XAI - Explanation in AI: From Machine Learning to Knowledge Representation & Reasoning and Beyond

Freddy Lécué
Inria, France
CortAlx@Thales, Canada
@freddylecue

November 8th, 2019
Alberta Machine Intelligence Institute
University of Alberta, Edmonton, Alberta, Canada

Scope
AI Adoption: Requirements

- Valid AI
- Responsible AI

- Privacy-preserving AI
- Explainable AI

What is the rational?

Human Interpretable AI
Machine Interpretable AI
Explanation in AI

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.
Outline
Outline

• Explanation in Artificial Intelligence
  • Motivation
  • Definitions
  • Evaluation (with role of the human in XAI systems)
  • The Role of Humans
  • Explanations in Different AI fields

• On the Role of Knowledge Graph in Explainable Machine Learning

• XAI Industrial Applications using Knowledge Graphs on Machine Learning

• Conclusion + Q&A
Motivation
Business to Customer
Critical Systems
Markets We Serve (Critical Systems)

Aerospace  Space  Ground Transportation  Defence  Security

Trusted Partner For A Safer World
But not Only Critical Systems
COMPAS recidivism black bias

**DYLAN FUGETT**

Prior Offense
1 attempted burglary

Subsequent Offenses
3 drug possessions

**BERNARD PARKER**

Prior Offense
1 resisting arrest without violence

Subsequent Offenses
None

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.
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

Oliver Ralph  MAY 16, 2017

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.

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, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, Noemie Elhadad: Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. KDD 2015: 1721-1730
XAI in a Nutshell
XAI in a Nutshell.

Today

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

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

Tomorrow

- Training Data
- New Learning Process
- Obstacle on rail train
- Explainable Model
- Explanation Interface
- User with a Task

- I understand why
- I understand why not
- I know when you’ll succeed
- I know when you’ll fail
- I know when to trust you
- I know why you erred

Source: https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf
An Example of an end-to-end XAI System

- Humans may have follow-up questions
- Explanations cannot answer all users’ concerns

How to Explain? Accuracy vs. Explanability

Learning

• Challenges:
  • Supervised
  • Unsupervised learning

• Approach:
  • Representation Learning
  • Stochastic selection

• Output:
  • Correlation
  • No causation

Interpretability

Non-Linear functions

Polynomial functions

Quasi-Linear functions
XAI Objective
Supporting Industrialization of AI at Scale
Explainability by Design for AI Products

- Model Diagnostics
  - Root Cause Analytics
- Debug
- Performance monitoring
  - Fairness monitoring
- Monitor
- Model Comparison
  - Cohort Analysis
- A/B Test
- Predict
- Explainable Decisions
  - API Support
- Feedback Loop
- Train
- Model Debugging
  - Model Visualization
- QA
- Model Evaluation
  - Compliance Testing
- Deploy
- Model Launch
  - Signoff
  - Model Release Management
XAI Definitions

Explanation vs. Interpretability
explanation | ekspla'neɪʃ(ə)n |

noun

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

interpret | in'teəprɪt |

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

1 explain the meaning of (information or actions): the evidence is difficult to interpret.
On Role of Data In XAI
Interpretable Data for Interpretable Models

Table of baby-name data
(baby-2010.csv)

<table>
<thead>
<tr>
<th>name</th>
<th>rank</th>
<th>gender</th>
<th>year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jacob</td>
<td>1</td>
<td>boy</td>
<td>2010</td>
</tr>
<tr>
<td>Isabella</td>
<td>1</td>
<td>girl</td>
<td>2010</td>
</tr>
<tr>
<td>Ethan</td>
<td>2</td>
<td>boy</td>
<td>2010</td>
</tr>
<tr>
<td>Sophia</td>
<td>2</td>
<td>girl</td>
<td>2010</td>
</tr>
<tr>
<td>Michael</td>
<td>3</td>
<td>boy</td>
<td>2010</td>
</tr>
</tbody>
</table>

2000 rows all told

Tabular

Images

Text
What about the Evaluation?
Perturbation-based Evaluation for Feature Attribution-based Approaches

Perturb top-k features by attribution and observe change in prediction

- Higher the change, better the method
- Perturbation may amount to replacing the feature with a random value
- Samek et al. formalize this using a metric: Area over perturbation curve
  - Plot the prediction for input with top-k features perturbed as a function of k
  - Take the area over this curve

![Graph showing prediction for perturbed inputs and area over perturbation curve](image-url)
Human (Role)-based Evaluation is Essential... but too often based on size!

**Evaluation criteria** for Explanations [Miller, 2017]
- Truth & probability
- Usefulness, relevance
- Coherence with prior belief
- Generalization

**Cognitive chunks** = basic explanation units (for different explanation needs)
- Which basic units for explanations?
- How many?
- How to compose them?
- Uncertainty & end users?

[Doshi-Velez and Kim 2017, Poursabzi-Sangdeh 18]
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
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?
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 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 actions are responsible of a plan?

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?

Dependency Plot
Feature Importance
Surrogate Model
Planning
Plan Refinement
Artificial Intelligence
MAS
Computer Vision
Saliency Map
Uncertainty Map
Game Theory
Search
Robotics
NLP
KRR
UAI
Multi-Agent Systems (MAS)
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?

Which complex features are responsible of classification?

XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches
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 agent strategy & plan?
- Which player contributes most?
- Why such a conversational flow?

- Which complex features are responsible of classification?

Saliency Map

Uncertainty Map

Computer Vision

KRR

UAI

NLP

Machine Learning

Surrogate Model

Dependency Plot

Game Theory

Search

Conlicts Resolution

Shapely Values

Robotics

Plan Refinement

Artificial Intelligence

Plan Refinement

Strategy Summarization

Surrogate Model

Feature Importance

Dependency Plot

Artificial Intelligence

Game Theory

Search

Conlicts Resolution

Uncertainty Map

Computer Vision

KRR

UAI

NLP

Machine Learning

Surrogate Model

Dependency Plot

Artificial Intelligence

Game Theory

Search

Conlicts Resolution

Shapely Values

Robotics

Plan Refinement

Strategy Summarization

Surrogate Model

Feature Importance

Dependency Plot

Artificial Intelligence

Game Theory

Search

Conlicts Resolution

Uncertainty Map

Computer Vision

KRR

UAI

NLP
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?

Artificial Intelligence

Strategy Summarization

Saliency Map

Computer Vision

Uncertainty Map

KRR

MAS

UAI

NLP

Dependency Plot

Feature Importance

Surrogate Model

Plan Refinement

Planning

Search

Conflicts Resolution

Game Theory

Robotics

Narrative-based

Surrogate Model

Shapely Values

Narrative-based
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?

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

Which entity is responsible for classification?

Artificial Intelligence

Plan Refinement

Search

Dependency Plot

Feature Importance

Surrogate Model

Game Theory

Robotics

NLP

Saliency Map

Uncertainty Map

KRR

MAS

UAI

Computer Vision

Strategy Summarization

Dependency Plot

Feature Importance

Surrogate Model

Game Theory

Robotics

NLP

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

Machine Learning

Surrogate Model

Planning

Search

Plan Refinement

Artificial Intelligence

Computer Vision

Strategy Summarization

Dependency Plot

Feature Importance

Surrogate Model

Game Theory

Robotics

NLP

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

Machine Learning
Which complex features are responsible of classification?

Which constraints can be relaxed?

Which combination of features is optimal?

Which features are responsible of classification?

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

Which actions are responsible of a plan?

Which entity is responsible for 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?

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

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

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

Which combination of features is optimal?
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

- Machine Learning
- Computer Vision
- Search
- Planning
- Surrogate Model
- Dependency Plot

Questions:
- Which complex features are responsible for classification?
- Which actions are responsible for a plan?
- Which constraints can be relaxed?
- Which combination of features is optimal?
- Which decisions, combination of multimodal decisions lead to an action?
- Which features are responsible for classification?
- Which agent strategy & plan?
- Which player contributes most?
- Why such a conversational flow?
- Which entity is responsible for classification?
- Uncertainty as an alternative to explanation
- Which axiom is responsible for inference (e.g., classification)?
- Abduction/Diagnostic: Find the right root causes (abduction)?

Artificial Intelligence

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

Planning

Robotics

Game Theory

Computer Vision

Search

Plan Refinement

NLP

Surrogate Model

KRR

Uncertainty Map

Saliency Map

Dependency Plot

Shapely Values

Narrative-based

UAI

Machine Learning based

Diagnosis

Abduction

Conflicts Resolution

Which entity is responsible for classification?

UAI

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

UAI

Machine Learning based

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

KRR

Uncertainty as an alternative to explanation

Planning

Which actions are responsible for a plan?

Which constraints can be relaxed?

Which combination of features is optimal?

Shapely Values

Narrative-based

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

Robotics

Which features are responsible for classification?
On the Role of Knowledge Graphs in Explainable AI
A Machine Learning Perspective

On the Role of Knowledge Graph in Explainable AI - under open review at the Semantic Web Journal -
http://www.semantic-web-journal.net/content/role-knowledge-graphs-explainable-ai
Knowledge Graph (1)

- Set of (subject, predicate, object — SPO) **triples** - subject and object are **entities**, and **predicate** is the **relationship** holding between them.
- Each SPO **triple** denotes a **fact**, i.e. the existence of an actual relationship between two entities.

<table>
<thead>
<tr>
<th>subject</th>
<th>predicate</th>
<th>object</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>is interested in</td>
<td>The Mona Lisa</td>
</tr>
<tr>
<td>Bob</td>
<td>is a friend of</td>
<td>Alice</td>
</tr>
<tr>
<td>The Mona Lisa</td>
<td>was created by</td>
<td>Leonardo Da Vinci</td>
</tr>
<tr>
<td>Bob</td>
<td>is a</td>
<td>Person</td>
</tr>
<tr>
<td>La Joconde à W.</td>
<td>is about</td>
<td>The Mona Lisa</td>
</tr>
<tr>
<td>Bob</td>
<td>is born on</td>
<td>14 July 1990</td>
</tr>
</tbody>
</table>

![Knowledge Graph Diagram](image)
## Knowledge Graph (2)

<table>
<thead>
<tr>
<th>Name</th>
<th>Entities</th>
<th>Relations</th>
<th>Types</th>
<th>Facts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freebase</td>
<td>40M</td>
<td>35K</td>
<td>26.5K</td>
<td>637M</td>
</tr>
<tr>
<td>DBpedia (en)</td>
<td>4.6M</td>
<td>1.4K</td>
<td>735</td>
<td>580M</td>
</tr>
<tr>
<td>YAGO3</td>
<td>17M</td>
<td>77</td>
<td>488K</td>
<td>150M</td>
</tr>
<tr>
<td>Wikidata</td>
<td>15.6M</td>
<td>1.7K</td>
<td>23.2K</td>
<td>66M</td>
</tr>
<tr>
<td>NELL</td>
<td>2M</td>
<td>425</td>
<td>285</td>
<td>433K</td>
</tr>
<tr>
<td>Google KG</td>
<td>570M</td>
<td>35K</td>
<td>1.5K</td>
<td>18B</td>
</tr>
<tr>
<td>Knowledge Vault</td>
<td>45M</td>
<td>4.5K</td>
<td>1.1K</td>
<td>271M</td>
</tr>
<tr>
<td>Yahoo! KG</td>
<td>3.4M</td>
<td>800</td>
<td>250</td>
<td>1.39B</td>
</tr>
</tbody>
</table>

- **Manual Construction** - curated, collaborative
- **Automated Construction** - semi-structured, unstructured

Right: **Linked Open Data cloud** - over 1200 interlinked KGs encoding more than 200M facts about more than 50M entities.

Spans a variety of domains - Geography, Government, Life Sciences, Linguistics, Media, Publications, Cross-domain..
Knowledge Graph Construction

Knowledge Graph construction methods can be classified in:

- **Manual** — *curated* (e.g. via experts), *collaborative* (e.g. via volunteers)
- **Automated** — *semi-structured* (e.g. from infoboxes), *unstructured* (e.g. from text)

Coverage is an issue:

- **Freebase** (40M entities) - 71% of persons without a birthplace, 75% without a nationality, even worse for other relation types [Dong et al. 2014]
- **DBpedia** (20M entities) - 61% of persons without a birthplace, 58% of scientists missing why they are popular [Krompaß et al. 2015]

**Relational Learning** can help us overcoming these issues.
Knowledge Graph in Machine Learning (1)

Augmenting (input) features with more semantics such as knowledge graph embeddings / entities

https://stats.stackexchange.com/questions/230581/decision-tree-too-large-to-interpret
Knowledge Graph in Machine Learning (2)

Augmenting machine learning models with more semantics such as knowledge graphs entities

https://stats.stackexchange.com/questions/230581/decision-tree-too-large-to-interpret
Training Data

Neurons respond to simple shapes

Neurons respond to more complex structures

Neurons respond to highly complex, abstract concepts

Input (unlabeled image)

1st Layer

2nd Layer

nth Layer

Low-level features to high-level features

Augmenting (intermediate) features with more semantics such as knowledge graph embeddings / entities
Knowledge Graph in Machine Learning (4)

- **Input Layer**: Input (unlabeled image)
- **Hidden Layer**: Neurons respond to simple shapes, more complex structures, and highly complex, abstract concepts.
- **Output Layer**: Augmenting (input, intermediate) features – output relationship with more semantics to capture causal relationship.

Layers:
- **1st Layer**: Low-level features to high-level features
- **2nd Layer**: Neurons respond to more complex structures
- **nth Layer**: Neurons respond to highly complex, abstract concepts
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

Augmenting models with semantics to support personalized explanation
“How to explain transfer learning with appropriate knowledge representation?

Augmenting input features and domains with semantics to support interpretable transfer learning.
On One Industrial Application in Thales
State of the Art
Machine Learning
Applied to Critical Systems
Object (Obstacle) Detection Task
Object (Obstacle) Detection Task State-of-the-art ML Result
Object (Obstacle) Detection Task State-of-the-art ML Result

Boulder - .09

Railway - .11

Lumbermill - .59
State of the Art
XAI
Applied to Critical Systems
Object (Obstacle) Detection Task
State-of-the-art XAI Result

Lumbermill - .59
Unfortunately, this is of NO use for a human behind the system
Let’s stay back

Why this Explanation?
(meta explanation)
After Human Reasoning...

Lumbermill - .59

A sawmill or lumber mill is a facility where logs are cut into lumber. Prior to the invention of the sawmill, boards were rived (split) and planed, or more often sawn by two men with a whipsaw, one above and another in a saw pit below. The earliest known mechanical mill is the Hierapolis sawmill, a Roman water-powered stone mill at Hierapolis, Asia Minor dating back to the 3rd century AD. Other water-powered mills followed and by the 11th century they were widespread in Spain and North Africa, the Middle East and Central Asia, and in the next few centuries, spread across Europe. The circular motion of the wheel was converted to a reciprocating motion at the saw blade. Generally, only the saw was powered, and the logs had to be loaded and moved by hand. An early improvement was the developm (wikipedia: Sawmill)
What is missing?
In geology, a boulder is a rock fragment with size greater than 25.6 centimetres (10.1 in) in diameter. Smaller pieces are called cobbles and pebbles, depending on their "grain size." While a boulder may be small enough to move or roll manually, others are extremely massive. In common usage, a boulder is too large for a person to move. Smaller boulders are usually just called rocks or stones. The word boulder is short for boulder stone, from Middle English bouldestan or Swedish bulsten. Boulder sized clasts are found in some sedimentary rocks, such as coarse conglomerate and boulder clay.

Context matters

About: Boulder
An Entity of Type: place, from Named Graph: http://dbpedia.org, within Data Space: dbpedia.org

About: Rail transport
An Entity of Type: software, from Named Graph: http://dbpedia.org, within Data Space: dbpedia.org

Rail transport is a means of conveyance of passengers and goods on wheeled vehicles running on rails, also known as tracks. It is also commonly referred to as train transport. In contrast to road transport, where vehicles run on a prepared flat surface, rail vehicles (rolling stock) are directionally guided by the tracks on which they run. Tracks usually consist of steel rails, installed on ties (sleepers) and ballast, on which the rolling stock, usually fitted with metal wheels, moves. Other variations are also possible, such as slab track, where the rails are fastened to a concrete foundation resting on a prepared subsurface. Rolling stock in a rail transport system generally encounters lower frictional resistance than road vehicles, so passenger and freight cars (trains and wagons) can be coupled into longer trains. The operation is carried out by a railway company, providing transport between train stations or freight customer facilities. Power is provided by locomotives which either draw electric power from a railway electrification system or produce their own power, usually by diesel engines. Most tracks are accompanied by a signaling system. Railways are a safe land transport system when compared to other forms of transport. Railway transport is capable of high levels of passenger and cargo utilization and energy efficiency, but is often less flexible and more capital intensive than road transport, when lower traffic levels are considered. The oldest, man-powered railways date back to the 6th century BC, with Roman, one of the Seven Wonders of the World.
XAI Thales Platform

• Higher accuracy with no intensive fine-tuning
• Human interpretable explanation
• Running on the edge at inference time
Hardware: High performance, scalable, generic (to different FPGA family) & portable CNN dedicated programmable processor implemented on an FPGA for real-time embedded inference.

Software: Knowledge graph extension of object detection.

This is an **Obstacle: Boulder** obstructing the train: XG142-R on **Rail_Track** from City: Cannes to City: Marseille at **Location: Tunnel VIX** due to **Landslide**.
Rail Track - Boulder - Railway - Obstacle - Tunnel

Train operating on Tunnel obstructing Landslide

Tunnel - .74

Boulder - .81

Railway - .90
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/
- CD-MAKE 2019 – Workshop on Explainable AI (#2) - https://cd-make.net/special-sessions/make-explainable-ai/
- IEEE FUZZ 2019 / Advances on eXplainable Artificial Intelligence (#2) - https://sites.google.com/view/xai-fuzzieee2019
- International Conference on NL Generation - Interactive Natural Language Technology for Explainable Artificial Intelligence (EU H2020 NL4XAI; #1) - https://sites.google.com/view/nl4xai2019/

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)
**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

**Evaluation Framework**

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

---

**Data Analytics**
- Multimedia Data

**Autonomy**
- ArduPilot & SITL Simulation

---

**Evaluator**
(Some) Initiatives: XAI in Canada

- DEEL (Dependable Explainable Learning) Project 2019-2024
- Research institutions
- Industrial partners

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
Why do we Need XAI by the Way?

• **To empower** individual against undesired effects of automated decision making
• **To reveal** and protect new vulnerabilities
• **To implement** the “right of explanation”
• **To improve** industrial standards for developing AI-powered products, increasing the trust of companies and consumers
• **To help** people make better decisions
• **To align** algorithms with human values
• **To preserve** (and expand) human autonomy
• **To scale and industrialize** AI
Why do we Need Knowledge Graphs to Achieve XAI?

Because this is not an explanation from an intelligent system.

This is even not interpretable, and then not actionable.
Conclusion

• Explainable AI is motivated by real-world applications in AI

• Not a new problem – a reformulation of past research challenges in AI

• Knowledge graphs should be foundational for XAI

• But they are facing challenges related to their integration (data mapping)

• Many industrial applications already – crucial for AI adoption in critical systems
Open Research Questions for the Semantic Web / Knowledge Graph Community

• [Data] Machine learning experts do not buy the data – knowledge mapping
• [Explanation] There is no agreement on what an explanation is
• [Explanation] There is not a formalism for explanations (neither model nor output)
• [Model] There is very limited work in machine learning modules composability – and none from a semantics perspective
• [Model] There is no work on describing and representing models
• [Model] What are disentangled representations and how can its factors be quantified and detected?
• [Human-in-the-loop] There is no work that seriously addresses the problem of quantifying the grade of comprehensibility of an explanation for humans
Research and Technology Applied AI (Artificial Intelligence) Scientist

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

Thales is a global technology leader for the Defence, Aeronautics, Space, Security and Transport sectors. As a leader in emerging technologies, the combined expertise of Thales and CortAlx, a key player in keeping the pub, is enabling Thales to innovate and protect the national security interests of count...

Established in 1972, Thales Canada has over 1,800 employees in Toronto and Vancouver working in Defence, Avionics, Space, Security, and Transportation. Our passion is imagining cutting-edge technologies that will shape the future of society. Thales is also co-located within Coi Intelligence Expertise, the new flagship program to work.

Job Description

An AI (Artificial Intelligence) Research and Technology Applied AI Scientist is responsible for developing innovative prototypes to demonstrate the potential of AI technology. To be successful in this role, one must have a broad understanding of AI technologies and be familiar with the latest research. Experience is essential as technical subject matter experts and their businesses. In addition to the implementation, individuals will be involved in initiative thinking and team work.

Professional Skill Requirements

- Good foundation in mathematics, statistics, and scientific computing.
- 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 organizations.
- 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 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 ability 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.

AUGUST 28TH, 2019

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

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