On the Role of Knowledge Graphs in Explainable AI
A Machine Learning Perspective

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

October 27th, 2019
International Semantic Web Conference – Workshop on Semantic Explainability
Auckland, New Zealand

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

• 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

Gary Chavez added a photo you might be in.
about a minute ago · 📸
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

LOW RISK 3

BERNARD PARKER
Prior Offense
1 resisting arrest without violence
Subsequent Offenses
None

HIGH RISK 10

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

Olive 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.

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

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.


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
XAI in a Nutshell
XAI in a Nutshell.

**Today**

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

  - Obstacle on rail train
  - 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 → Explainable Model → Explanation Interface → User with a Task

  - Obstacle on rail train
  - 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](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
- Performance monitoring
  Fairness monitoring
- Model Comparison
  Cohort Analysis
- A/B Test
- Debug
- Monitor
- Train
- Feedback Loop
- Model Debugging
  Model Visualization
- QA
- Deploy
- Model Evaluation
  Compliance Testing
- Model Launch
  Signoff
  Model Release Mgmt
- Explainable Decisions
  API Support

XAI Definitions

Explanation vs. Interpretability
explanation  |  ekspləˈneɪʃ(ə)n |

noun

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

interpret  |  ɪnˈtɛrpri:t |

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

Images

Tabular

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

![Diagram showing prediction drop and area over perturbation curve](image)
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?

Saliency Map

Uncertainty Map

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

- Artificial Intelligence
- Planning
- Search
- Robotics
- Machine Learning
- Feature Importance
- Surrogate Model

Saliency Map

Uncertainty Map

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

• Computer Vision
• MAS
• KRR
• UAI
• NLP

Dependency Plot

Game Theory
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?
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 complex features are responsible of classification?

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

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?

Computer Vision

Uncertainty Map

Saliency Map

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

Artificial Intelligence

MAS

KRR

UAI

NLP

Dependency Plot

Feature Importance

Surrogate Model

Plan Refinement

Planning

Search

Conflicts Resolution

Game Theory

Surrogate Model

Robotics

Machine Learning

Game Theory

Dependency Plot

Game Theory

Computer Vision

NLP

KRR

UAI

MAS

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

Shapely Values
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 complex features are responsible of classification?

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

Artificial Intelligence

Plan Refinement

Computer Vision

KRR

UAI

Machine Learning based

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

Dependency Plot

Feature Importance

Surrogate Model

Search

Game Theory

Robotics

Narrative-based

Shapely Values

Strategy Summarization

Uncertainty Map

NLP

Uncertainty Map

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

Which entity is responsible for classification?
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

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

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

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

- Strategy Summarization
- Artificial Intelligence
- Plan Refinement
- MAS
- KRR
- UAI
- Computer Vision
- Diagnosis
- Abduction
- Uncertainty Map
- NLP
- Machine Learning based
- Uncertainty Map
- Saliency Map
- XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches
XAI: One Objective, Many ‘AI’s, Many Definitions, Many Approaches

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

Artificial Intelligence

Plan Refinement

Surrogate Model

Feature Importance

Dependency Plot

Search

Robotics

Game Theory

Conflicts Resolution

Computer Vision

Computation

KRR

UAI

Abduction

Uncertainty Map

Surrogate Model

Dependency Plot

Feature Importance

Machine Learning

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?
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 \((subject, predicate, object \rightarrow SPO)\) **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 (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
Neurons respond to simple shapes

Neurons respond to more complex structures

Neurons respond to highly complex, abstract concepts

1st Layer

2nd Layer

n-th Layer

Input (unlabeled image)

Low-level features to high-level features

Augmenting (input, intermediate) features – output relationship with more semantics to capture causal relationship
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?"


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

- **Huajun Chen**
  College of Computer Science, Zhejiang University, China
  Alibaba-Zhejiang University Frontier Technology Research Center

- **Ian Horrocks**
  Department of Computer Science
  University of Oxford, UK

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

Lumbermill - .59
Object (Obstacle) Detection Task State-of-the-art ML Result

Lumbermill - .59
Boulder - .09
Railway - .11
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...

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 (en)
What is missing?
Context matters

Boulder - .09

Railway - .11
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**


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)
(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

**Deep Learning Teams**

**TA 2: Psychological Model of Explanation**
- Psychological theories of explanation and develop a computational model of explanation from those theories

**Evaluator**

**TA 1: Explainable Learners**
- Explainable learning systems that include both an explainable model and an explanation interface

**TA 2: Psychological Model of Explanation**
- Psychological theories of explanation and develop a computational model of explanation from those theories

**Evaluation Framework**

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

**Interpretable Model Teams**

**Model Induction Teams**
(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 builds smarter solutions. Everywhere.

Thales is a global technology leader for the Defence, Aeronautics, Space, Security and Mobility markets. We provide solutions that contribute to the protection of lives and the safety and well-being of citizens, to the support of military operations and to the optimization of the functional performance of transportation systems.

Our combined expertise in operational analysis, cyber security, and the latest technologies enables us to provide unique and trusted solutions to protect the national security interests of countires worldwide.

Established in 1972, Thales Canada has over 1,800 employees working in Defence, Avionics, and Transportation, with three key sites in Toronto, Montreal and Vancouver. Thales currently has over 400 research and development professionals at five locations worldwide, with expertise in Machine Learning, Artificial Intelligence, Cyber Security, and others.

This is a unique opportunity to play a key role on Thales’ Technology (T2T) team. T2T is a group of experts at five locations worldwide, delivering cutting-edge AI technologies. Our passion is imagining new technology and making it a reality. To be successful in this role, you will work with a high profile team, to develop cutting-edge AI technologies, and contribute to the national security interests of Canada.

Job Description

An AI (Artificial Intelligence) Research and Technology position involves developing innovative prototypes to demonstrate the capabilities of our AI technologies. To be successful in this role, you will need to be: a) a strong AI scientist with relevant experience in developing machine learning algorithms, b) someone with an understanding of how AI can be used to solve complex problems in a variety of domains, and c) someone who can communicate effectively with both technical and non-technical stakeholders.

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
- Good foundation in mathematics, statistics, and programming
- 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.)

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 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