Explainable AI – The Story So Far

August 29th, 2019 - @Inha University

Freddy Lecue
Chief AI Scientist, CortAlx, Thales, Montreal – Canada
Inria, Sophia Antipolis - France

@freddylecue
https://tinyurl.com/freddylecue
Motivation
Gary Chavez added a photo you might be in.
about a minute ago · 📸
Critical Systems
But not Only Critical Systems
Motivation (1)

Criminal Justice
- People wrongly denied parole
- Recidivism prediction
- Unfair Police dispatch

Opinion

When a Computer Program Keeps You in Jail
By Rebecca Wexler
June 13, 2017
nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

How We Analyzed the COMPAS Recidivism Algorithm
by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin
May 23, 2016
propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

STATEMENT OF CONCERN ABOUT PREDICTIVE POLICING BY ACLU AND 16 CIVIL RIGHTS PRIVACY, RACIAL JUSTICE, AND TECHNOLOGY ORGANIZATIONS
aclu.org/other/statement-concern-about-predictive-policing-aclu-and-16-civil-rights-privacy-racial-justice
Motivation (2)

- Finance:
  - Credit scoring, loan approval
  - Insurance quotes

The Big Read Artificial intelligence

Insurance: Robots learn the business of covering risk

Artificial intelligence could revolutionise the industry but may also allow clients to calculate if they need protection

Oliver Ralph May 16, 2017

https://www.ft.com/content/e07cee0c-3949-11e7-821a-6027b8a20f23

Community.fico.com/s/explainable-machine-learning-challenge
Motivation (3)

Healthcare

- Applying ML methods in medical care is problematic.
- AI as 3rd-party actor in physician-patient relationship
- Responsibility, confidentiality?
- Learning must be done with available data.
  Cannot randomize cares given to patients!
- Must validate models before use.

Researchers say use of artificial intelligence in medicine raises ethical questions

In a perspective piece, Stanford researchers discuss the ethical implications of using machine-learning tools in making health care decisions for patients.


Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission

Rich Caruana Microsoft Research rcaruana@microsoft.com
Yin Lou LinkedIn Corporation ylou@linkedin.com
Paul Koch Microsoft Research paulkoch@microsoft.com
Marc Sturm NewYork-Presbyterian Hospital mas9161@nyp.org
Johannes Gehrke Microsoft johannes@microsoft.com
Noémie Elhadad Columbia University noemie.elhadad@columbia.edu

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, Noémie Elhadad: Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. KDD 2015: 1721-1730
The need for explainable AI rises with the potential cost of poor decisions

- **Industrial / Military**
  - Cyber Threat Detection
  - Jet Engine Predictive Maintenance
  - Self-Driving Vehicles
  - Flight Trajectory Optimization

- **Enterprise**
  - Quality Inspection
  - Medical Diagnosis
  - Project Risk Monitoring
  - Credit Risk Profiling
  - Incident Investigation
  - Fraud Detection
  - Case Load Processing
  - Product Pricing
  - Auditing
  - Production Scheduling

- **Consumer**
  - Machine Translation
  - Face Recognition
  - Speech Recognition
  - Friends Recommendations
  - Search Result Ranking
  - Ad Placement

- **Professional**
  - Fashion Recommendation
  - Medical Image Interpretation
  - Mentor Recommendation
  - Automated Trading
  - Data Labeling
  - Fitness Coaching
  - Compliance Monitoring

Most prominent successes of AI to date

Most impactful successes of AI to come

Source: Accenture Point of View. Understanding Machines: Explainable AI. Freddy Lecue, Dadong Wan
XAI in a Nutshell
XAI in a Nutshell

Today

- 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?

Tomorrow

- 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

Source: https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf
How to Explain? Accuracy vs. Explanability

- **Challenges:**
  - Supervised
  - Unsupervised learning

- **Approach:**
  - Representation Learning
  - Stochastic selection

- **Output:**
  - **Correlation**
  - **No causation**
Trustable AI
AI Adoption: Requirements

- Valid AI
- Responsible AI
- Explainable AI
- Privacy-preserving AI

Trustable AI

What is the rational?
Definitions
Explanation in AI aims to create a suite of techniques that produce more explainable models, while maintaining a high level of searching, learning, planning, reasoning performance: optimization, accuracy, precision; and enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI systems.
explanation | ˌekspləˈneɪʃ(ə)n |

noun

a statement or account that makes something clear: the birth rate is central to any explanation of population trends.

interpret | ɪnˈtɛːprɪt |

verb (interprets, interpreting, interpreted) [with object]

1 explain the meaning of (information or actions): the evidence is difficult to interpret.
XAI in AI
XAI: One Objective, Many ‘Al’s, Many Definitions, Many Approaches

- Artificial Intelligence
  - Machine Learning
  - Game Theory
  - Planning
  - Search
  - Robotics
  - KRR
  - UAI
  - MAS
  - Computer Vision
  - NLP
XAI: One Objective, Many ‘Al’s, Many Definitions, Many Approaches

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?
Which features are responsible of classification?

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

Which complex features are responsible of classification?

Which features are responsible of classification?

Artificial Intelligence

Dependency Plot

Feature Importance

Surrogate Model

Machine Learning

Planning

Search

Game Theory

Robotics

Computer Vision

KRR

UAI

NLP

Saliency Map

Uncertainty Map

Game Theory

Dependency Plot

Feature Importance

Surrogate Model

Machine Learning

Planning

Search

Game Theory

Robotics

Computer Vision

KRR

UAI

NLP

Saliency Map

Uncertainty Map

This document may not be reproduced, modified, adapted, published, translated, in any way, in whole or in part or disclosed to a third party without the prior written consent of Thales.
Which features are responsible of classification?

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

- Which agent strategy & plan?
- Which player contributes most?
- Why such a conversational flow?

Which complex features are responsible of classification?
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

Which features are responsible of classification?
Which actions are responsible of a plan?
Which agent strategy & plan?
Which player contributes most?
Why such a conversational flow?

Computer Vision

Machine Learning

Surrogate Model

Feature Importance

Dependency Plot

Artificial Intelligence

MAS

Plan Refinement

Search

Game Theory

Robotics

Uncertainty Map

KRR

UAI

NLP

Dependency Plot

Feature Importance

Surrogate Model

Saliency Map

Plan Refinement

Which complex features are responsible of classification?
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

Which features are responsible of classification?

Which actions are responsible of a plan?

Which constraints can be relaxed?

Which complex features are responsible of classification?

Which agent strategy & plan?

Which player contributes most?

Why such a conversational flow?

Plan Refinement

Strategy Summarization

Conflicts Resolution

Surrogate Model

Dependency Plot

Feature Importance

Artificial Intelligence

MAS

Computer Vision

XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

Saliency Map

Dependency Plot

Feature Importance

Surrogate Model

Artificial Intelligence

MAS

Computer Vision

XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

Saliency Map

Dependency Plot

Feature Importance

Surrogate Model

Artificial Intelligence

MAS

Computer Vision

XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

Saliency Map

Dependency Plot

Feature Importance

Surrogate Model

Artificial Intelligence

MAS

Computer Vision

XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

Saliency Map

Dependency Plot

Feature Importance

Surrogate Model

Artificial Intelligence

MAS

Computer Vision

XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

Saliency Map

Dependency Plot

Feature Importance

Surrogate Model

Artificial Intelligence

MAS

Computer Vision

XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

Saliency Map

Dependency Plot

Feature Importance

Surrogate Model

Artificial Intelligence

MAS

Computer Vision
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

Which combination of features is optimal?
Which features are responsible of classification?
Which actions are responsible of a plan?
Which constraints can be relaxed?

Artificial Intelligence

MAS

Computer Vision

Dependency Plot
Feature Importance
Surrogate Model

Plan Importance
Plan Refinement

Plan Refinement

KRR

UAI

NLP

Search

Game Theory

Robotics

Machine Learning

Saliency Map

Dependency Plot
Feature Importance
Surrogate Model

Saliency Map

Surrogate Model

Computer Vision

Search

Game Theory

Robotics

Which complex features are responsible of classification?

Which agent strategy & plan?
Which player contributes most?
Why such a conversational flow?

Plan Refinement

Conflicts Resolution

Which agent strategy & plan?
Which player contributes most?
Why such a conversational flow?

Robotics

NLP

MAS

Computer Vision

Game Theory

Which combination of features is optimal?
Which features are responsible of classification?
Which actions are responsible of a plan?
Which constraints can be relaxed?

Artificial Intelligence

MAS

Computer Vision

Dependency Plot
Feature Importance
Surrogate Model

Saliency Map

Plan Importance
Plan Refinement

Plan Refinement

KRR

UAI

NLP

Search

Game Theory

Robotics

Machine Learning

Saliency Map
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

- Which combination of features is optimal?
- Which features are responsible of classification?
- Which actions are responsible of a plan?
- Which constraints can be relaxed?
- How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?
- Which decisions, combination of multimodal decisions lead to an action?
- Which complex features are responsible of classification?
- Which agent strategy & plan?
- Which player contributes most?
- Why such a conversational flow?
Which combination of features is optimal?

Which features are responsible of classification?

Which actions are responsible of a plan?

Which constraints can be relaxed?

Which decisions, combination of multimodal decisions lead to an action?
Which complex features are responsible of classification?

Which actions are responsible of a plan?

Which entity is responsible for classification?

Which combination of features is optimal?

Which constraints can be relaxed?

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

Which agent strategy & plan?

Which player contributes most?

Why such a conversational flow?

XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

- Which decisions, combination of multimodal decisions lead to an action?
- Which agent strategy & plan?
- Which player contributes most?
- Why such a conversational flow?

- Which complex features are responsible of classification?
- Which entity is responsible for classification?
- Which combination of features is optimal?
- Which constraints can be relaxed?

- Artificial Intelligence
- MAS
- Computer Vision
- KRR
- UAI

- Strategy Summarization
- Uncertainty Map
- Machine Learning based
- NLP

- Feature Importance
- Surrogate Model
- Dependency Plot
- Machine Learning
- Search
- Planning
- Conflicts Resolution
- Game Theory
- Robotics
- Shapely Values
- Narrative-based
- Machine Learning based Narrative-based Robotics

How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?

- Abduction/Diagnostic: Find the right root causes (abduction)?
- Which axiom is responsible of inference (e.g., classification)?
- Which complex features are responsible of classification?
- Which entity is responsible for classification?

- Abduction
- Diagnosis
- Plan Refinement
- Search
- which actions are responsible of a plan?
Deep Dive
Overview of explanation in different AI fields (1)

Machine Learning (except Artificial Neural Network)

Interpretable Models:
- Linear regression,
- Logistic regression,
- Decision Tree,
- GLMs,
- GAMs
- KNNs
Overview of explanation in different AI fields (1)

### Machine Learning (except Artificial Neural Network)

**Interpretable Models:**
- Linear regression,
- Logistic regression,
- Decision Tree,
- GLMs,
- GAMs,
- KNNs

---

**Naive Bayes model**

Overview of explanation in different AI fields (1)

Machine Learning (except Artificial Neural Network)

Interpretable Models:
- Linear regression,
- Logistic regression,
- Decision Tree,
- GLMs,
- GAMs
- KNNs

Counterfactual
What-if


Naive Bayes model
Overview of explanation in different AI fields (1)

### Machine Learning (except Artificial Neural Network)

**Interpretable Models:**
- Linear regression,
- Logistic regression,
- Decision Tree,
- GLMs,
- GAMs,
- KNNs

**Counterfactual What-if**


**Naive Bayes model**


---

**Data:** titanic

**Model:** NB

**Prediction:** $p(\text{survived} = \text{yes}|x) = 0.671$

**Actual class label for this instance:** yes

<table>
<thead>
<tr>
<th>Feature</th>
<th>Contribution</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
<td>-0.344</td>
<td>3rd</td>
</tr>
<tr>
<td>Age</td>
<td>-0.034</td>
<td>adult</td>
</tr>
<tr>
<td>Sex</td>
<td>1.194</td>
<td>female</td>
</tr>
</tbody>
</table>
Overview of explanation in different AI fields (2)

Machine Learning (only Artificial Neural Network)

Network $f(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

Integrated gradients: $x_1 = 1.5, x_2 = -0.5$
DeepLift: $x_1 = 1.5, x_2 = -0.5$
LRP: $x_1 = 1.5, x_2 = -0.5$

Network $g(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$

Integrated gradients: $x_1 = 1.5, x_2 = -0.5$
DeepLift: $x_1 = 2, x_2 = -1$
LRP: $x_1 = 2, x_2 = -1$

Attribution for Deep Network (Integrated gradient-based)


Overview of explanation in different AI fields (2)

Machine Learning (only Artificial Neural Network)

Network $f(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$
- Integrated gradients: $x_1 = 1.5, x_2 = -0.5$
- DeepLift: $x_1 = 1.5, x_2 = -0.5$
- LRP: $x_1 = 1.5, x_2 = -0.5$

Network $g(x_1, x_2)$
Attributions at $x_1 = 3, x_2 = 1$
- Integrated gradients: $x_1 = 1.5, x_2 = -0.5$
- DeepLift: $x_1 = 2, x_2 = -1$
- LRP: $x_1 = 2, x_2 = -1$

Attribution for Deep Network (Integrated gradient-based)


Auto-encoder / Prototype

Overview of explanation in different AI fields (2)

### Machine Learning (only Artificial Neural Network)

**Network** $f(x_1, x_2)$
- Attributes at $x_1 = 3, x_2 = 1$
  - Integrated gradients: $x_1 = 1.5, x_2 = -0.5$
  - DeepLift: $x_1 = 1.5, x_2 = -0.5$
  - LRP: $x_1 = 1.5, x_2 = -0.5$

**Network** $g(x_1, x_2)$
- Attributes at $x_1 = 3, x_2 = 1$
  - Integrated gradients: $x_1 = 1.5, x_2 = -0.5$
  - DeepLift: $x_1 = 2, x_2 = -1$
  - LRP: $x_1 = 2, x_2 = -1$

**Attribution for Deep Network (Integrated gradient-based)**


**Auto-encoder / Prototype**


**Surrogate Model**

Overview of explanation in different AI fields (2)

**Machine Learning (only Artificial Neural Network)**

Network $f(x_1, x_2)$

- Attributions at $x_1 = 3, x_2 = 1$
  - **Integrated gradients** $x_1 = 1.5, x_2 = -0.5$
  - DeepLift $x_1 = 1.5, x_2 = -0.5$
  - LRP $x_1 = 1.5, x_2 = -0.5$

Network $g(y_1, y_2)$

- Attributions at $x_1 = 1, x_2 = 1$
  - **Integrated gradients** $x_1 = 1.5, x_2 = -0.5$
  - DeepLift $x_1 = 2, x_2 = 1$
  - LRP $x_1 = 2, x_2 = 1$

**Attribution for Deep Network (Integrated gradient-based)**


**Auto-encoder / Prototype**


**Attention Mechanism**


**Surogate Model**

Overview of explanation in different AI fields (3)

Computer Vision

Interpretable Units

Overview of explanation in different AI fields (3)

Computer Vision

Interpretable Units


Uncertainty Map

Alex Kendall, Yarin Gal: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? NIPS 2017: 5580-5590
Overview of explanation in different AI fields (3)

**Computer Vision**

Interpretable Units

Uncertainty Map
Alex Kendall, Yarin Gal: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? NIPS 2017: 5580-5590

Visual Explanation
Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell: Generating Visual Explanations. ECCV (4) 2016: 3-19
Overview of explanation in different AI fields (3)

Computer Vision

Interpretable Units


Visual Explanation

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell: Generating Visual Explanations. ECCV (4) 2016: 3-19

Uncertainty Map

Alex Kendall, Yarin Gal: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? NIPS 2017: 5580-5590

Saliency Map

Overview of explanation in different AI fields (4)

**Game Theory**

Shapley Additive Explanation

Overview of explanation in different AI fields (4)

Game Theory

Shapley Additive Explanation


L-Shapley and C-Shapley (with graph structure)

Jianbo Chen, Le Song, Martin J. Wainwright, Michael I. Jordan: L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data. ICLR 2019
Overview of explanation in different AI fields (4)

### Game Theory

<table>
<thead>
<tr>
<th>Status = Married-div-spouse</th>
<th>Hours per week = 55</th>
<th>Occupation = Exec-managerial</th>
<th>Relationship = Husband</th>
<th>Education = 12</th>
<th>Age = 20</th>
<th>Capital Gain = 0</th>
<th>Race = Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.369</td>
<td>0.369</td>
<td>-0.363</td>
<td>-0.363</td>
<td>-0.363</td>
<td>-0.3628</td>
<td>0.337</td>
<td>0.337</td>
</tr>
</tbody>
</table>

Shapley Additive Explanation


L-Shapley and C-Shapley (with graph structure)

Jianbo Chen, Le Song, Martin J. Wainwright, Michael I. Jordan: L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data. ICLR 2019

~ instancewise feature importance (causal influence)
Overview of explanation in different AI fields (5)

Search and Constraint Satisfaction

Conflicts resolution

Barry O'Sullivan, Alexandre Papadopoulos, Boi Faltings, Pearl Pu: Representative Explanations for Over-Constrained Problems. AAAI 2007: 323-328

Robustness Computation

Overview of explanation in different AI fields (5)

Search and Constraint Satisfaction

Conflicts resolution

Barry O’Sullivan, Alexandre Papadopoulos, Boi Faltings, Pearl Pu: Representative Explanations for Over-Constrained Problems. AAAI 2007: 323-328

Robustness Computation


Constraints relaxation

Overview of explanation in different AI fields (6)

Knowledge Representation and Reasoning

Explaining Reasoning (through Justification) e.g., Subsumption

Deborah L. McGuinness, Alexander Borgida: Explaining Subsumption in Description Logics. IJCAI (1) 1995: 816-821
Overview of explanation in different AI fields (6)

Knowledge Representation and Reasoning

1. (at-least 3 grape) \implies (at-least 2 grape)\hspace{1cm} \text{AtLst}
2. (and (at-least 3 grape) \langle \text{prim GOOD WINE} \rangle) \implies (at-least 2 grape)\hspace{1cm} \text{AndL1}
3. (\text{prim GOOD WINE}) \implies (\text{prim WINE})\hspace{1cm} \text{Prim}
4. (and (at-least 3 grape) (\text{prim GOOD WINE})) \implies (\text{prim WINE})\hspace{1cm} \text{AndL1.3}
5. A \equiv (\text{and} (\text{at-least 3 grape}) (\text{prim GOOD WINE}))\hspace{1cm} \text{Told}
6. A \implies (\text{prim WINE})\hspace{1cm} \text{Eq.4.5}
7. (\text{prim WINE}) \equiv (\text{and} (\text{prim WINE}))\hspace{1cm} \text{AndE1}
8. A \implies (\text{and} (\text{prim WINE}))\hspace{1cm} \text{Eq.7.6}
9. A \equiv (\text{at-least 2 grape})\hspace{1cm} \text{Eq.5.8}
10. A \equiv (\text{and} (at-least 2 grape) (\text{prim WINE})) \hspace{1cm} \text{AndR9.8}

Abduction Reasoning (in Bayesian Network)


Explaining Reasoning (through Justification) e.g., Subsumption

Deborah L. McGuinness, Alexander Borgida: Explaining Subsumption in Description Logics. IJCAI (1) 1995: 816-821
Overview of explanation in different AI fields (6)

Knowledge Representation and Reasoning

Explaining Reasoning (through Justification) e.g.,
Subsumption

Deborah L. McGuinness, Alexander Borgida: Explaining Subsumption in Description Logics. IJCAI (1) 1995: 816-821

Abduction Reasoning (in Bayesian Network)


Diagnosis Inference

Overview of explanation in different AI fields (7)

Multi-agent Systems

<table>
<thead>
<tr>
<th>MAS INFRASTRUCTURE</th>
<th>INDIVIDUAL AGENT INFRASTRUCTURE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MAS INTEROPERATION</strong></td>
<td><strong>INTEROPERATION</strong></td>
</tr>
<tr>
<td>Translation Services</td>
<td>Interopration Modules</td>
</tr>
<tr>
<td>Interoperation Services</td>
<td></td>
</tr>
<tr>
<td><strong>CAPABILITY TO AGENT MAPPING</strong></td>
<td><strong>CAPABILITY TO AGENT MAPPING</strong></td>
</tr>
<tr>
<td>Middle Agents</td>
<td>Middle Agents Components</td>
</tr>
<tr>
<td><strong>NAME TO LOCATION MAPPING</strong></td>
<td><strong>NAME TO LOCATION MAPPING</strong></td>
</tr>
<tr>
<td>AMS</td>
<td>AMS Component</td>
</tr>
<tr>
<td><strong>SECURITY</strong></td>
<td><strong>SECURITY</strong></td>
</tr>
<tr>
<td>Certificate Authority</td>
<td>Security Module</td>
</tr>
<tr>
<td>Cryptographic Services</td>
<td>private/public Keys</td>
</tr>
<tr>
<td><strong>PERFORMANCE SERVICES</strong></td>
<td><strong>PERFORMANCE SERVICES</strong></td>
</tr>
<tr>
<td>MAS Monitoring</td>
<td>Performance Services Modules</td>
</tr>
<tr>
<td>Reputation Services</td>
<td></td>
</tr>
<tr>
<td><strong>MULTIAGENT MANAGEMENT SERVICES</strong></td>
<td><strong>MANAGEMENT SERVICES</strong></td>
</tr>
<tr>
<td>Logging, Activity Visualization, Launching</td>
<td>Logging and Visualization Components</td>
</tr>
<tr>
<td><strong>ACL INFRASTRUCTURE</strong></td>
<td><strong>ACL INFRASTRUCTURE</strong></td>
</tr>
<tr>
<td>Public Ontology</td>
<td>ACL Parser</td>
</tr>
<tr>
<td>Protocols Servers</td>
<td>Private Ontology</td>
</tr>
<tr>
<td><strong>COMMUNICATION INFRASTRUCTURE</strong></td>
<td><strong>COMMUNICATION MODULES</strong></td>
</tr>
<tr>
<td>Discovery</td>
<td>Discovery Component</td>
</tr>
<tr>
<td>Message Transfer</td>
<td>Message Transfer Module</td>
</tr>
<tr>
<td><strong>OPERATING ENVIRONMENT</strong></td>
<td></td>
</tr>
<tr>
<td>Machines, OS, Network</td>
<td>Multicast</td>
</tr>
<tr>
<td>Transport Layer: TCP/IP, Wireless, Infrared, SSL</td>
<td></td>
</tr>
</tbody>
</table>

Explanation of Agent Conflicts & Harmful Interactions

Overview of explanation in different AI fields (7)

Multi-agent Systems

<table>
<thead>
<tr>
<th>MAS INFRASTRUCTURE</th>
<th>INDIVIDUAL AGENT INFRASTRUCTURE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MAS INTEROPERATION</strong></td>
<td><strong>INTEROPERATION</strong></td>
</tr>
<tr>
<td>Translation Services</td>
<td>Interoperatio Modules</td>
</tr>
<tr>
<td>Interoperatio Services</td>
<td></td>
</tr>
<tr>
<td><strong>CAPABILITY TO AGENT MAPPING</strong></td>
<td><strong>CAPABILITY TO AGENT MAPPING</strong></td>
</tr>
<tr>
<td>Middle Agents</td>
<td>Middle Agents Components</td>
</tr>
<tr>
<td><strong>NAME TO LOCATION MAPPING</strong></td>
<td><strong>NAME TO LOCATION MAPPING</strong></td>
</tr>
<tr>
<td>ANS</td>
<td>ANS Component</td>
</tr>
<tr>
<td><strong>SECURITY</strong></td>
<td><strong>SECURITY</strong></td>
</tr>
<tr>
<td>Certificate Authority</td>
<td>Security Module</td>
</tr>
<tr>
<td>Cryptographic Services</td>
<td>Private/public Keys</td>
</tr>
<tr>
<td><strong>PERFORMANCE SERVICES</strong></td>
<td><strong>PERFORMANCE SERVICES</strong></td>
</tr>
<tr>
<td>MAS Monitoring</td>
<td>Performance Services Modules</td>
</tr>
<tr>
<td>Reputation Services</td>
<td></td>
</tr>
<tr>
<td><strong>MULTIAGENT MANAGEMENT SERVICES</strong></td>
<td><strong>MANAGEMENT SERVICES</strong></td>
</tr>
<tr>
<td>Logging</td>
<td>Logging and Visualization Components</td>
</tr>
<tr>
<td>Activity Visualization</td>
<td></td>
</tr>
<tr>
<td>Launching</td>
<td></td>
</tr>
<tr>
<td><strong>ACL INFRASTRUCTURE</strong></td>
<td><strong>ACL INFRASTRUCTURE</strong></td>
</tr>
<tr>
<td>Public Ontology</td>
<td>ACL Parser</td>
</tr>
<tr>
<td>Protocols Servers</td>
<td>Private Ontology</td>
</tr>
<tr>
<td><strong>COMMUNICATION INFRASTRUCTURE</strong></td>
<td><strong>COMMUNICATION MODULES</strong></td>
</tr>
<tr>
<td>Discovery</td>
<td>Discovery Component</td>
</tr>
<tr>
<td>Message Transfer</td>
<td>Message Transfer Module</td>
</tr>
<tr>
<td><strong>OPERATING ENVIRONMENT</strong></td>
<td></td>
</tr>
<tr>
<td>Machines, OS, Network</td>
<td>Multicast</td>
</tr>
<tr>
<td>Transport Layer: TCP/IP, Wireless,</td>
<td></td>
</tr>
<tr>
<td>Infrared, SSL</td>
<td></td>
</tr>
</tbody>
</table>

Explanation of Agent Conflicts & Harmful Interactions

Overview of explanation in different AI fields (7)

Multi-agent Systems

<table>
<thead>
<tr>
<th>Multi-agent Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MAS INFRASTRUCTURE</strong></td>
</tr>
<tr>
<td><strong>MAS INTEROPERATION</strong></td>
</tr>
<tr>
<td>Translation Services</td>
</tr>
<tr>
<td><strong>CAPABILITY TO AGENT MAPPING</strong></td>
</tr>
<tr>
<td>Middle Agents</td>
</tr>
<tr>
<td><strong>NAME TO LOCATION MAPPING</strong></td>
</tr>
<tr>
<td>AWS</td>
</tr>
<tr>
<td><strong>SECURITY</strong></td>
</tr>
<tr>
<td>Certificate Authority</td>
</tr>
<tr>
<td><strong>PERFORMANCE SERVICES</strong></td>
</tr>
<tr>
<td>MAS Monitoring</td>
</tr>
<tr>
<td><strong>MULTIAGENT MANAGEMENT SERVICES</strong></td>
</tr>
<tr>
<td>Logging, Activity Visualization, Launching</td>
</tr>
<tr>
<td><strong>ACL INFRASTRUCTURE</strong></td>
</tr>
<tr>
<td>Public Ontology</td>
</tr>
<tr>
<td><strong>COMMUNICATION INFRASTRUCTURE</strong></td>
</tr>
<tr>
<td>Discovery</td>
</tr>
<tr>
<td><strong>OPERATING ENVIRONMENT</strong></td>
</tr>
<tr>
<td>Machines, OS, Network</td>
</tr>
</tbody>
</table>

Explanation of Agent Conflicts & Harmful Interactions


Agent Strategy Summarization


Explainable Agents


W. Lewis Johnson: Agents that Learn to Explain Themselves. AAAI 1994: 1257-1263
Overview of explanation in different AI fields (8)

NLP

Fine-grained explanations are in the form of:
- texts in a real-world dataset;
- Numerical scores

Explainable NLP

Fine-grained explanations are in the form of:

- texts in a real-world dataset;
- Numerical scores

### Explainable NLP

Overview of explanation in different AI fields (8)

**NLP**

Fine-grained explanations are in the form of:
- texts in a real-world dataset;
- Numerical scores

**LIME for NLP**


---


---

NLP Debugger

Overview of explanation in different AI fields (9)

### Planning and Scheduling

<table>
<thead>
<tr>
<th>Explanation Type</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plan Patch Explanation / VAL</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Model Patch Explanation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Minimally Complete Explanation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>?</td>
</tr>
<tr>
<td>Minimally Monotonic Explanation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>?</td>
</tr>
<tr>
<td>(Approximate) Minimally Complete Explanation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Overview of explanation in different AI fields (9)

### Planning and Scheduling

<table>
<thead>
<tr>
<th>Explanation Type</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plan Patch Explanation / VAL</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Model Patch Explanation</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Minimally Complete Explanation</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>?</td>
</tr>
<tr>
<td>Minimally Monotonic Explanation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>?</td>
</tr>
<tr>
<td>(Approximate) Minimally Complete Explanation</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>


Human-in-the-loop Planning


(Manual) Plan Comparison

Overview of explanation in different AI fields (10)

### Robotics

#### Narration of Autonomous Robot Experience


Overview of explanation in different AI fields (10)

Robotics

Narration of Autonomous Robot Experience


Robot: I have decided to turn left.
Human: Why did you do that?
Robot: I believe that the correct action is to turn left BECAUSE:
I’m being asked to go forward
AND This area in front of me was 20 cm higher than me
*highlights area*
AND the area to the left has maximum protrusions of less than 5 cm *highlights area*
AND I’m tilted to the right by more than 5 degrees.
Here is a display of the path through the tree that lead to this decision. *displays tree*

Human: How confident are you in this decision?
Robot: The distribution of actions that reached this leaf node is shown in this histogram. *displays histogram*
This action is predicted to be correct 67% of the time.

Human: Where did the threshold for the area in front come from?
Robot: Here is the histogram of all training examples that reached this leaf. 80% of examples where this area was above 20 cm predicted the appropriate action to be “drive forward”.

From Decision Tree to human-friendly information

Raymond Ka-Man Sheh: "Why Did You Do That?" Explainable Intelligent Robots. AAAI Workshops 2017
Overview of explanation in different AI fields (11)

Reasoning under uncertainty

Probabilistic Graphical Models

Evaluation
XAI: One Objective, Many Metrics

- **Comprehensibility**: How much effort for correct human interpretation?
- **Succinctness**: How concise and compact is the explanation?
- **Actionability**: What can one action, do with the explanation?
- **Reusability**: Could the explanation be personalized?
- **Accuracy**: How accurate and precise is the explanation?
- **Completeness**: Is the explanation complete, partial, restricted?

Source: Accenture Point of View. Understanding Machines: Explainable AI. Freddy Lecue, Dadong Wan
On the role of Knowledge Graphs in Explainable Machine Learning
Knowledge Graph Embeddings in Machine Learning

https://stats.stackexchange.com/questions/230581/decision-tree-too-large-to-interpret
Knowledge Graph for Decision Trees

https://stats.stackexchange.com/questions/230581/decision-tree-too-large-to-interpret
Knowledge Graph for Deep Neural Network (1)

- **Input Layer**
- **Training Data**
- **Hidden Layer**
  - Neurons respond to simple shapes
  - Neurons respond to more complex structures
  - Neurons respond to highly complex, abstract concepts
- **Output Layer**

Image descriptions:
- 1st Layer: Low-level features to high-level features
- 2nd Layer: Low-level features to high-level features
- n-th Layer: Low-level features to high-level features

Legend:
- **Input (unlabeled image)**
- **10% WOLF**
- **90% DOG**

For more details, please refer to the full document.
Neurons respond to simple shapes
Neurons respond to more complex structures
Neurons respond to highly complex, abstract concepts

What is the causal relationship between the input / hidden / output layers

Knowledge Graph for Deep Neural Network (2)
Knowledge Graph for Personalized XAI

Description 1: This is an orange train accident

Description 2: This is an train accident between two speed merchant trains of characteristics X43-B and Y33-C in a dry environment

Description 3: This is a public transportation accident
"How to explain transfer learning with appropriate knowledge representation?"


Knowledge-Based Transfer Learning Explanation

Jiaoyan Chen  
Department of Computer Science  
University of Oxford, UK

Freddy Lecue  
INRIA, France  
Accenture Labs, Ireland

Jeff Z. Pan  
Department of Computer Science  
University of Aberdeen, UK

Ian Horrocks  
Department of Computer Science  
University of Oxford, UK

Huajun Chen  
College of Computer Science, Zhejiang University, China  
Alibaba-Zhejian University Frontier Technology Research Center
Applications
Challenge: Object detection is usually performed from a large portfolio of Artificial Neural Networks (ANNs) architectures trained on large amount of labelled data. Explaining object detections is rather difficult due to the high complexity of the most accurate ANNs.

AI Technology: Integration of AI related technologies i.e., Machine Learning (Deep Learning / CNNs), and knowledge graphs / linked open data.

XAI Technology: Knowledge graphs and Artificial Neural Networks

Fig. 2. Left image: results from baseline Faster RCNN: Paddle: 50% confidence, Person: 66%, Man: 46%. Right image: results from the semantic augmentation: Paddle: 74% confidence, Person: 66%, Man: 56%, Boat: 58% with explanation: Person, Paddle, Water as part of the context in the image and knowledge graph of concept Boat. (color print).
Debugging Artificial Neural Networks – Industry Agnostic

**Challenge:** Designing Artificial Neural Network architectures requires lots of experimentation (i.e., training phases) and parameters tuning (optimization strategy, learning rate, number of layers...) to reach optimal and robust machine learning models.

**AI Technology:** Artificial Neural Network

**XAI Technology:** Artificial Neural Network, 3D Modeling and Simulation Platform For AI
**Challenge:** What is the robustness of Visual Question Answering models? What is the impact of semantics?

**AI Technology:** Artificial Neural Networks.

**XAI Technology:** Integrated Gradients

---

**Tabular QA**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Nation</th>
<th>Gold</th>
<th>Silver</th>
<th>Bronze</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>India</td>
<td>102</td>
<td>58</td>
<td>37</td>
<td>197</td>
</tr>
<tr>
<td>2</td>
<td>Nepal</td>
<td>32</td>
<td>10</td>
<td>24</td>
<td>66</td>
</tr>
<tr>
<td>3</td>
<td>Sri Lanka</td>
<td>16</td>
<td>42</td>
<td>62</td>
<td>120</td>
</tr>
<tr>
<td>4</td>
<td>Pakistan</td>
<td>10</td>
<td>36</td>
<td>30</td>
<td>76</td>
</tr>
<tr>
<td>5</td>
<td>Bangladesh</td>
<td>10</td>
<td>35</td>
<td>13</td>
<td>68</td>
</tr>
<tr>
<td>6</td>
<td>Bhutan</td>
<td>1</td>
<td>6</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>Maldives</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Q: How many medals did India win? A: 197

Neural Programmer (2017) model 33.5% accuracy on WikiTableQuestions

Kazemi and Elqursh (2017) model. 61.1% on VQA 1.0 dataset (state of the art = 66.7%)

---

**Visual QA**

Q: How symmetrical are the white bricks on either side of the building? A: very

Yu et al (2018) model. 84.6 F-1 score on SQuAD (state of the art)

---

**Reading Comprehension**

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager.

Q: Name of the quarterback who was 38 in Super Bowl XXXIII? A: John Elway

---

**Explaining Visual Question Answering – Industry Agnostic**

What is the man doing? → What is the **tweet** doing?
How many children are there? → How many **tweet** are there?

VQA model’s response remains the same 75.6% of the time on questions that it originally answered correctly.

---

Source: Explainable AI in Industry. KDD 2019 Tutorial. Ankur Taly, Mukund Sundararajan, Kedar Dhamdhere, Pramod Mudrakarta
Relevance Debugging and Explaining – Industry Agnostic

Challenge: A Machine Learning system can fail in many different points e.g., data features selection, construction, inconsistencies. How to debug bad performance in machine learning models and prediction?

AI Technology: Artificial Neural Networks.

XAI Technology: Model / Prediction comparison

Source: Explainable AI in Industry. KDD 2019 Tutorial. Daniel Qiu, Yucheng Qian
Challenge: Public transportation is getting more and more self-driving vehicles. Even if trains are getting more and more autonomous, the human stays in the loop for critical decision, for instance in case of obstacles. In case of obstacles trains are required to provide recommendation of action i.e., go on or go back to station. In such a case the human is required to validate the recommendation through an explanation exposed by the train or machine.

AI Technology: Integration of AI related technologies i.e., Machine Learning (Deep Learning / CNNs), and semantic segmentation.

XAI Technology: Deep learning and Epistemic uncertainty
Explaining Flight Performance - Transportation

Challenge: Predicting and explaining aircraft engine performance

AI Technology: Artificial Neural Networks

XAI Technology: Shapely Values
Challenge: Globally 323,454 flights are delayed every year. Airline-caused delays totaled 20.2 million minutes last year, generating huge cost for the company. Existing in-house technique reaches 53% accuracy for predicting flight delay, does not provide any time estimation (in minutes as opposed to True/False) and is unable to capture the underlying reasons (explanation).

AI Technology: Integration of AI related technologies i.e., Machine Learning (Deep Learning / Recurrent neural Network), Reasoning (through semantics-augmented case-based reasoning) and Natural Language Processing for building a robust model which can (1) predict flight delays in minutes, (2) explain delays by comparing with historical cases.

XAI Technology: Knowledge graph embedded Sequence Learning using LSTMs
Model Explanation for Sales Prediction - Sales

Challenge: How to predict and explain upsell / churn for a company?

AI Technology: Artificial Neural Networks.

XAI Technology: Features importance (contribution, influence), LIME.

Company: CompanyX
Upsell LCP (LinkedIn Career Page)

Top Feature Contributor
- f1: 430.5
- f2: 216
- f3: 10097.57
- f4: 15

Top Feature Influencer (Positive)
- f5: 0 → 5.4, 0.03
- f6: 168 → 0, 0.03
- f7: 0 → 0.24, 0.02

Top Feature Influencer (Negative)
- f1: 430.5 → 148.7, 0.20
- f2: 216 → 0, 0.17
- f8: 423 → 146.0, 0.07

Source: Explainable AI in Industry. KDD 2019 Tutorial. Jilei Yang, Wei Di, Songtao Guo
**Challenge:** Accenture is managing every year more than 80,000 opportunities and 35,000 contracts with an expected revenue of $34.1 billion. Revenue expectation does not meet estimation due to the complexity and risks of critical contracts. This is, in part, due to the (1) large volume of projects to assess and control, and (2) the existing non-systematic assessment process.

**AI Technology:** Integration of AI technologies i.e., Machine Learning, Reasoning, Natural Language Processing for building a robust model which can (1) predict revenue loss, (2) recommend corrective actions, and (3) explain why such actions might have a positive impact.

**XAI Technology:** Knowledge graph embedded Random Forrest
**Challenge:** Predicting and explaining abnormally employee expenses (as high accommodation price in 1000+ cities).

**AI Technology:** Various techniques have been matured over the last two decades to achieve excellent results. However, most methods address the problem from a statistic and pure data-centric angle, which in turn limit any interpretation. We elaborated a web application running live with real data from (i) travel and expenses from Accenture, (ii) external data from third party such as Google Knowledge Graph, DBPedia (relational Database version of Wikipedia) and social events from Eventful, for explaining abnormalities.

**XAI Technology:** Knowledge graph embedded Ensemble Learning

Freddy Lécué, Jiewen Wu: Explaining and predicting abnormal expenses at large scale using knowledge graph based reasoning. J. Web Sem. 44: 89-103 (2017)
Local, post-hoc, contrastive explanations of black-box classifiers

Required minimum change in input vector to flip the decision of the classifier.

Interactive Contrastive Explanations

Challenge: We predict loan applications with off-the-shelf, interchangeable black-box estimators, and we explain their predictions with counterfactual explanations. In counterfactual explanations the model itself remains a black box; it is only through changing inputs and outputs that an explanation is obtained.

AI Technology: Supervised learning, binary classification.

XAI Technology: Post-hoc explanation, Local explanation, Counterfactuals, Interactive explanations

Counterfactual Explanations for Credit Decisions (1) - Finance

Counterfactual Explanations for Credit Decisions (2) - Finance

**Sorry, your loan application has been rejected.**

**Our analysis:**

The following features were too high:
- PercentInstallTrad...
- NetFractionRevol...
- NetFractionInstall...
- NumRevolvingTra...
- NumBank2NetTra...
- PercentTradesWB...

The following features were too low:
- MSinceOldestTrad...
- AverageMlnFile...
- NumTotalTrades

The following features require changes:
- MaxDelq2PublicR...
- MaxDelqEver

**Counterfactuals suggest where to increase (green, dashed) or decrease (red, striped) each feature.**

Counterfactual Explanations for Credit Decisions (3) - Finance

RECOMMENDED CHANGES

- M Since Oldest Trade Open
- Average M In File
- Num Satisfactory Trades

Select categorical constraints.

Max Delq 2 Public Rec Last 12M
Current: unknown delinquency
10 selected

Max Delq Ever
Current: 60 days delinquent

Explaining Talent Search Results – Human Resources

**Challenge:** How to rationalize a talent search for a recruiter when looking for candidates for a given role. Features are dynamic and costly to compute. Recruiters are interested in discriminating between two candidates to make a selection.

**AI Technology:** Generalized Linear Mixed Models, Artificial Neural Networks, XGBoost

**XAI Technology:** Generalized Linear Mixed Models (inherently explainable), Integrated Gradient, Features Importance in XGBoost

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Difference (1 vs 2)</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature</td>
<td>Description</td>
<td>-2.0476928</td>
<td>-2.144455602</td>
</tr>
<tr>
<td>Feature</td>
<td>Description</td>
<td>-2.3223877</td>
<td>1.903594818</td>
</tr>
<tr>
<td>Feature</td>
<td>Description</td>
<td>0.1166667</td>
<td>0.2114946752</td>
</tr>
<tr>
<td>Feature</td>
<td>Description</td>
<td>-2.1442587</td>
<td>0.2060414469</td>
</tr>
<tr>
<td>Feature</td>
<td>Description</td>
<td>-14</td>
<td>0.1215354111</td>
</tr>
<tr>
<td>Feature</td>
<td>Description</td>
<td>1</td>
<td>0.1000282466</td>
</tr>
<tr>
<td>Feature</td>
<td>Description</td>
<td>-92</td>
<td>-0.085286277</td>
</tr>
<tr>
<td>Feature</td>
<td>Description</td>
<td>0.9333333</td>
<td>0.0568533262</td>
</tr>
<tr>
<td>Feature</td>
<td>Description</td>
<td>-1</td>
<td>-0.051796317</td>
</tr>
<tr>
<td>Feature</td>
<td>Description</td>
<td>-1</td>
<td>-0.050895940</td>
</tr>
</tbody>
</table>

Explanation of Medical Condition Relapse – Health

Challenge: Explaining medical condition relapse in the context of oncology.

AI Technology: Relational learning

XAI Technology: Knowledge graphs and Artificial Neural Networks
Results

These results are for women who have already had surgery. This table shows the percentage of women who survive at least 5, 10, and 15 years after surgery, based on the information you have provided.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Additional Benefit</th>
<th>Overall Survival %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgery only</td>
<td>-</td>
<td>72%</td>
</tr>
<tr>
<td>+ Hormone therapy</td>
<td>0%</td>
<td>72%</td>
</tr>
</tbody>
</table>

If death from breast cancer were excluded, 82% would survive at least 10 years.

Challenge: Predict is an online tool that helps patients and clinicians see how different treatments for early invasive breast cancer might improve survival rates after surgery.

AI Technology: competing risk analysis

XAI Technology: Interactive explanations, Multiple representations.

David Spiegelhalter, Making Algorithms trustworthy, NeurIPS 2018 Keynote predict.nhs.uk/tool
More on XAI
(Some) Tutorials, Workshops, Challenge

**Tutorial:**
- ICCV 2019 Tutorial on Interpretable Machine Learning for Computer Vision (#2) - [https://interpretablevision.github.io/](https://interpretablevision.github.io/)

**Workshop:**
- IJCAI 2019 Workshop on Optimisation and Explanation in AI (#1) - [https://www.doc.ic.ac.uk/~kc2813/OXAI/](https://www.doc.ic.ac.uk/~kc2813/OXAI/)
- SIGIR 2019 Workshop on Explainable Recommendation and Search (#2) - [https://ears2019.github.io/](https://ears2019.github.io/)
- [https://openreview.net/group?id=icaps-conference.org/ICAPS/2019/Workshop/XAIP](https://openreview.net/group?id=icaps-conference.org/ICAPS/2019/Workshop/XAIP)
- CD-MAKE 2019 – Workshop on Explainable AI (#2) - [https://cd-make.net/special-sessions/make-explainable-ai/](https://cd-make.net/special-sessions/make-explainable-ai/)

**Challenge:**
(Some) Software Resources

- iNNvestigate: A toolbox to iNNvestigate neural networks’ predictions. [github.com/albermax/iNNvestigate](https://github.com/albermax/iNNvestigate)
- SHAP: SHapley Additive exPlanations. [github.com/slundberg/shap](https://github.com/slundberg/shap)
- Microsoft Explainable Boosting Machines. [https://github.com/Microsoft/interpret](https://github.com/Microsoft/interpret)
- GANDissect: Pytorch-based tools for visualizing and understanding the neurons of a GAN. [https://github.com/CSAILVision/GANDissect](https://github.com/CSAILVision/GANDissect)
- ELi5: A library for debugging/inspecting machine learning classifiers and explaining their predictions. [github.com/TeamHG-Memex/eli5](https://github.com/TeamHG-Memex/eli5)
- Skater: Python Library for Model Interpretation/Explanations. [github.com/datascienceinc/Skater](https://github.com/datascienceinc/Skater)
- Yellowbrick: Visual analysis and diagnostic tools to facilitate machine learning model selection. [github.com/DistrictDataLabs/yellowbrick](https://github.com/DistrictDataLabs/yellowbrick)
- Lucid: A collection of infrastructure and tools for research in neural network interpretability. [github.com/tensorflow/lucid](https://github.com/tensorflow/lucid)
- Sklearn_explain: model individual score explanation for an already trained scikit-learn model. [https://github.com/antoinecarme/sklearn_explain](https://github.com/antoinecarme/sklearn_explain)
- Heatmapping: Prediction decomposition in terms of contributions of individual input variables
- Deep Learning Investigator: Investigation of Saliency, Deconvnet, GuidedBackprop and more. [https://github.com/albermax/iNNvestigate](https://github.com/albermax/iNNvestigate)
- Google PAIR What-if: Model comparison, counterfactual, individual similarity. [https://pair-code.github.io/what-if-tool/](https://pair-code.github.io/what-if-tool/)
- IBM AI Fairness: Set of fairness metrics for datasets and ML models, explanations for these metrics. [https://github.com/IBM/aif360](https://github.com/IBM/aif360)
- Blackbox auditing: Auditing Black-box Models for Indirect Influence. [https://github.com/algofairness/BlackBoxAuditing](https://github.com/algofairness/BlackBoxAuditing)
- AXA Interpretability and Robustness: [https://axa-rev-research.github.io/](https://axa-rev-research.github.io/) (more on research resources – not much about tools)
(Some) Initiatives: XAI in USA

**Challenge Problem Areas**

- Data Analytics
  - Multimedia Data
- Autonomy
  - ArduPilot & SITL Simulation

**TA 1: Explainable Learners**

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

**TA 2: Psychological Model of Explanation**

- Psych. Theory of Explanation
- Computational Model
- Consulting

**Evaluator**

**Deep Learning Teams**

**Interpretable Model Teams**

**Model Induction Teams**

**TA1: Explainable Learners**

- Explainable learning systems that include both an explainable model and an explanation interface

**TA2: Psychological Model of Explanation**

- Psychological theories of explanation and develop a computational model of explanation from those theories

**Evaluation Framework**

- Learning Performance
  - Explanation Effectiveness
  - User Satisfaction
  - Mental Model
  - Task Performance
  - Trust Assessment
  - Correctability
(Some) Initiatives: XAI in Canada

**DEEL (Dependable Explainable Learning) Project 2019-2024**

- **Research institutions**
  - CRIAQ
  - IVADO
  - CRGNS NSERC

- **Industrial partners**
  - Bell Helicopter
  - BOMBARDIER
  - CAF
  - THALES

- **Academic partners**
  - Science and technology to develop new methods towards Trustable and Explainable AI

- **System Robustness**
  - To biased data
  - Of algorithm
  - To change
  - To attacks

- **Certificability**
  - Structural warranties
  - Risk auto evaluation
  - External audit

- **Explicability & Interpretability**

- **Privacy by design**
  - Differential privacy
  - Homomorphic coding
  - Collaborative learning
  - To attacks
(Some) Initiatives: XAI in EU
Conclusion

- Explainable AI is motivated by real-world applications in AI
- Not a new problem – a reformulation of past research challenges in AI
- Multi-disciplinary: multiple AI fields, HCI, social sciences (multiple definitions)
- In AI (in general): many interesting / complementary approaches
- Many industrial applications already – crucial for AI adoption in critical systems
Future Challenges

- Creating awareness! Success stories!
- Foster multi-disciplinary collaborations in XAI research.
- Help shaping industry standards, legislation.
- More work on transparent design.
- Investigate symbolic and sub-symbolic reasoning.

**Evaluation:**

- *We need benchmark* - Shall we start a task force?
- *We need an XAI challenge* - Anyone interested?
- *Rigorous, agreed upon, human-based evaluation protocols*
Job Openings

Research and Technology Applied AI (Artificial Intelligence) Scientist

Wherever safety and security are critical, Thales can build smarter solutions. Everywhere.

As a Research and Technology leader for the Defence Technology, the combined expertise of Thales Ireland, Thales UK and Thales Canada is a key player in keeping the peace and protecting the national security interests of a number of countries in the region.

Established in 1972, Thales Canada has over 1,800 employees in locations in Toronto, Ottawa, Montreal, and Vancouver working in Defence, Avionics and Transportation sectors.

This is a unique opportunity to play a key role on the Defence Technology (RTT) in Canada (Quebec and Montreal) as part of a team of 950 experts at five locations worldwide. We are actively researching and developing cutting-edge AI technologies. Our passion is to imagine technologies that will not only make you more efficient but also more effective in your work.

Job Description
An AI (Artificial Intelligence) Research and Technology Applied Scientist will be part of a cross-functional team that is developing innovative prototypes to demonstrate intelligence. To be successful in this role, one must have a good understanding of machine learning and a strong ability to learn new technologies and apply them to real-world problems.

Job Requirements
- Strong knowledge of Machine Learning foundations
- Strong development skills with Machine Learning frameworks e.g., Scikit-learn, Tensorflow, PyTorch, Theano
- Knowledge of mainstream Deep Learning architectures (MLP, CNN, RNN, etc.)
- Strong Python programming skills
- Working knowledge of Linux OS
- Eagerness to contribute in a team-oriented environment
- Demonstrated leadership abilities in school, civil or business organisations
- Ability to work creatively and analytically in a problem-solving environment
- Proven verbal and written communication skills in English (talks, presentations, publications, etc.)

Basic Qualifications
- Master’s degree in computer science, engineering or mathematics fields
- Prior experience in artificial intelligence, machine learning, natural language processing, or advanced analytics

Preferred Qualifications
- Minimum 3 years of experience working with Python and machine learning
- Demonstrated abilities in designing large-scale AI systems
- Demonstrated interest in Explainable AI and/or learning to learn
- Work experience with programming languages such as C, C++, Java, scripting languages (Perl/Python/Ruby) or similar
- Hands-on experience with data visualization, analytics tools/languages
- Demonstrated teamwork and collaboration in professional settings
- Ability to establish credibility with clients and other team members

@freddylecue
https://tinyurl.com/freddylecue
Freddy.lecue.e@thalesdigital.io