

Okay but Really... What is Explainable AI?

Notions and Conceptualizations of the Field

Derek Doran

Associate Professor

Dept. of Computer Science and Engineering

Wright State University, Dayton, OH, USA

<https://derk--.github.io>

<https://wsu-wacs.github.io>

November 28, 2018

Collaboration Credit

This presentation includes joint work with:

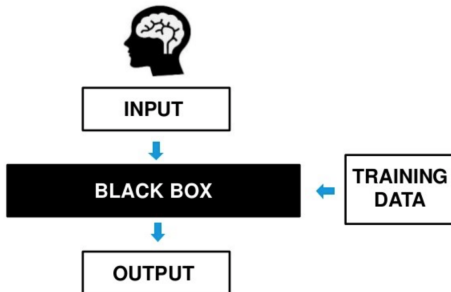
- ▶ Tarek Besold, Alpha Labs, Telefonica, Bacerlona, Spain
- ▶ Sarah Schulz, Institute for Natural Language Processing, University of Stuttgart

D. Doran, S. Schulz, and T. R. Besold. *What does Explainable AI Really Mean? A new Conceptualization of Perspectives.*
arxiv:170.00795 [cs.AI], arXiv.org, 2017

Context

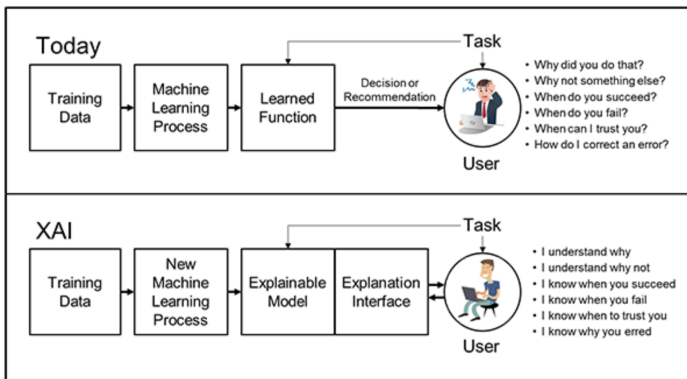
- ▶ To the layman (or expert), and **worse, to professionals** making or interpreting a decision by a **machine learning algorithm**, the view looks like...

MACHINE LEARNING



Context

Rationalizing the decisions an AI makes is crucial!



<http://www.darpa.mil/program/explainable-artificial-intelligence>



WRIGHT STATE
UNIVERSITY

Peering in

- ▶ **Symbolic systems** are inherently opaque: I can prove a fact is true given a knowledge base of facts, and show a user an inference sequence to prove the fact holds.

But not all is so clear for **sub-symbolic systems (statistical ML)**:

- ▶ **White-box**: You can see model mechanisms simple enough to be able to **trace how inputs map to outputs**
- ▶ **Grey-box**: You have **some vision** mechanisms, but parameters are numerous, decisions are probabilistic, or inputs get “lost” (transformed)
- ▶ **Black-box**: The model is so complex with so large a parameter space that it is not easily decipherable.

Peering in

Types of sub-symbolic systems:

- ▶ White Box: regression, DTs, rule mining, linear SVMs
- ▶ Grey Box: clustering, Bayesian nets, genetic algs, logic programming
- ▶ Black Box: DNNs, matrix factorizations, non-linear dimensionality reduction

The problem: Most white, some grey box models provide quantitative explanations (“variance explained”) that is, generally, not useful to a layman responsible for an action recommended by the model.

The Automatic Statistician

2.2 Component 2 : An approximately periodic function with a period of 1.0 years and with linearly increasing amplitude

This component is approximately periodic with a period of 1.0 years and varying amplitude. Across periods the shape of this function varies very smoothly. The amplitude of the function increases linearly. The shape of this function within each period has a typical lengthscale of 6.0 weeks.

This component explains 89.9% of the residual variance; this increases the total variance explained from 85.4% to 98.5%. The addition of this component reduces the cross validated MAE by 63.45% from 34.03 to 12.44.

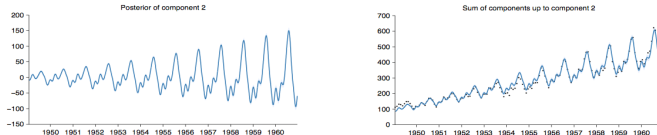


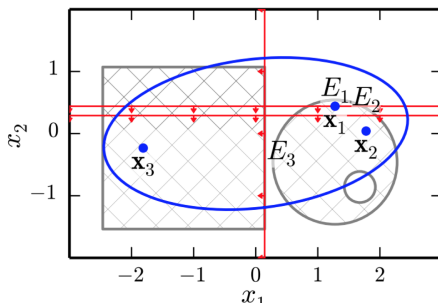
Figure 4: Pointwise posterior of component 2 (left) and the posterior of the cumulative sum of components with data (right)

Is the machine helping me understand the data?

Steinruecken, Christian, et al. "The Automatic Statistician."

“Explanations”: Linear Models

“Explainability” in ML models from linear decision boundaries
(think regression, decision trees, SVMs)



Red lines are the “explanation” (essentially a rule).
Classify x_3 as positive because x_3 satisfies E_2 and E_3 .

Turner, Ryan. “A Model Explanation System: Latest Updates and Extensions.” arXiv preprint arXiv:1606.09517 (2016).

“Explanations”: DNNs

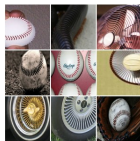
For DNNs: **emphasis on explaining the cause of activations** at feature learning and mixing network components.

Dataset Examples show us what neurons respond to in practice

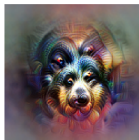
Tell me what inputs maximize the neuron's activation

Optimization isolates the causes of behavior from mere correlations. A neuron may not be detecting what you initially thought.

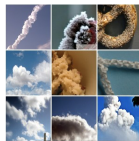
Optimize an input to maximize the neuron's activation



Baseball—or stripes?
mixed4a, Unit 6



Animal faces—or
snouts?
mixed4a, Unit 240



Clouds—or fluffiness?
mixed4a, Unit 453



Buildings—or sky?
mixed4a, Unit 492

Olah, et al., “Feature Visualization”, Distill, 2017.

What is an explanation, really?

Oxford Dict: *A statement or account that makes something clear; a **reason** or justification given for an action or belief.*

- ▶ “Innately explainable” white and grey box methods shed light into the **how**, not **why**, decisions are made. (Classify input as y because feature $x_1 < 10$ and $x_2 > 50$ and $x_3 = 0$).
- ▶ Methods enabling ‘XAI’ really just answer **how** (LIME and other methods that build local linear classifiers, for example); feature visualizations.
- ▶ **How** requires human-driven post processing under their own line of reasoning to create a *biased* explanation.

- 

WRIGHT STATE
on UNIVERSITY

-

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡

Our Theses

1. Explainable AI defines a **field of research**, *not define a particular “type” of AI*.
2. There exist unique notions of ‘explanations’, making problem or context-independent explainable AI work inappropriate.
3. ‘Strong’ and ‘Ultrastrong’ AI systems (by Michie’s definition) demands the integration of **symbolic systems and reasoning** to provide operationally effective communications of the internal state and operation of an AI.
4. Meaningful ontologies and knowledge bases, and Symbolic AI methods are a **fundamental** (but right now, mostly ignored) to the most important XAI use-case: when in practice, within some context, a layman must understand, trust, and be responsible for the conclusions an AI draws.



Our Theses

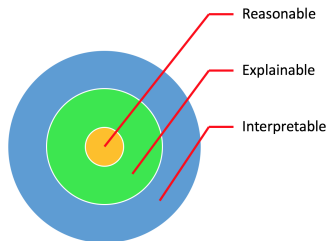
1. Explainable AI defines a **field of research**, *not define a particular “type” of AI.*
2. There exist unique notions of ‘explanations’, making problem or context-independent explainable AI work inappropriate.
3. ‘Strong’ and ‘Ultrastrong’ AI systems (by Michie’s definition) demands the integration of **symbolic systems and reasoning** to provide operationally effective communications of the internal state and operation of an AI.
4. **Meaningful ontologies and knowledge bases, and Symbolic AI methods are a fundamental** (but right now, mostly ignored) **to the most important XAI use-case: when in practice, within some context, a layman must understand, trust, and be responsible for the conclusions an AI draws.**



Notions of Explanation

Perhaps there is a hierarchy of explanation 'types' an AI system can exhibit:

- ▶ **Interpretable**: I can **identify** why an input goes to an output.
- ▶ **Explainable**: I can **explain** how inputs are mapped to outputs.
- ▶ **Reasonable**: I can **conceptualize** how inputs are mapped to outputs.



So you were denied a loan.

You go to a bank and are denied a loan. “Why was I denied?”

- ▶ (Interpretation): “Well the *account balance* covariate in the logistic regression model we use to make decisions explains 89.9% of residual variance.”

“What.... does that mean?”

- ▶ (Explanation): “It means our system denies loan applicants with low bank account balances.”

“What is the rationale for that?”

- ▶ (Reasoning): “Because the system does not want to award loans to those who do not show evidence of being able to pay them off.”

So you were denied a loan.

- ▶ **(Explanation)**: “It means our system denies loan applicants with low bank account balances.”

“What is the rationale for that?”

- ▶ **(Reasoning)**: “Because the system does not want to award loans to those who do not show evidence of being able to pay them off.”

Note how **reasoning** is a key element to obtain explanations and reasons! Here, the **explanation** comes from statistical interpretation, while **reasons** follow by the banker's logic!

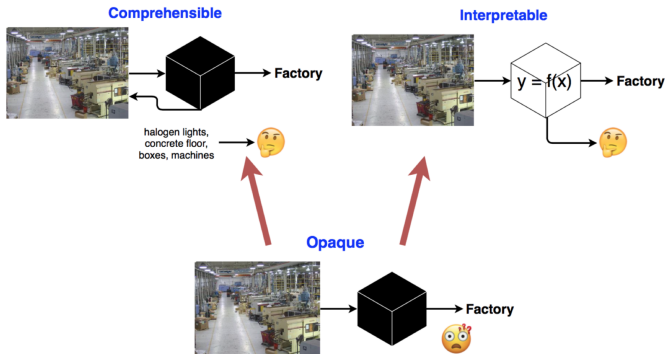
To Comprehend and Interpret

The history and present art of XAI for ML focuses on **how decisions are reached**. Consider a classification of ML systems into the following:

- ▶ **Opaque.** Like an oracle; all mechanisms mapping input to output are invisible to the user.
- ▶ **Interpretable.** User can see, study, and evaluate **how** mechanisms cause an input/output mapping. Interpretability implies *transparency*.
- ▶ **Comprehensible.** An opaque or interpretable model that emits **symbols** with its output that relate properties of the input to an output.

Comprehensibility is a graded notion; the ability to comprehend depends strongly on the type and quality of symbols.

To Comprehend and Interpret



Interpretable and Comprehensible AI systems only provide an understanding of **how** an AI processed an input.

Interpretable vs Comprehensible: A Doctor Analogy

To patients, physicians should be like a comprehensible model.

- ▶ Deliver prognosis by presenting high-level indicators (**symbols**) revealed in tests.

To other doctors and staff, physicians should be like an interpretable model.

- ▶ Present technicals of test results and how they link to a prognosis.
- ▶ Other doctors and staff can interpret technical details of test results and verify that the conclusions are supported by reasonable evaluation functions and evidence weighting.

What about to Reason?

- ▶ Interpretable and comprehensible systems lack in their ability to formulate a line of **reasoning** that explains why **symbols** or mathematical model details are connected to a particular model decision
- ▶ Interpretable and comprehensible systems **enable** explanations - but do not yield explanations themselves!
- ▶ A **reasoner** (of some sort) is needed to connect the input, output, and the how.

The Punch Line

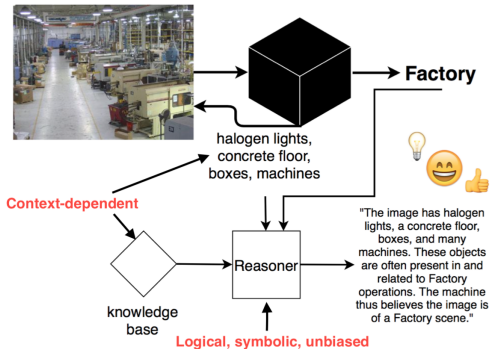
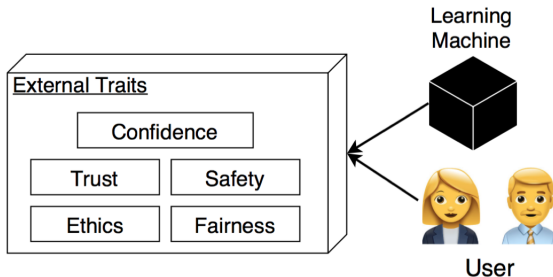


Fig. 5: Augmenting comprehensible models with a reasoning engine. This engine can combine symbols emitted by a comprehensible machine with a (domain specific) knowledge base encoding relationships between concepts represented by the symbols. The relationships between symbols in the knowledge based can yield a logical deduction about their relationship to the machine's decision.

Traits of Reasoning XAI Systems

Typical desiderata of XAI traits are only enabled through reasoning!



Pointers to what we are up to...

- WWW19 (Under Review)** Ebrahimi, Monireh, et al. "Reasoning over RDF Knowledge Bases using Deep Learning." arXiv preprint arXiv:1811.04132 (2018).
- NIPS18 (VIGiL Workshop)** Xie, Ning, et al. "Visual Entailment Task for Visually-Grounded Language Learning." arXiv preprint arXiv:1811.10582 (2018).
- ECDA18** Xie, Ning and Doran, Derek. "A Manifold Perspective for Debugging and Interpreting Deep Neural Networks", European Conference on Data Analysis, Paderborn, Germany, Jul. 2018.
- NeSy17** Sarker, Md Kamruzzaman, et al. "Explaining trained neural networks with semantic web technologies: First steps." arXiv preprint arXiv:1710.04324 (2017).
- CeX17** Doran, Derek, Sarah Schulz, and Tarek R. Besold. "What does explainable AI really mean? A new conceptualization of perspectives." arXiv preprint arXiv:1710.00794 (2017).

Thanks!

Thank you!