Artur Kiulian
July 28, 2017

NOVA Next
Ghosts in the Machine
Christina Couch
October 25, 2017

Jane’s
DARPA’s XAI seeks explanations from autonomous systems
Geoff Fein
November 16, 2017

COMPUTERWORLD
Oracle quietly researching ‘Explainable AI’
George Notter
May 5, 2017

SCIENTIFIC AMERICAN
Demystifying the Black Box That Is AI
Ariel Bleicher
August 9, 2017

Approved for public release: distribution unlimited.
Deep Learning Neural Networks
Architecture and How They Work

Automatic algorithm
(feature extraction and classification)

Deep Learning Neural Network

- Input Layer
- Hidden Layer
  - Low-level features to high-level features
- Output Layer

Training Data

Input (unlabeled image)

1st Layer

- Neurons respond to simple shapes

2nd Layer

- Neurons respond to more complex structures

nth Layer

- Neurons respond to highly complex, abstract concepts

10% Wolf
90% Dog

XenonStack ©

https://www.xenonstack.com/

Approved for public release: distribution unlimited.
What Are We Trying To Do?

**Today**

- **Training Data**
- **Learning Process**
- **Learned Function**
- **Output**
- **User with a Task**

  - Why did you do that?
  - Why not something else?
  - When do you succeed?
  - When do you fail?
  - When can I trust you?
  - How do I correct an error?

This is a cat: (p = .93)

**Tomorrow**

- **Training Data**
- **New Learning Process**
- **Explainable Model**
- **Explanation Interface**
- **User with a Task**

  - I understand why
  - I understand why not
  - I know when you'll succeed
  - I know when you'll fail
  - I know when to trust you
  - I know why you erred
## Challenge Problems

<table>
<thead>
<tr>
<th>Learn a model</th>
<th>Explain decisions</th>
<th>Use the explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Analytics</strong></td>
<td><strong>Autonomy</strong></td>
<td><strong>Explainable Model</strong></td>
</tr>
<tr>
<td>Classification Learning Task</td>
<td>Reinforcement Learning Task</td>
<td>Explanation Interface</td>
</tr>
<tr>
<td>Multimedia Data</td>
<td>ArduPilot &amp; SITL Simulation</td>
<td></td>
</tr>
<tr>
<td>Classifies items of interest in large data set</td>
<td>Learns decision policies for simulated missions</td>
<td>Explains why/why not for recommended items</td>
</tr>
</tbody>
</table>

An analyst is looking for items of interest in massive multimedia data sets.

An operator is directing autonomous systems to accomplish a series of missions.

---

Approved for public release: distribution unlimited.
Goal: Performance and Explainability

- XAI will create a suite of machine learning techniques that
  - Produce more explainable models, while maintaining a high level of learning performance (e.g., prediction accuracy)
  - Enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners

![Graph showing Performance vs. Explainability](image)
Measuring Explanation Effectiveness

Explanation Framework

XAI System
The system takes input from the current task and makes a recommendation, decision, or action

Explainable Model

Explanation Interface

Explanation
The system provides an explanation to the user that justifies its recommendation, decision, or action

Task
Recommendation, Decision or Action

Decision
The user makes a decision based on the explanation

Measure of Explanation Effectiveness

User Satisfaction
- Clarity of the explanation (user rating)
- Utility of the explanation (user rating)

Mental Model
- Understanding individual decisions
- Understanding the overall model
- Strength/weakness assessment
- ‘What will it do’ prediction
- ‘How do I intervene’ prediction

Task Performance
- Does the explanation improve the user’s decision, task performance?
- Artificial decision tasks introduced to diagnose the user’s understanding

Trust Assessment
- Appropriate future use and trust

Correctability (Extra Credit)
- Identifying errors
- Correcting errors
- Continuous training

Approved for public release: distribution unlimited.
Performance vs. Explainability

Learning Techniques (today)
- Neural Nets
- Statistical Models
- SVMs
- Deep Learning
- Graphical Models
- Bayesian Belief Nets
- Decision Trees
- Ensemble Methods
- SRL
- CRFs
- HBNs
- MLNs
- Markov Models
- Random Forests
- SRL
- CRFs
- HBNs
- MLNs
- Markov Models

Explainability (notional)

Approved for public release: distribution unlimited.
Performance vs. Explainability

New Approach

Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance.

Learning Techniques (today)

- Neural Nets
- Graphical Models
- Ensemble Methods
- Bayesian Belief Nets
- Random Forests
- Markov Models
- Decision Trees
- Deep Learning
- Statistical Models
- AOGs
- SVMs

Explainability (notional)

Deep Explanation

Modified deep learning techniques to learn explainable features

Interpretable Models

Techniques to learn more structured, interpretable, causal models

Model Induction

Techniques to infer an explainable model from any model as a black box

Approved for public release: distribution unlimited.
**XAI Concept and Technical Approaches**

**New Learning Process**

**Explainable Model**

**Explanation Interface**

---

<table>
<thead>
<tr>
<th>Training Data</th>
<th>New Learning Process</th>
<th>Explainable Model</th>
<th>Explanation Interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC Berkeley</td>
<td>Deep Learning</td>
<td>Reflexive and Rational</td>
<td></td>
</tr>
<tr>
<td>Charles River Analytics</td>
<td>Causal Modeling</td>
<td>Narrative Generation</td>
<td></td>
</tr>
<tr>
<td>UCLA</td>
<td>Pattern Theory+</td>
<td>3-Level Explanation</td>
<td></td>
</tr>
<tr>
<td>Oregon State</td>
<td>Adaptive Programs</td>
<td>Acceptance Testing</td>
<td></td>
</tr>
<tr>
<td>PARC</td>
<td>Cognitive Modeling</td>
<td>Interactive Training</td>
<td></td>
</tr>
<tr>
<td>CMU</td>
<td>Explainable RL (XRL)</td>
<td>XRL Interaction</td>
<td></td>
</tr>
<tr>
<td>SRI International</td>
<td>Deep Learning</td>
<td>Show and Tell Explanations</td>
<td></td>
</tr>
<tr>
<td>Raytheon BBN</td>
<td>Deep Learning</td>
<td>Argumentation and Pedagogy</td>
<td></td>
</tr>
<tr>
<td>UT Dallas</td>
<td>Probabilistic Logic</td>
<td>Decision Diagrams</td>
<td></td>
</tr>
<tr>
<td>Texas A&amp;M</td>
<td>Mimic Learning</td>
<td>Interactive Visualization</td>
<td></td>
</tr>
<tr>
<td>Rutgers</td>
<td>Model Induction</td>
<td>Bayesian Teaching</td>
<td></td>
</tr>
</tbody>
</table>

---

**IHMC Psychological Model of Explanation**

- Explanation
  - Explanation Quality
  - User Satisfaction
- User
  - User's Mental Model
  - Better Performance
- Trust or Mistrust
  - User Comprehension
  - Appropriate Trust
  - Appropriate Use

---

Approved for public release: distribution unlimited.
Approaches to Deep Explanation
(Berkeley, SRI, Raytheon BBN, OSU, CRA, PARC)

Attention Mechanisms

Top-down Caption Saliency
[Ramanishka et al. CVPR17]
Caption: A man in a jacket is standing at the slot machine

Feature Identification

Modular Networks

Neural module networks
[Andreas et al. CVPR16, EMNLP16] [Hu et al. CVPR17]
Q: Can you park here? Prediction
Neural module network
Decision path

Learn to Explain

Downy Woodpecker Definition:
This bird has a white breast, black wings, and a red spot on its head.

Image Explanation:
This is a Downy Woodpecker because it is a black and white bird with a red spot on its crown.
Network Dissection Quantifying Interpretability of Deep Representations (MIT)

Audit trail: for a particular output unit, the drawing shows the most strongly activated path

Interpretation of several units in pool5 of AlexNet trained for place recognition.
Causal Model Induction (CRA): Experiment with the learned model (as a grey box) to learn an explainable, causal, probabilistic programming model.
Explanation by Selection of Teaching Examples

Bayesian Teaching for optimal selection of examples for machine explanation

Training Data

This face is Angry because it is similar to these examples and dissimilar to these examples.

Bayesian Teaching

Approved for public release: distribution unlimited.
Common Ground Learning and Explanation (COGLE)
An interactive sensemaking system to explain the learned performance capabilities of a UAS flying in an ArduPilot simulation testbed

Common Ground Builder
- Explain
- Train
- Evaluate

COGLE

EXPLANATION LAYER
- Explanations for mission performance and for assessing skills, risks & coverage.

COGNITIVE LAYER
- Cast learned abstractions, policies & clusters into explainable form.

LEARNING LAYER
- Learn policies from the sensed world.

TEST BED LAYER

Deep Adaptive Program

Decision Net

Procedure main()

......

AdaptiveChoice

strategyChoice();

......

Explanation-Informed Acceptance Testing of Deep Adaptive Programs (xACT)
Tools for explaining deep adaptive programs and discovering best principles for designing explanation user interfaces

xFSM

xNN

Game Engine

Saliency Visualizer

Interactive Naming Interface

Visual Words

Robotics Curriculum

Series 1. Primitives: Navigating with Constraints and Lookahead
- Lesson 1.1: Taking off .................................................. 7
- Lesson 1.2: Taking off and Landing .................................. 9
- Lesson 1.3: Reconnaissance Over a Point (3 Months) .......... 11
- Lesson 1.4: Looking Ahead to Avoid Crashing into Mountains .... 13
- Lesson 1.5: Choosing a Safe Descent Approach for Landing ....... 15
- Lesson 1.6: Provisioning a Hiker (6 months) ....................... 17

Series 2. Behaviors: Managing Competing Goals and Foraging
- Lesson 2.1: Provisioning a Hiker in a Box Canyon (opt) .......... 19
- Lesson 2.2: Taking an Inventory of a Region and Refueling (opt) .. 22
- Lesson 2.3: Foraging Around a Point for a Hiker (opt) ............ 24
- Lesson 2.4: Foraging Around a Point with an Interfering Obstacle .... 26

Series 3. Missions: Harder Missions and Heavy Testing
- Lesson 3.1: Double Hiker Jeopardy (9 months) .................... 28
- Lesson 3.2: Bear on the Runway ...................................... 30
- Lesson 3.3: Auto-Generated Missions with Testing (12 months) .. 32
DARPA

Four Modes of Explanation (Raytheon BBN)

Explanation Modes

**Analytic (didactic) statements**
in natural language that describe the elements and context that support a choice

**Visualizations**
that directly highlight portions of the raw data that support a choice and allow viewers to form their own perceptual understanding

**Cases**
that invoke specific examples or stories that support the choice

**Rejections of alternative choices**
(or “common misconceptions” in pedagogy) that argue against less preferred answers based on analytics, cases, and data
**XAI Program Structure**

**Evaluator**

**Naval Research Laboratory**

---

### Challenge Problem Areas

**TA1: Explainable Learners**
- Teams that provide prototype systems with both components
  - Explainable Model
  - Explanation Interface

**TA2: Psychological Model of Explanation**
- Psychological Theory of Explanation
- Computational Model Consulting

---

**Data Analytics**
- Multimedia Data

**Autonomy**
- ArduPilot & SITL Simulation

---

**Evaluation Framework**

- Learning Performance
  - Explanation Effectiveness
  - Explanation Measures
    - User Satisfaction
    - Mental Model
    - Task Performance
    - Trust Assessment
    - Correctability

---

- **TA1: Explainable Learners**
  - Multiple TA1 teams will develop prototype explainable learning systems that include both an explainable model and an explanation interface

- **TA2: Psychological Model of Explanation**
  - At least one TA2 team will summarize current psychological theories of explanation and develop a computational model of explanation from those theories
Challenge Problem Candidates

Analytics

- Visual Question Answering
  - MovieQA
  - CLEVR

- Activity Recognition
  - ActivityNet

Autonomy

- Strategy Games
  - Starcraft2
  - ELF-MiniRTS

- Vehicle Control
  - ArduPilot
  - Driving Simulator

Approved for public release: distribution unlimited.
Psychological Model of Explanation (IHMC)

Model of the Explanation Process and Possible Metrics

- **XAI Process**
- **XAI Metrics**

**User** receives **Explanation**
- **User’s Mental Model**
  - **Test of Comprehension**
  - **Test of Satisfaction**
  - **“Goodness” Criteria**

**Appropriate Trust**
- **Better Performance**
  - **Test of Performance**

- **Trust or Mistrust**
  - **Appropriate Use**
    - **System**
      - **is assessed by**
        - **User**
          - **receives**
            - **Explanation**
              - **revises**
                - **User’s Mental Model**
                  - **enables**
                    - **Better Performance**
                      - **Test of Performance**

- **may initially**
  - **can engender**
    - **given way to**
      - **appropriate Use**

Approved for public release: distribution unlimited.
### Technical Area 1 (Explainable Learners) Milestones
- Demonstrate the explainable learners against problems proposed by the developers (Phase 1)
- Demonstrate the explainable learners against common problems (Phase 2)
- Deliver software libraries and toolkits (at the end of Phase 2)

### Technical Area 2 (Psychology of Explanation) Milestones
- Deliver an interim report on psychological theories (after 6 months during Phase 1)
- Deliver a final report on psychological theories (after 12 months, during Phase 1)
- Deliver a computational model of explanation (after 24 months, during Phase 2)
- Deliver the computational model software (at the end of Phase 2)
<table>
<thead>
<tr>
<th>CP</th>
<th>Performer</th>
<th>Explainable Model</th>
<th>Explanation Interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both</td>
<td>UC Berkeley</td>
<td>Deep Learning</td>
<td>Reflexive and Rational</td>
</tr>
<tr>
<td></td>
<td>Charles River</td>
<td>Causal Modeling</td>
<td>Narrative Generation</td>
</tr>
<tr>
<td></td>
<td>UCLA</td>
<td>Pattern Theory+</td>
<td>3-level Explanation</td>
</tr>
<tr>
<td>Autonomy</td>
<td>Oregon State</td>
<td>Adaptive Programs</td>
<td>Acceptance Testing</td>
</tr>
<tr>
<td></td>
<td>PARC</td>
<td>Cognitive Modeling</td>
<td>Interactive Training</td>
</tr>
<tr>
<td></td>
<td>CMU</td>
<td>Explainable RL (XRL)</td>
<td>XRL Interaction</td>
</tr>
<tr>
<td>Analytics</td>
<td>SRI International</td>
<td>Deep Learning</td>
<td>Show and Tell Explanation</td>
</tr>
<tr>
<td></td>
<td>Raytheon BBN</td>
<td>Deep Learning</td>
<td>Argumentation and Pedagogy</td>
</tr>
<tr>
<td></td>
<td>UT Dallas</td>
<td>Probabilistic Logic</td>
<td>Decision Diagrams</td>
</tr>
<tr>
<td></td>
<td>Texas A&amp;M</td>
<td>Mimic Learning</td>
<td>Interactive Visualization</td>
</tr>
<tr>
<td></td>
<td>Rutgers</td>
<td>Model Induction</td>
<td>Bayesian Teaching</td>
</tr>
</tbody>
</table>
Deeply Explainable Artificial Intelligence

**Explainable Model**

**Deep Learning**
- Explain *implicit* (latent) nodes by training additional DL models
- Explain *explicit* nodes thru Neural Module Networks (NMNs)

**Explanation Interface**

**Reflexive & Rational**
- Reflexive explanations (that arise directly from the model)
- Rational explanations (that come from reasoning about user’s beliefs)

**Challenge Problem**

**Autonomy**
- ArduPilot and OpenAI Gym Simulations

**Data Analytics**
- Visual QA and Multimedia Event QA

**PI**: Trevor Darrell (Berkeley)

- Pieter Abbeel (Berkeley)
- Tom Griffiths (Berkeley)
- Kate Saenko (BU)
- Zeynep Akata (U. Amsterdam)
- Dan Klein (Berkeley)
- John Canny (Berkeley)
- Anca Dragan (Berkeley)
- Anthony Hoogs (Kitware)
CAMEL: Causal Models to Explain Learning

**Explainable Model**

Model Induction Causal Models
- Experiment with the learned model (as a grey box) to learn an explainable, causal, probabilistic programming model

**Explanation Interface**

Narrative Generation
- Interactive visualization based on the generation of temporal, spatial narratives from the causal, probabilistic models

**Challenge Problem**

Autonomy
- Minecraft, Starcraft

Data Analytics
- Pedestrian Detection (INRIA), Activity Recognition (ActivityNet)

- **PI**: Brian Ruttenberg (CRA)
- Avi Pfeffer (CRA)
- David Jensen (U. Mass)
- Michael Littman (Brown)
- James Niehaus (CRA)
- Emilie Roth (Roth Cognitive Engineering)
- Joe Gorman (CRA)
- James Tittle (CRA)
Learning and Communicating Explainable Representations for Analytics and Autonomy

**Pattern Theory+**
- Integrated representation across an entropy spectrum:
  - Deep Neural Nets
  - Stochastic And-Or-Graphs (AOG)
  - Predicate Calculus

**3-Level Explanation**
- Integrate 3 levels of explanation:
  - Concept compositions
  - Causal and counterfactual reasoning
  - Utility explanations

**Challenge Problem**
- **Autonomy**
  - Humanoid robot behavior and VR simulation platform
- **Data Analytics**
  - Understanding complex multimedia events

**PI:** Song-Chun Zhu (UCLA)
- Ying Nian Wu (UCLA)
- Sinisa Todorovic (OSU)
- Joyce Chai (Michigan State)
**xACT: Explanation-Informed Acceptance Testing of Deep Adaptive Programs**

**Explainable Model**

**Adaptive Programs**
- Explainable Deep Adaptive Programs (xDAPs) – a new combination of Adaptive Programs, Deep Learning and explainability

**Explanation Interface**

**Acceptance Testing**
- Provides a visual & NL explanation interface for acceptance testing by test pilots based on Information Foraging Theory

**Challenge Problem**

**Autonomy**
- Real-Time Strategy Games based on custom designed game engine designed to support explanation
- Possible use of Starcraft

**PI**: Alan Fern (OSU)

- Tom Dietterich (OSU)
- Fuxin Li (OSU)
- Prasad Tadepalli (OSU)
- Weng-Keen Wong (OSU)

- Margaret Burnett (OSU)
- Martin Erwig (OSU)
- Liang Huang (OSU)

---

Approved for public release: distribution unlimited.
COGLE: Common Ground Learning and Explanation

**Explainable Model**

**Cognitive Model**
- 3-layer architecture:
  - Learning Layer (DNNs)
  - Cognitive Layer (ACT-R Cog. Model)
  - Explanation Layer (HCI)

**Explanation Interface**

**Interactive Training**
- Interactive visualization of states, actions, policies & values
- Includes a module for test pilots to refine and train the system

**Challenge Problem**

**Autonomy**
- ArduPilot simulation environment
- Value of Explanation (VoE) framework for measuring explanation effectiveness

**PI**: Mark Stefik (PARC)

- Honglak Lee (U. Mich.)
- Subramanian Ramamoorthy (U. Edinburgh)

- Christian Lebiere (CMU)
- John Anderson (CMU)
- Robert Thomson (USMA)

- Michael Youngblood (PARC)
XRL: Explainable Reinforcement Learning for AI Autonomy

**XRL Models**
- Create a new scientific discipline for Explainable Reinforcement Learning with work on new algorithms and representations

**XRL Interaction**
- Interactive explanations of dynamic systems
- Human-machine interaction to improve performance

**Challenge Problem**
- Open AI Gym
- Autonomy in the electrical grid
- Mobile service robots
- Self-improving educational software

**PI**: Geoff Gordon (CMU)

- Zico Kolter (CMU)
- Pradeep Ravikumar (CMU)

- Manuela Veloso (CMU)
- Emma Brunskill (Stanford)
DARE: Deep Attention-based Representations for Explanation

**Deep Learning**
- Multiple deep learning techniques:
  - Attention-based mechanisms
  - Compositional NMNs
  - GANs

**Show-and-Tell Explanations**
- DNN visualization
- Query evidence that explains DNN decisions
- Generate natural language justifications

**Challenge Problem**
- Visual Question Answering (VQA) using Visual Gnome, Flickr30
- MovieQA

**PIs**
- Giedrius Burachas (SRI), Mohamed Amer (SRI)
- Shalini Ghosh (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)
EQUAS: Explainable Question Answering System

**Deep Learning**
- Semantic labelling of DNN neurons
- DNN audit trail construction
- Gradient-weighted Class Activation Mapping

**Explanation Interface**
- **Argumentation Theory**
  - Comprehensive strategy based on argumentation theory
  - NL generation
  - DNN visualization

**Challenge Problem**
- **Data Analytics**
  - Visual Question Answering (VQA), beginning with images and progressing to video

**PI**: William Ferguson (Raytheon BBN)

- Antonio Torralba (MIT)
- Ray Mooney (UT Austin)
- Devi Parikh (GA Tech)
- Dhruv Batra (GA Tech)
Tractable Probabilistic Logic Models: A New, Deep Explainable Representation

### Explainable Model

**Probabilistic Logic**
- Tractable Probabilistic Logic Models (TPLMs)
  - an important class of (non-deep learning) interpretable models

### Explanation Interface

**Probabilistic Decision Diagrams**
- Enables users to explore and correct the underlying model as well as add background knowledge

### Challenge Problem

**Data Analytics**
- Infer activities in multimodal data (video and text)
- Using the 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)
Transforming Deep Learning to Harness the Interpretability of Shallow Models: An Interactive End-to-End System

**Mimic Learning**
- Develop a mimic learning framework that combines deep learning models for prediction and shallow models for explanations

**Interactive Visualization**
- Interactive visualization over multiple views, using heat maps & topic modeling clusters to show predictive features

**Data Analytics**
- Multiple tasks using data from Twitter, Facebook, ImageNet, UCI, NIST and Kaggle
- Metrics for explanation effectiveness

**PI:** Xia Hu (Texas A&M)

**Shuiwang Ji** (Wash. State)  
**Eric Ragan** (Texas A&M)
Model Explanation by Optimal Selection of Teaching Examples

**Explainable Model**
- **Model Induction**
  - Select the optimal training examples to explain model decisions based on Bayesian Teaching

**Explanation Interface**
- **Bayesian Teaching**
  - Example-based explanation of:
    - the full model
    - user-selected sub-structure
    - user submitted examples

**Challenge Problem**
- **Data Analytics**
  - Movie descriptions
  - Image processing
  - Caption data
  - Movie events
  - Human motion events

- **PI**: Patrick Shafto (Rutgers)

- Scott Cheng-Hsin Yang (Rutgers)
Naturalistic Decision Making Foundations of Explainable AI

**Literature Review**
- Extensive review of relevant psychological theories
- Extend the theory of Naturalistic Decision Making to cover explanation

**Computational Model**
- Represent reductionist mental models that humans develop as part of the explanatory process
- Including mental simulation

**Model Validation**
- Conduct interactive assessment and formal human experiments
- Validate the model
- Develop metrics of explanation effectiveness

**Experiments**

**Naturalistic Theory**
- Gary Klein (MacroCognition)
- Shane T. Mueller (Michigan Tech)

**Bayesian Framework**
- William J. Clancey (IHMC)
- COL Timothy M. Cullen (SAASS)

**Model Validation**
- Jordan Litman (IHMC Psychometrician)
- Simon Attfield (Middlesex University-London)
- Peter Pirolli (IHMC)

**PI:** Robert R. Hoffman (IHMC)
**XAI Evaluation**

**Challenge Problems**
- Analytics
- Autonomy

**Evaluation Framework**
- Evaluation protocols
- Training environment
  - Training data
  - Simulation environment
- Testing environment
  - Subjects
  - Web infrastructure
- Baseline systems

**Measurement**

- PI: David Aha (NRL)
- Justin Karneeb (Knexus)
- Matt Molineaux (Knexus)
- Leslie Smith (NRL)
- Mike Pazzani (UC Riverside)