XAI – The Story So Far

May 2nd, 2019

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

@freddylecue
https://tinyurl.com/freddylecue

www.thalesgroup.com
Context
Gary Chavez added a photo you might be in.
about a minute ago · 📸
Markets we serve

Aerospace  Space  Ground Transportation  Defence  Security

Trusted Partner For A Safer World
Trustable AI
AI Adoption: Requirements

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

What is the rational?
Motivations
Motivation (1)

Criminal Justice

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

---

nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

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

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.

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

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

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
Motivation (4)

Critical Systems

https://www.sncf.com/sncv1/ressources/presskit__train_auto__s__v2.pdf
The need for explainable AI rises with the potential cost of poor decisions

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
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** | ˌɛkspləˈneɪʃ(ə)n |

noun

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

Models, Outputs of the Intelligent System

**interpret** | ɪnˈtɛprɪt |

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

1 explain the meaning of (information or actions): the evidence is difficult to interpret.

Models, Outputs of the Intelligent System
XAI in AI
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?

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

Artificial Intelligence

Machine Learning

Dependency Plot

Feature Importance

Surrogate Model

Game Theory

Search

Planning

Robotics

Computer Vision

KRR

UAI

MAS

NLP

NLP

UAI

KRR

Game Theory

Search

Planning

Robotics

Dependency Plot

Feature Importance

Surrogate Model

Machine Learning

Computer Vision

NLP

UAI

KRR

MAS

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?

Saliency Map

Uncertainty Map

Dependency Plot

Feature Importance

Surrogate Model

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

Artificial Intelligence

MAS

Planning

Search

Game Theory

Robotics

Computer Vision

KRR

UAI

NLP

Which features are responsible of classification?
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 complex features are responsible of classification?

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

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 complex features are responsible of classification?

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

Surrogate Model

Plan Refinement

Dependency Plot

Feature Importance

Artificial Intelligence

Robotics

Search

Game Theory

Computer Vision

KRR

UAI

MAS

NLP

Uncertainty Map

Saliency Map

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

Dependency Plot

Feature Importance

Surrogate Model

Artificial Intelligence

Plan Refinement

Search

Robotics

Game Theory

Computer Vision

KRR

UAI

MAS

NLP

Uncertainty Map

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

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

Computer Vision

- Saliency Map
- Uncertainty Map

Machine Learning

- Surrogate Model
- Feature Importance
- Dependency Plot

Game Theory

- Search
- Conflicts Resolution

Artificial Intelligence

Plan Refinement

Robotics

- KRR
- UAI

NLP

- MAS

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?

Which complex features are responsible of classification?

Which agent strategy & plan?

Which player contributes most?

Why such a conversational flow?

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

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

Conflicts Resolution

Game Theory

Shapely Values

Narrative-based

Robotics

Which decisions, combination of multimodal decisions lead to an action?

Search

Plan Refinement

MARS

KRR

UAI

Machine Learning based

What agent strategy & plan?

Why such a conversational flow?

Which complex features are responsible of classification?

Which entity is responsible for classification?

Machine Learning based

AI's

Many Definitions

Many Approaches

Computer Vision

Uncertainty Map

Dependency Plot

Feature Importance

Surrogate Model

Explain reasons/motivation, justify, and explain the causes of their decisions.

How to summarize the reasons/motivation, justify, and explain the causes of their decisions?

Strategy Summarization

Artificial Intelligence

Which player contributes most?

Why such a conversational flow?
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

Artificial Intelligence

- Which complex features are responsible of classification?
- Which actions are responsible of a plan?
- Which constraints can be relaxed?
- Which combination of features is optimal?
- Which decisions, combination of multimodal decisions lead to an action?

Machine Learning

- Which features are responsible of classification?

Computer Vision

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

Search

- Which constraints can be relaxed?

Game Theory

- Which combination of features is optimal?

Conflicts Resolution

Plan Refinement

Planning

KRR

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

UAI

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

Abduction

Uncertainty Map

Surrogate Model

Dependency Plot

Surrogate Model

Strategy Summarization

MAS

Surrogate Model

Which entity is responsible for classification?

Robotics

Narrative-based

Shapely Values

Mathematical

Dependency Plot

Feature Importance

Game Theory

Which actions are responsible of a plan?

Search

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

Game Theory

- Which combination of features is optimal?

Narrative-based

Shapely Values

- Which entity is responsible for classification?
- Which agent strategy & plan?
- Which player contributes most?
- Why such a conversational flow?

Robotics

Narrative-based

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

Which axiom is responsible of inference (e.g., classification)?

Abduction/Diagnostic: Find the right root causes (abduction)?

Which combination of features is optimal?

Which decisions, combination of multimodal decisions lead to an action?

Which constraints can be relaxed?

Uncertainty as an alternative to explanation

Which entity is responsible for classification?

Uncertainty Map

Which combinations are optimal?

Which features are responsible of classification?
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

**Naive Bayes model**


**Counterfactual What-if**


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 (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)**


Overview of explanation in different AI fields (2)

### Machine Learning (only Artificial Neural Network)

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


### 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)


---


---

#### Surogate 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 $g(x_1, x_2)$
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: $x_1 = 2, x_2 = -1$

Attention Mechanism


Attribution for Deep Network (Integrated gradient-based)


Auto-encoder


Surogate Model

Overview of explanation in different AI fields (3)

Computer Vision

Interpretable Units

Overview of explanation in different AI fields (3)

### Computer Vision

<table>
<thead>
<tr>
<th>Interpretable Units</th>
</tr>
</thead>
</table>

### 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


#### 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
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

#### Saliency Map


#### 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 (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**

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


If A+1 then NEW Conflicts on X and Y

Constraints relaxation

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

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)

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>Interoporation Modules</td>
</tr>
<tr>
<td>Interoporation 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, 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>
</tbody>
</table>

**OPERATING ENVIRONMENT**
- Machines, OS, Network
- Multicast
- Transport Layer: TCP/IP, Wireless, Infrared, SSL

Explanation of Agent Conflicts & Harmful Interactions

Overview of explanation in different AI fields (7)

Multi-agent Systems

MAS INFRASTRUCTURE
- MAS INTEROPERATION: Translation Services, Interopration Services
- CAPABILITY TO AGENT MAPPING: Middle Agents
- NAME TO LOCATION MAPPING: ANS
- SECURITY: Certificate Authority, Cryptographic Services
- PERFORMANCE SERVICES: Monitoring, Regulation Services
- MULTIAGENT MANAGEMENT SERVICES: Logging, Activity Visualization, Launching
- ACL INFRASTRUCTURE: Public Ontology, Protocol Servers
- COMMUNICATION INFRASTRUCTURE: Discovery, Message Transfer

INDIVIDUAL AGENT INFRASTRUCTURE
- INTEROPERATION: Interopration Modules
- CAPABILITY TO AGENT MAPPING: Middle Agents Components
- NAME TO LOCATION MAPPING: ANS Components
- SECURITY: Security Module, Private/Public Keys
- PERFORMANCE SERVICES: Performance Services Modules
- MANAGEMENT SERVICES: Logging and Visualization Components
- ACL INFRASTRUCTURE: ACL Parser, Private Ontology, Protocol Engine
- COMMUNICATION MODULES: Discovery Component, Message Transfer Module

OPERATING ENVIRONMENT
- Machines, OS, Network
- Multicast, Transport Layer: TCP/IP, Wireless, Infrared, SSL

Agent Strategy Summarization


Explanation of Agent Conflicts & Harmful Interactions

### 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>Interoperaton 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 Components</td>
<td>AMS Components</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>MULTIAGENT MANAGEMENT SERVICES</strong></td>
</tr>
<tr>
<td>Logging, Auditing 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>Protocol 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>
</tbody>
</table>

**OPERATING ENVIRONMENT**

- MultiCast
- Transport Layer: TCP/IP, Wireless, Infrared, SSL

### Overview of explanation in different AI fields (7)

- **W. Lewis Johnson**: Agents that Learn to Explain Themselves. AAAI 1994: 1257-1263

### Explanation of Agent Conflicts & Harmful Interactions


### Agent Strategy Summarization


### Explainable Agents

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**


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


LIME for 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

**Exploreable NLP**


**LIME for NLP**

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>


XAI Plan

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


XAI Plan


(Manual) Plan Comparison
Overview of explanation in different AI fields (10)

Robotics

![Diagram of robot path]

**Narration of Autonomous Robot Experience**


Robotics

Overview of explanation in different AI fields (10)

Narration of Autonomous Robot Experience


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

Neurons respond to simple shapes

Neurons respond to more complex structures

Neurons respond to highly complex, abstract concepts

1st Layer

Input (unlabeled image)

Low-level features to high-level features

2nd Layer

nth Layer

Output Layer

73
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 / output / training data?
Knowledge Graph for Personalized XAI

Description 1: This is an orange train accident

Description 2: This is a 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
Knowledge Graph for Explaining Transfer Learning

“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-Zhejiang University Frontier Technology Research Center
Applications
Obstacle Identification Certification (Trust) - Transportation

**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
**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


Nicholas McCarthy, Mohammad Karzand, Freddy Lecue: Amsterdam to Dublin Eventually Delayed? LSTM and Transfer Learning for Predicting Delays of Low Cost Airlines: AAAI 2019
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.
Data analysis for spatial interpretation of abnormalities: abnormal expenses

Semantic explanation (structured in classes: fraud, events, seasonal) of abnormalities

Detailed semantic explanation (structured in sub classes e.g. categories for events)

**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)
Counterfactual Explanations for Credit Decisions (1) - Finance

- 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 (2) - Finance

Sorry, your loan application has been rejected.

Our analysis:

The following features were too high:
- PercentInstallTrad
- NetFractionRevol...
- NetFractionInstall...
- NumRevolvingTra...
- NumBank2NatTra...
- PercentTradesWB...

The following features were too low:
- MSinceOldestTrad...
- AverageMInFile
- 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

Drag sliders to change constraints.
External Risk Estimate
M Since Oldest Trade Open
M Since Most Recent Trade...M
Average M In File
Num Satisfactory Trades
Select categorical constraints.
Max Delq 2 Public Rec Last 12M
Max Delq Ever

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.

Results

These results are for women who have already had surgery. This table shows the percentage of women who survive at least 5, 10, 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.

David Spiegelhalter, Making Algorithms trustworthy, NeurIPS 2018 Keynote

predict.nhs.uk/tool
More on XAI
(Some) Tutorials, Workshops, Challenge

**Tutorial:**

**Workshop:**
- IJCAI 2019 Workshop on Optimisation and Explanation in AI (#1) - https://www.doc.ic.ac.uk/~kc2813/OXAI/
- ICAPS 2019 Workshop on Explainable Planning (#2) - https://kcl-planning.github.io/XAIP-Workshops/ICAPS_2019
- CD-MAKE 2019 – Workshop on Explainable AI (#2) - https://cd-make.net/special-sessions/make-explainable-ai/

**Challenge:**

**iNNvestigate**: A toolbox to iNNvestigate neural networks’ predictions. [github.com/albermax/ininvestigate](https://github.com/albermax/ininvestigate)

**SHAP**: SHapley Additive exPlanations. [github.com/slundberg/shap](https://github.com/slundberg/shap)

**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)

**LIME**: Agnostic Model Explainer. [https://github.com/marcotcr/lime](https://github.com/marcotcr/lime)

**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/ininvestigate](https://github.com/albermax/ininvestigate)

**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)

**Model describer**: Basic statistical metrics for explanation (visualisation for error, sensitivity). [https://github.com/DataScienceSquad/model-describer](https://github.com/DataScienceSquad/model-describer)
**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
(Some) Initiatives: XAI in Canada

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

- **Research institutions**
  - CRIAQ
  - IVADO
  - CRSNG NSERC

- **Industrial partners**
  - Bell
  - Helicopter Company
  - BOMBARDIER
  - CAE
  - 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
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*
Research and Technology Applied AI (Artificial Intelligence) Scientist

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

As a leading technology leader for the Defence Industry, the combined expertise of Thales R&T, Thales Canada and its Montreal-based AI team is a key player in keeping the pulse of technology, working to protect the national security interests of countres around the world.

Established in 1972, Thales Canada has over 1,800 employees in Toronto, Vancouver and Avon. This is a unique opportunity to play a key role on the Technology (R&T) in Canada (Quebec and Montreal) and work with an R&T experts at five locations worldwide developing cutting-edge AI technologies. Our passion is imagining and developing AI technologies that not only will you join but also integrate within our co-intelligence expertise (i.e., the new flagship program).

Job Description

An AI (Artificial Intelligence) Research and Technology (R&T) scientist is a key player at Thales Canada to develop innovative prototypes to demonstrate intelligence. To be successful in this role, one must possess a strong technical background and be familiar with the latest technology trends. A strong ability to learn new technologies and stay up-to-date with the latest developments is also necessary. In addition to technical skills, the successful candidate will also bring a strong business acumen, with the ability to understand and translate complex technical concepts into clear and concise business plans.

As a Research and Technology Applied AI Scientist, you will:

- Be part of a dynamic team focused on developing innovative AI technologies.
- Work closely with cross-functional teams to design and implement AI solutions.
- Participate in the development of AI-driven solutions for various sectors, including defence, security, and intelligence.
- Stay updated with the latest advancements in AI technology and continuously improve the existing solutions.
- Collaborate with internal and external stakeholders to identify and deliver AI-enabled solutions.

Professional Skill Requirements

- Good foundation in mathematics, statistics, and computer science.
- 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 analytic experience in Python with interest in artificial intelligence with working structured and unstructured data (SQL, Cassandra, MongoDB, Hive, etc.).
- A track record of outstanding AI software development with Github (or similar) evidence.
- Demonstrated abilities in designing large scale AI systems.
- Demonstrated interest in Explainable AI and/or relational learning.
- 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.

MAY 2ND, 2019

Freddy Lecue
Chief AI Scientist, CortAIx, Thales, Montreal – Canada

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