skema BUSINESS SCHOOL | AI R&D CENTER CANADA

What Is An Interpretation in AI?

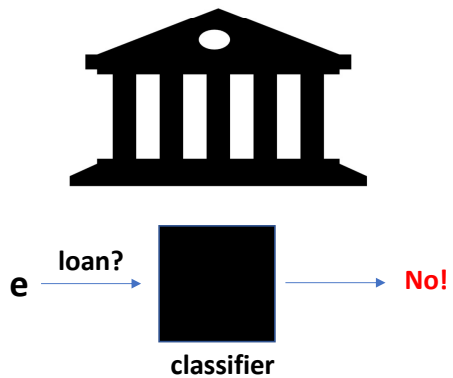
Leopoldo Bertossi

leopoldo.bertossi@skema.edu

Explanations in Machine Learning

- Bank client $\mathbf{e} = \langle \text{john}, 18, \text{plumber}, 70\text{K}, \text{harlem}, \ldots \rangle$

  As an entity represented as a record of values for features Name, Age, Activity, Income, ...
- **e** requests a loan from a bank, which uses a classifier



- The client asks *Why?*
- What kind of *explanation?*
  How?
  From what?

Explanations in AI

- This problem is representative of a more general situation in applications of AI systems
- Users and those affected by results from AI systems, the stakeholders, request explanations
- A whole new area of AI has emerged: *Explainable AI* (XAI)
- It is part of AI: (as opposed to about AI)

(a) AI systems should be extended with the capability to provide explanations

(b) AI researchers and professionals are those who understand these systems

So as mathematical logicians study the methods and scope of Math (with the methods of Math)

(c) Humans give explanations as part of their intelligent activities

Hence, explanation building should be a capability of AI agents

- Then, explanations have to be understood, modeled, implemented, ... as part of AI
- XAI is of interest to many other people
- We talked about stakeholders being affected by *outcomes from AI systems*

  Assessments (e.g. a credit score), classifications (good/bad client), decisions (approve/reject loan), etc.
- A whole discipline has emerged: *Ethical AI*
- It touches many others, including AI itself, but beyond: Law, Sociology, Philosophy, ..., Business, ...
- Naturally emerges, and motivated by need for more *fair, transparent, trustable, responsible, unbiased, ...* AI systems

- New legislation forces (owners of) AI systems affecting users to provide explanations and guarantee all the above
- There is a request for *interpretable AI systems*

← It may really be a "black box"!

A system so complex (after ML) that is practically a black box

E.g. Complex Neural Networks, Large Language Models, ...

- What is an interpretation of an AI system?
- What is an interpretation?

We will come back ...

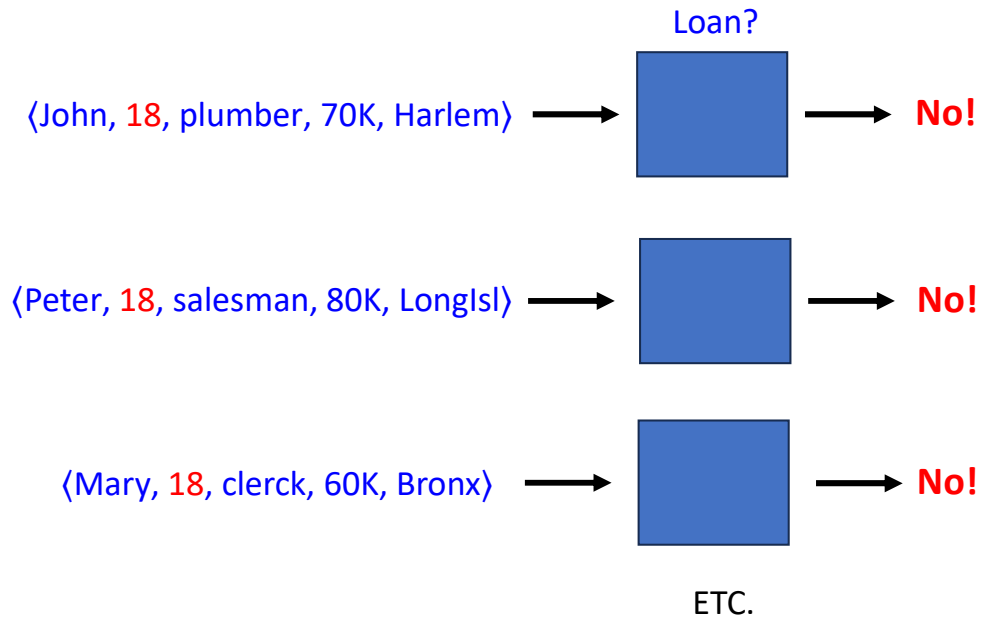
Explanations (in AI)

- Search for explanations belongs to the nature of human beings
- The quest has been around since the inception of humans
- Ancient Greeks already concerned with *causes* and *effects*
- Studied as such by Philosophers, Logicians, Physicists, ...
- Are explanations a new subject in AI?
- Yes and No
- Explanations have been studied in AI for some decades by now, and in related disciplines, e.g. Logic, Statistics
- Model-Based Diagnosis, Causality, etc.
- Some forms of explanations are new in AI and ML

  Others have roots in already existing ones

Some Forms of Explanation

- Abductive or Sufficient Explanations:

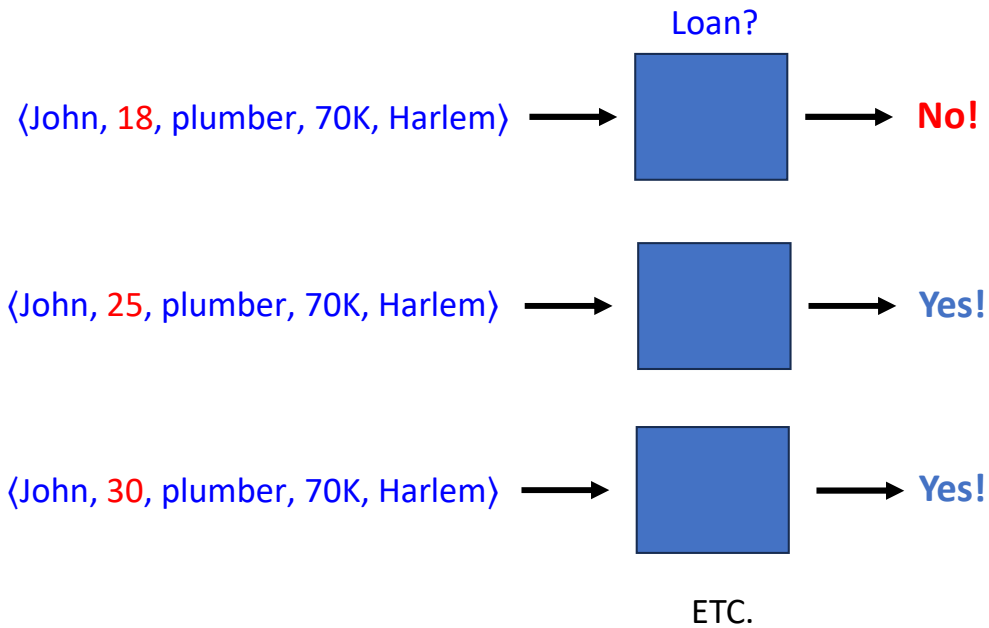


Fixed feature value $Age = 18$, no matter how other feature values change, the label does not change

$Age = 18$ is an abductive or sufficient explanation

It implies the observed label!

- Counterfactual Explanations:



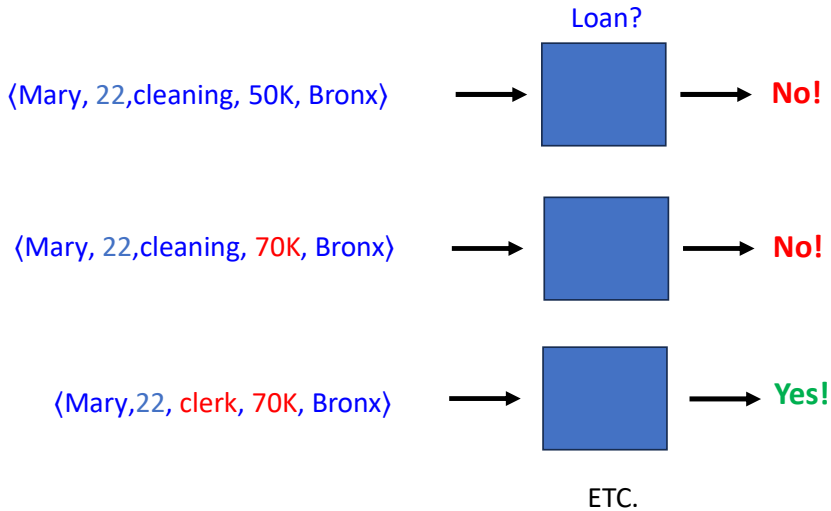
Fixed all other feature values, if feature *Age* were higher, the label would switch

$Age = 18$ is a counterfactual explanation

It is necessary for the observed label!

- Score-Based Explanations: (one of them here, simplified, for the gist)

Sometimes changing one feature value, no matter how, is not enough to switch the label, it may need company



Feature value *Salary = 70K* is actual counterfactual cause, with contingent value *Occupation = cleaning*

$$\mathit{ResponsibilityScore}(\mathit{Salary} = \mathit{70K}) = \frac{1}{1 + \text{minimum \# of cont. changes}} = \frac{1}{2}$$

A counterfactual cause (previous example) has responsibility 1

Feature values are ranked according to their causal strength

Interpretations

- Explanations as above, no matter how useful, can hardly be called "interpretations"

  If the latter refers to the AI system as a whole

- An interpretation has to do with the overall behaviour of the system

  As opposed to its input/output relation alone

- Interpretations have to do with elusive notions, such as:

  Understanding

  Meaning

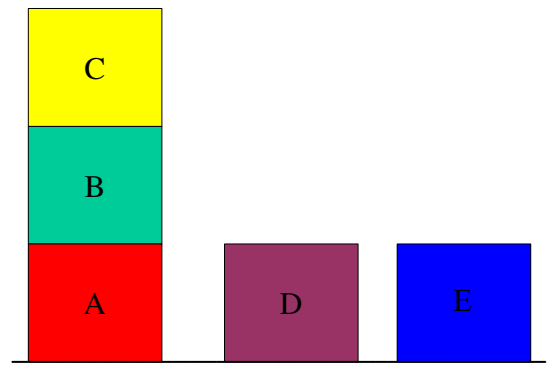
  Putting in context

  Making sense, etc.

- However, there is research in Logic and Computer Science

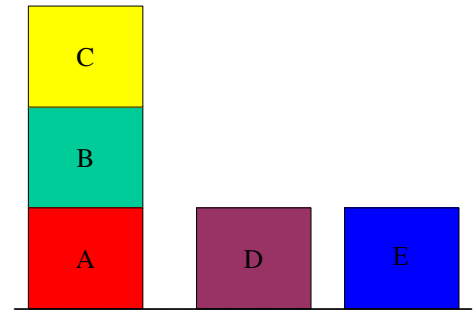
Interpretation Structures

- The classic example for a long time in AI: the Blocks World
- Intelligent agents should be able to reason about this world
- A robot should be able to move blocks around to reach a goal configuration, etc.
- In order to do this one needs a logical model
- Represented as a Knowledge Base (KB) of symbolic statements

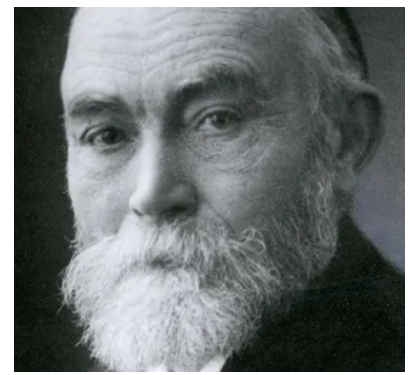


- To say things like:
  "Every object that is on top of a block is not on the floor", *"C is a yellow block"*, *"C is on top of B"*, *"A is to the left of D"*, *"There is a blue block"*
- Define new or extend old properties: *"A first object is to the left of a second object if it is on top of a third object that is to the left of the second"*

- To do computational reasoning from the representation To conclude (entail), e.g. *"B is to the left of D"*



- What kind of formulas in the KB?
- A language of First-Order Predicate Logic *"Begrifftsschrift und andere Aufsätze"* (Gottlob Frege, 1879)



- No wonder: Mathematical Logic is at the root of Computer Science, and most of initial and several of current approaches to AI
  - $Block(A)$, $On(B, A)$, $LeftOf(A, D)$, $A = A$, ...
  - $\neg\, A = B$, $(Block(A) \wedge On(B, A))$,

    $\forall x \forall y \forall z((LeftOf(x, y) \wedge On(z, x)) \rightarrow LeftOf(z, y))$

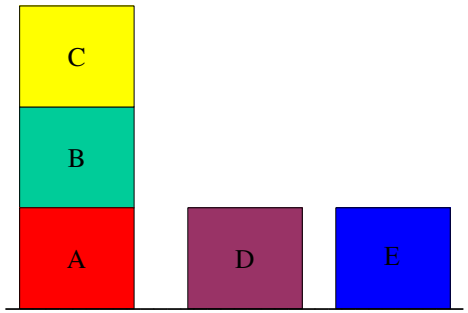
    "for all three objects, if ..."

- $\forall x((\exists y\, Block(y) \wedge On(x, y)) \rightarrow \neg On(x, floor))$

  "for every object, if there is a block ...."

- $\exists x(Block(x) \wedge \forall y(Block(y) \rightarrow \neg On(y, x)))$

  "there is a block that has no other block on top"



- This is all symbolic so far (except for the picture)
- There are automated reasoning systems that can do symbolic logical reasoning from this KB
- What is we want to verify that a symbolic entailment is a real consequence of the KB?

  In the usual mathematical/scientific (everyday) sense?
- What if we want to determine the truth of a symbolic statement?

- These questions have to do with the semantics of the symbolic formulas

  They are about meaning in a general sense

- We need to interpret symbols and formulas

  How? Where?

- In the external reality the symbolic statements are talking about?

  The picture of the Blocks World (BW)?

- We need a model of the BW

- An abstract representation of the essentials of the BW

  ... where the symbolic elements can be interpreted

- Mathematical Logic and Mathematics can help us ...
- Mathematician/Logician Alfred Tarski
  "The Notion of Truth in Formalized Languages"
  (1935)

- Which quickly led to the use of semantic structures to model an external reality or domain of discourse
- A set-theoretic structure that stays in correspondence with the symbolic language ...

  ... and can be used to interpret it
- Structures are representations in set-theoretic terms
- They have been and are commonly used in Mathematics

  Widely used since the late 30s (Nicolas Bourbaki)

A structure $\mathfrak{B}$ representing our BW:

- Domain/Universe: $\mathcal{U} = \{A, B, C, D, E, \mathit{green}, \mathit{yellow}, \mathit{red}, \mathit{purple}, \mathit{blue}, \ldots, \mathit{floor}, \ldots\}$
- Relations:

$$
\begin{array}{rcll}
\mathit{Block}^{\mathfrak{B}} & := & \{A, B, C, D, E\} & \text{(unary, i.e. } \subseteq \mathcal{U}) \\
\mathit{On}^{\mathfrak{B}} & := & \{(A, \mathit{floor}), (B, A), (C, B), (D, \mathit{floor}), (E, \mathit{floor})\} & \\
 & & & \text{(binary, } \subseteq \mathcal{U} \times \mathcal{U}) \\
\mathit{Color}^{\mathfrak{B}} & := & \{(A, \mathit{red}), \ldots\} & \\
\mathit{LeftOf}^{\mathfrak{B}} & := & \{(A, D), (D, E)\} & \\
=^{\mathfrak{B}} & := & \{(A, A), (B, B), \ldots, (\mathit{floor}, \mathit{floor})\} & \text{(usually left implicit)}
\end{array}
$$

- Distinguished individuals: $A, B, C, D, E, \mathit{green}, \ldots$
- $\mathfrak{B} = \langle \mathcal{U}, \mathit{Block}^{\mathfrak{B}}, \mathit{On}^{\mathfrak{B}}, \mathit{Color}^{\mathfrak{B}}, \mathit{LeftOf}^{\mathfrak{B}}, A, B, C, D, E, \mathit{green}, \ldots \rangle$ is a set-theoretic structure modeling BW
- Now we can put the formal language in correspondence with the structure

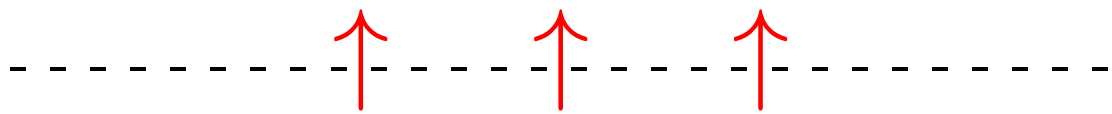
In general:

Meta-level, structural, interpretation level

$$\mathfrak{B} = \langle \mathcal{U}, Block^{\mathfrak{B}}, On^{\mathfrak{B}}, Color^{\mathfrak{B}}, LeftOf^{\mathfrak{B}}, A, B, C, D, E, green, \ldots \rangle$$

"there are $e_1, e_2 \in \mathcal{U}$ such that $(e_1, e_2) \in On^{\mathfrak{B}}$ and $(e_2, red) \in Color^{\mathfrak{B}}$"

Statement in the meta-language of usual Math



$$\exists x \exists y (On(x, y) \wedge Color(y, red))$$

Statement in the symbolic language

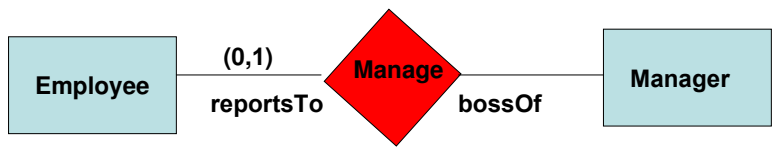
Symbolic, formal, object level

This formal statement should be true once interpreted in the BW structure

Ontological Interpretations

- Except for rare useful cases, e.g. Relational DBs, structures cannot be computationally represented or processed
- A weaker, non-equivalent alternative is to use an ontology as a model of an external reality

  Both a model and a KB describing it
- Ontology: a representation of a domain in terms of concepts (classes, entities) and relationships (roles) between concepts
- Example:

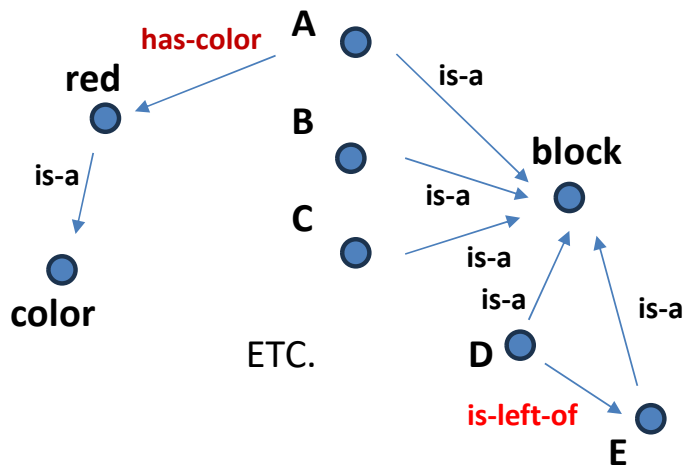


  - Unary predicates for concepts: $Employee(\cdot)$, $Manager(\cdot)$
  - Binary predicates for roles: $ReportsTo(\cdot,\cdot)$, $BossOf(\cdot,\cdot)$

  Symbolic statements in the ontology, e.g.

$$\forall x \forall y (BossOf(x,y) \rightarrow Employee(x)), \text{ etc.}$$

- Very common today: Ontologies as Knowledge Graphs (KGs)
- Multiple applications in Business (and other areas)
- Our BW as a KG:



- Easily stored, processed and queried inside a computer
- One can add rules, e.g. $LeftOf(x, y) \wedge LeftOf(y, z) \rightarrow LeftOf(x, z)$
  Transitivity of *LeftOf*
- We can think of using ontologies as interpretation "structures"
  Better, interpretation models

Back to Interpretations

- In our research we have used ontologies to model contexts

  For data quality (that is context-dependent)
- Interpretations could be achieved by mapping an AI system into a set-theoretic structure or an ontology
- An ontology could be used to interpret such a system
- There are extensions of Predicate Logic that can be used for describing dynamic processes

  As those involved in AI-based decision making or classification
- A full description of internals and behavior of an AI system may be out of reach (or not needed)

- The user or application domain may need only certain aspects are relevant to understand or make sense of the outcomes
- These relevant aspects can be modeled in structural or ontological interpretations
- They could be *mapped* into an ontology, for further computational use
- This sets a research agenda that has been very little developed
- These ideas can be applied to other AI processes that require understanding
- E.g. feature engineering: Why and what for are we choosing these features (and not those others) to build an ML-based system?

  An ontology could specify a *preference* relation, and other relations among features, etc.

Research Directions

(1) Explanation scores commonly use the classifier plus a probability distribution over the underlying entity population

Imposing or using explicit and additional domain semantics or domain knowledge is relevant to explore

Can we modify *Shap*'s definition and computation accordingly?

Or the probability distribution?

(2) Shapley values satisfy desirable properties for general coalition game theory

Specific properties for Explanations Scores (in AI)?

Existing scores have been criticized or under-explored in terms of general properties

(3) Features (in ML and in general) may be hierarchically ordered according to categorical dimensions

$$\text{address} \rightarrow \text{neighborhood} \rightarrow \textit{city} \rightarrow \cdots$$

We may want to define and compute explanations (scores) at different levels of abstraction

How to do this in a systematic way, possibly reusing results at different levels?

Multi-dimensional explanations?

(4) There is a need for principled and sensible algorithms for explanations and score aggregation

At the individual level as in (3) or at the group level, e.g. categories of instances

Hopefully guided by a declarative and flexible specifications (about what to aggregate and at which level)

(5) More informative and usable explanations

E.g. recommender systems may leave users puzzled by their recommendations

Provide explicit, declarative, and computable explanation (KG or ontology-like style)

(6) There is much research on fairness in data science and AI

Different approaches have been proposed

It would be good to have systems accepting and computing with different specifications of fairness

Algorithms Are Deciding Who Gets Organ Transplants. Are Their Decisions Fair?
Financial Times
Madhumita Murgia
November 9, 2023

The National Liver Offering Scheme was rolled out in the U.K. in 2018 to match livers with patients waiting for transplants based on their Transplant Benefit Score. However, concerns have been raised by some transplant patients and medical professionals due to a lack of understanding of how the algorithm works and the absence of an appeals process. Although the goal is to make transplant decisions fairer, an analysis by the Liver Advisory Group to the U.K. National Health Service found that patients aged 26 to 39 were waiting longer than they had before the algorithm, and longer than patients over 60. University of Cambridge's David Spiegelhalter said, "A range of subtle statistical issues appear to have unintentionally biased the algorithm against certain classes of patients."