INTRODUCTION TO MACHINE LEARNING EXPLAINABILITY

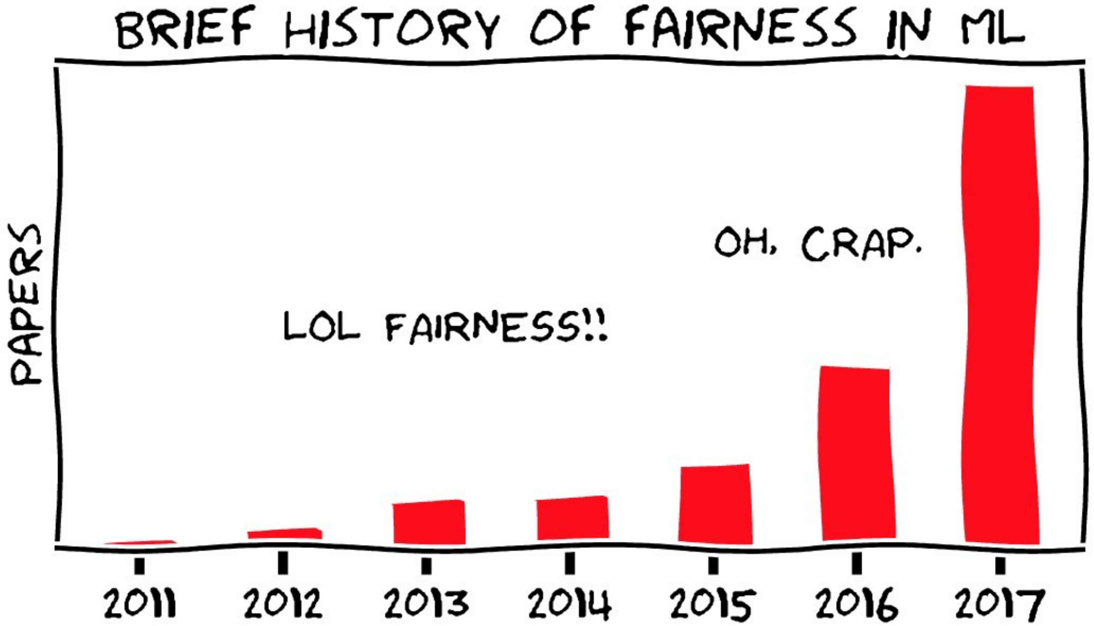
Part I

Kacper Sokol

TOPICS

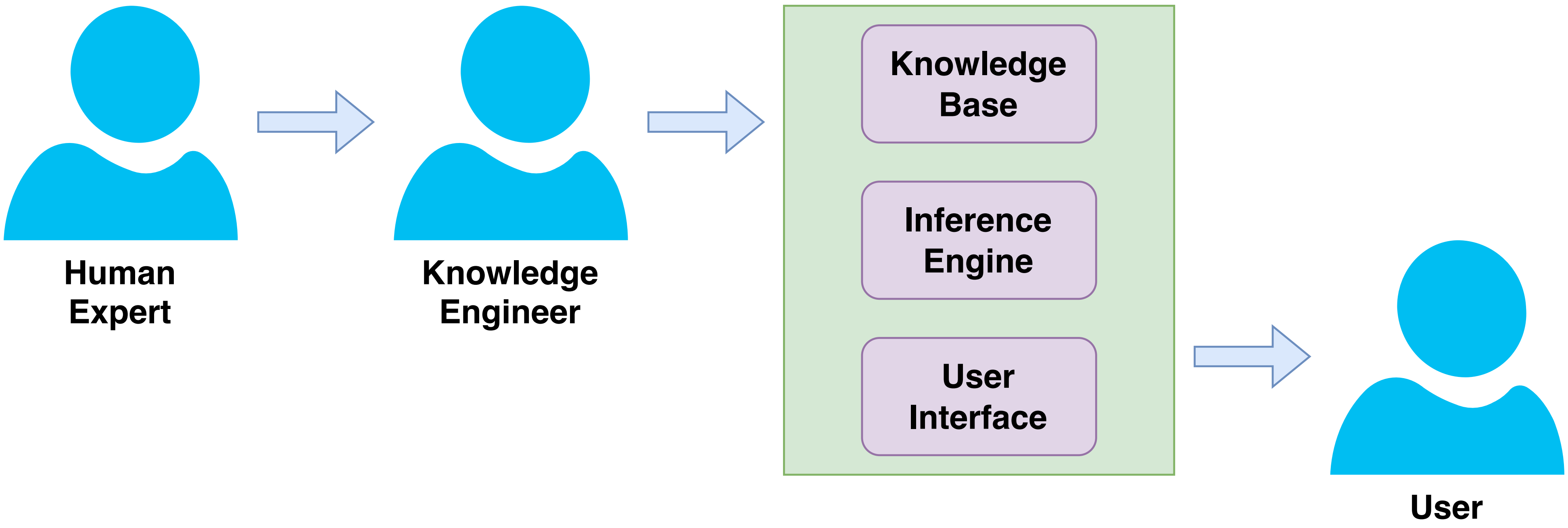
- Brief History of Explainability
- Why We Need Explainability
- Example of Explainability
- Important Developments
- Taxonomy of Explainable AI
- What Is Explainability?
- Evaluating Explainability
- Take-home Messages
- Useful Resources

BRIEF HISTORY OF EXPLAINABILITY

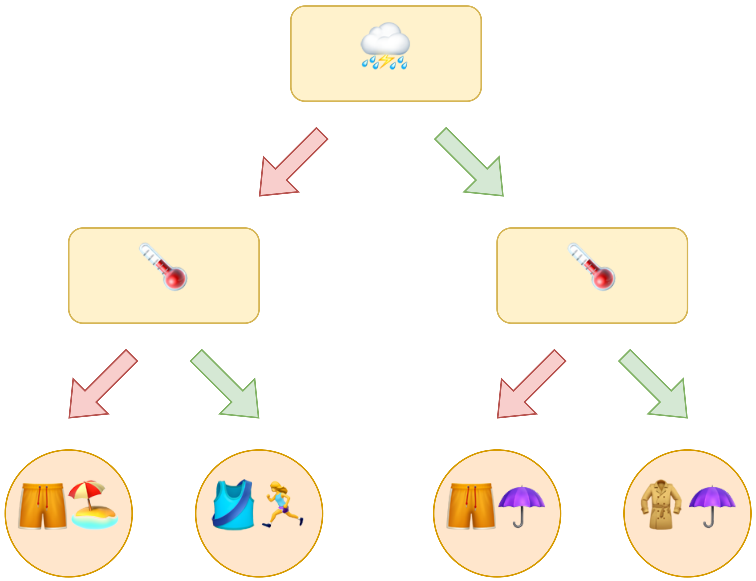


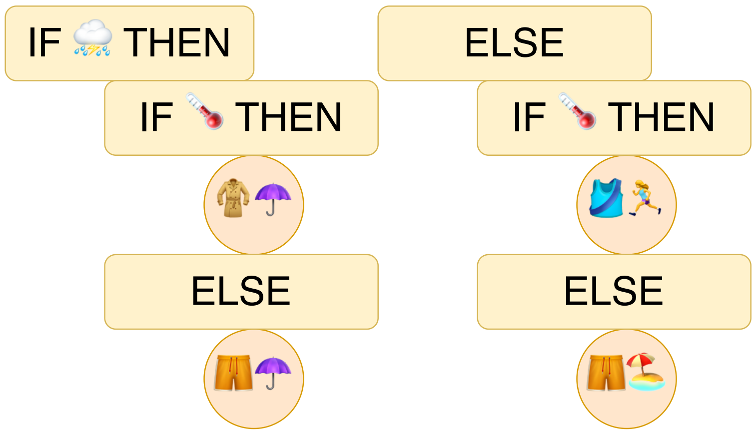
https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

EXPERT SYSTEMS (1970S & 1980S)

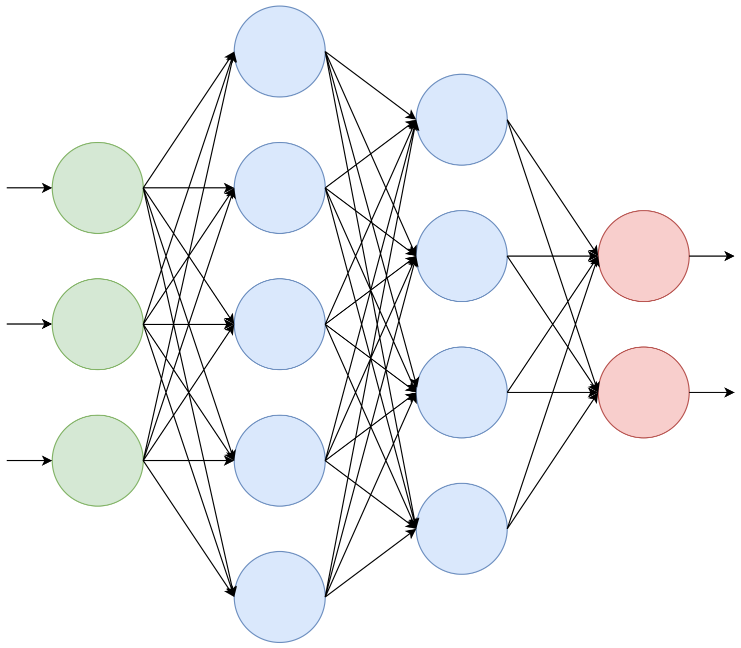


TRANSPARENT MACHINE LEARNING MODELS



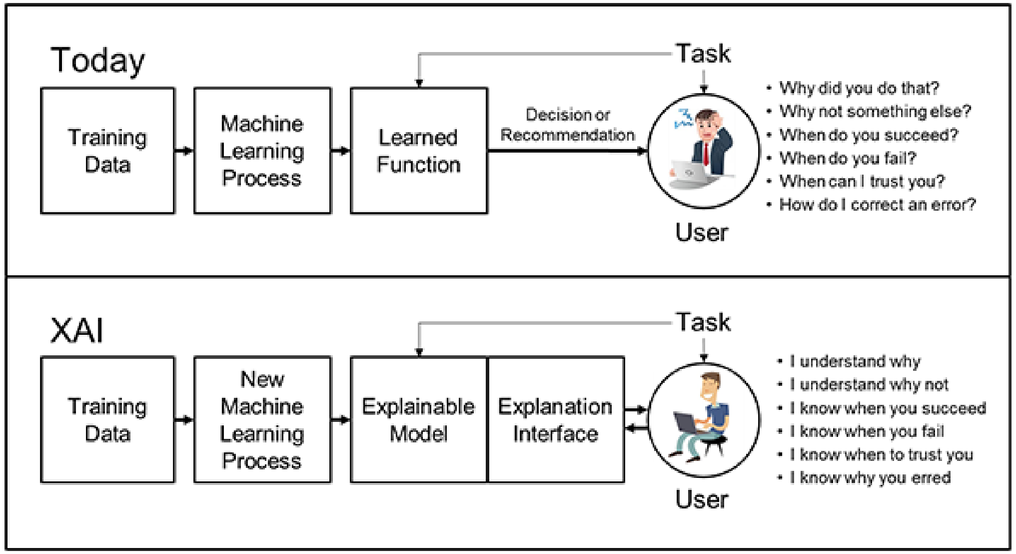


RISE OF THE DARK SIDE (DEEP NEURAL NETWORKS)



- No need to engineer features (by hand)
- High predictive power
- Black-box modelling

DARPA'S XAI CONCEPT



https://www.darpa.mil/program/explainable-artificial-intelligence

WHY WE NEED EXPLAINABILITY

BENEFITS

- Trustworthiness

> No silly mistakes

- Fairness

> Does not discriminate

- New knowledge

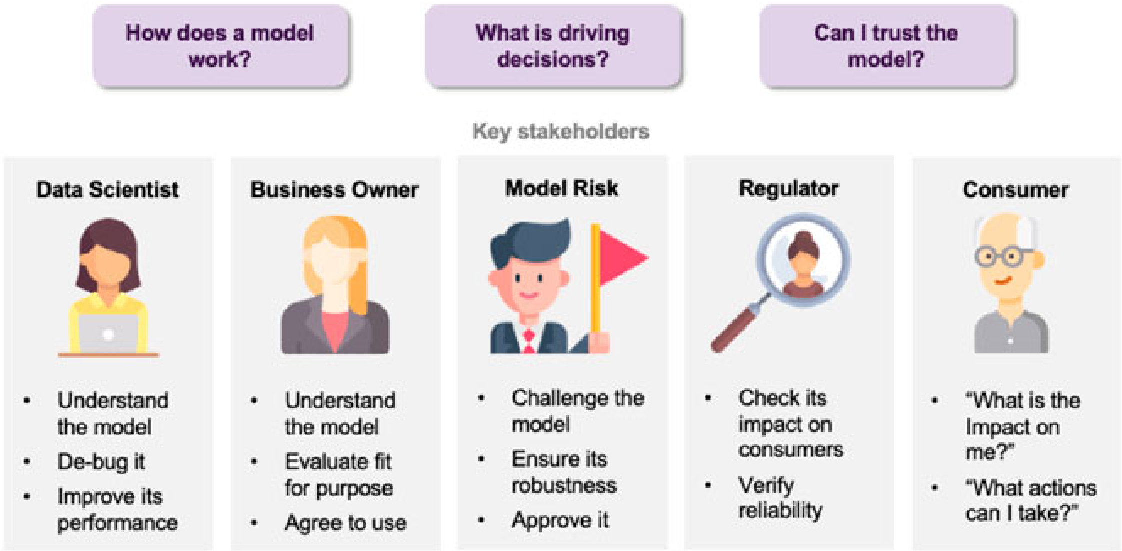
> Aids in scientific discovery

- Legislation

> Does not break the law

  - EU’s General Data Protection Regulation
  - California Consumer Privacy Act

STAKEHOLDERS



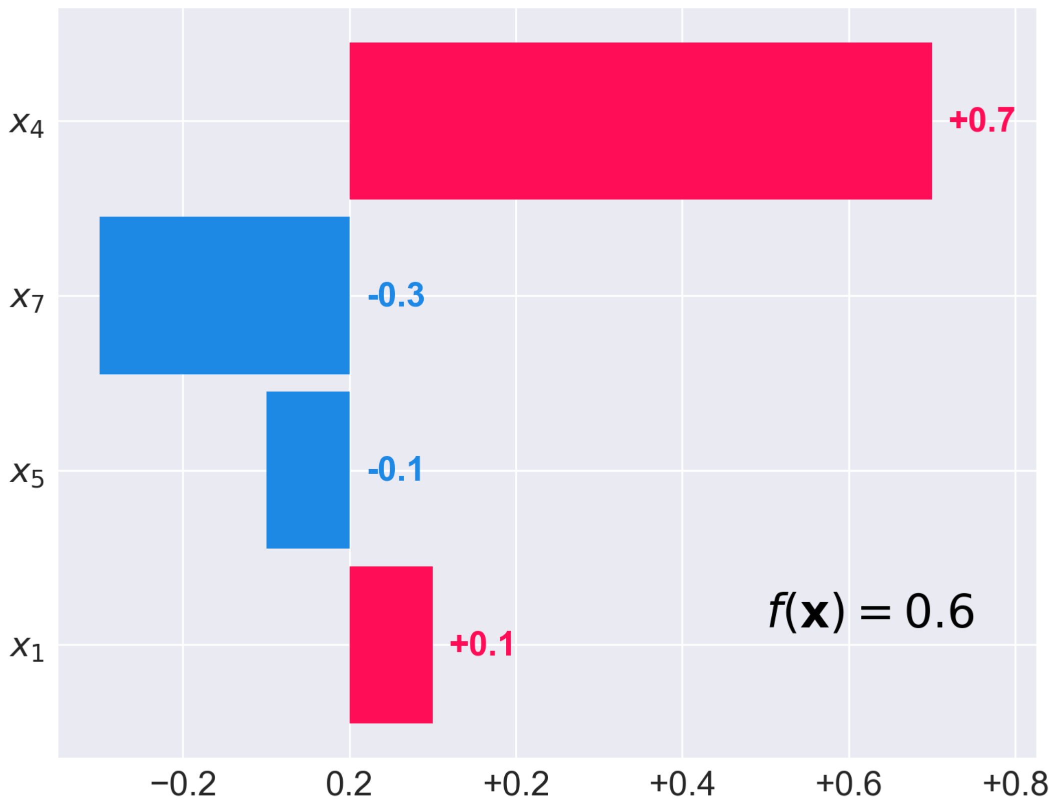
Belle and Papantonis, 2021. Principles and Practice of Explainable Machine Learning

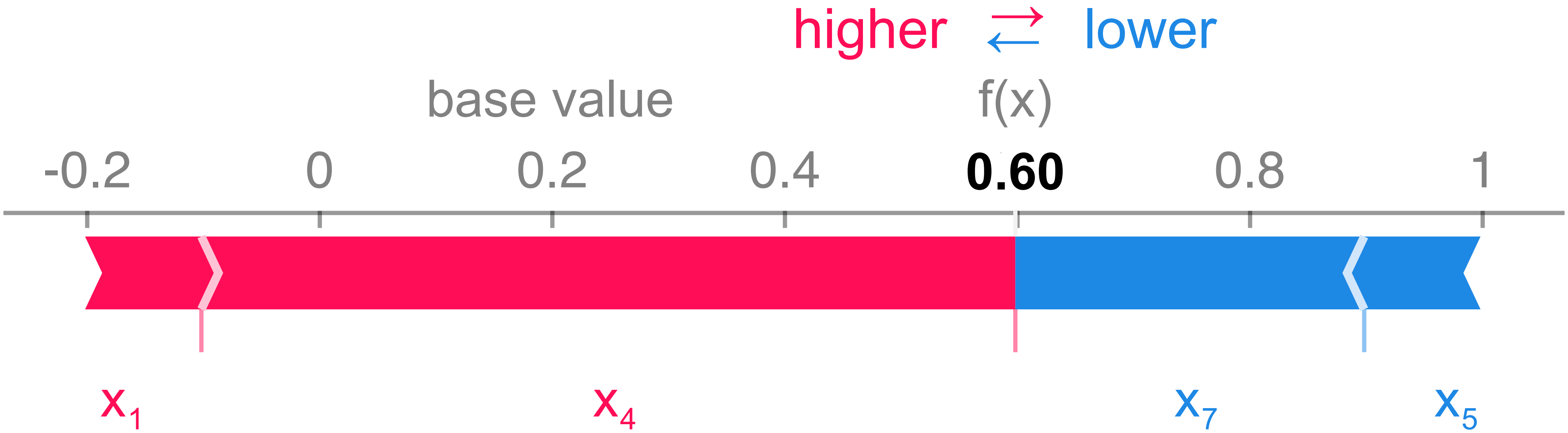
EXAMPLE OF EXPLAINABILITY

$$f(\mathbf{x}) = 0.2 \quad + \quad 0.25 \times x_1 \quad + \quad 0.7 \times x_4 \quad - \quad 0.2 \times x_5 \quad - \quad 0.9 \times x_7$$

$$\mathbf{x} = (0.4, \ldots, 1, \frac{1}{2}, \ldots \frac{1}{3})$$

$$f(\mathbf{x}) = 0.2 \quad \underbrace{+0.1}_{x_1} \quad \underbrace{+0.7}_{x_4} \quad \underbrace{-0.1}_{x_5} \quad \underbrace{-0.3}_{x_7} \quad = \quad 0.6$$


x_4
x_7
x_5
x_1
+0.7
-0.3
-0.1
+0.1
$f(\mathbf{x}) = 0.6$
−0.2
0.2
+0.2
+0.4
+0.6
+0.8


higher ⇄ lower
base value
f(x)
-0.2
0
0.2
0.4
0.60
0.8
1
x_1
x_4
x_7
x_5

IMPORTANT DEVELOPMENTS

WHERE IS THE HUMAN? (CIRCA 2017)

ELSEVIER

Artificial Intelligence
Volume 267, February 2019, Pages 1-38

Explanation in artificial intelligence: Insights from the social sciences

Tim Miller

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.artint.2018.07.007 Get rights and content

Abstract

There has been a recent resurgence in the area of explainable artificial intelligence as researchers and practitioners seek to provide more transparency to their algorithms. Much of this research is focused on explicitly explaining decisions or actions to a human observer, and it should not be controversial to say that looking at how humans explain to each other can serve as a useful starting point for explanation in artificial intelligence. However, it is fair to say that most work in explainable artificial intelligence uses only the researchers' intuition of what constitutes a 'good' explanation. There exist vast and valuable bodies of research in philosophy, psychology, and cognitive science of how people define, generate, select, evaluate, and present explanations, which argues that people employ certain

Abstract

There has been a recent resurgence in the area of explainable artificial intelligence as researchers and practitioners seek to provide more transparency to their algorithms. Much of this research is focused on explicitly explaining decisions or actions to a human observer, and it should not be controversial to say that looking at how humans explain to each other can serve as a useful starting point for explanation in artificial intelligence. However, it is fair to say that most work in explainable artificial intelligence uses only the researchers' intuition of what constitutes a 'good' explanation. There exist vast and valuable bodies of research in philosophy, psychology, and cognitive science of how people define, generate, select, evaluate, and present explanations, which argues that people employ certain

Miller, 2019. Explanation in artificial intelligence: Insights from the social sciences

HUMANS AND EXPLANATIONS

- Human-centred perspective on explainability
- Infusion of explainability insights from social sciences
  - Interactive dialogue (bi-directional explanatory process)
  - Contrastive statements (e.g., counterfactual explanations)

EXPLODING COMPLEXITY (2019)

Perspective | Published: 13 May 2019

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Nature Machine Intelligence **1**, 206–215 (2019) | Cite this article

41k Accesses | **476** Citations | **290** Altmetric | Metrics

Abstract

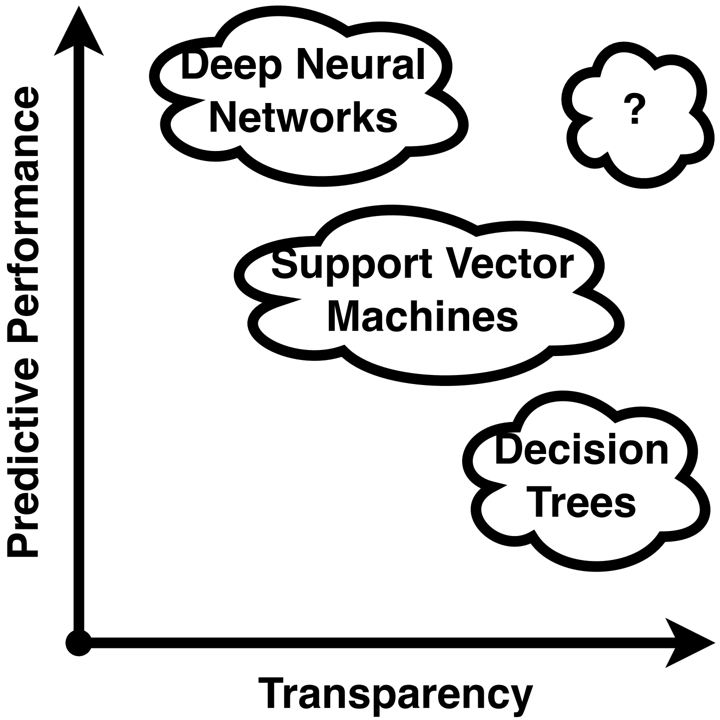
Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.

Abstract

Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.

Rudin, 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

ANTE-HOC VS. POST-HOC



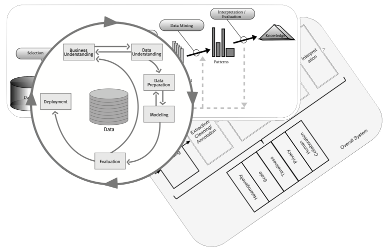
BLACK BOX + POST-HOC EXPLAINER

1. Chose a well-performing black-box model
2. Use explainer that is
   - *post-hoc* (can be retrofitted into pre-existing predictors)
   - and possibly *model-agnostic* (works with any black box)

CAVEAT: THE NO FREE LUNCH THEOREM

POST-HOC EXPLAINERS HAVE POOR FIDELITY

- Explainability needs a **process** similar to *KDD*, *CRISP-DM* or *BigData*



- Focus on engineering **informative features** and **inherently transparent models**

> **It requires effort**

XAI PROCESS

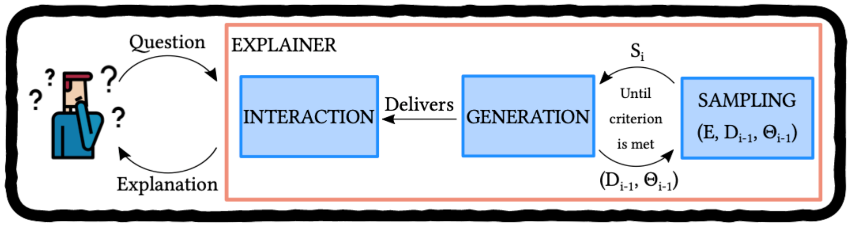
A **generic** eXplainable Artificial Intelligence process is *beyond our reach* at the moment

- **XAI Taxonomy** spanning social and technical desiderata:
  • Functional • Operational • Usability • Safety • Validation •

  (Sokol and Flach, 2020. Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches)

- **Framework** for black-box explainers

  (Henin and Le Métayer, 2019. Towards a generic framework for black-box explanations of algorithmic decision systems)



TAXONOMY OF EXPLAINABLE AI

(Explainability Fact Sheets)

Social and *technical* explainability desiderata spanning five dimensions

1. **functional** – algorithmic requirements
2. **usability** – user-centred properties
3. **operational** – deployment setting
4. **safety** – robustness and security
5. **validation** – evaluation, verification and validation

Sokol and Flach, 2020. Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches

Audience

- Researchers (*creators*)
- Practitioners (*users*): engineers & data scientists
- Compliance Personnel (*evaluators*): policymakers & auditors

Operationalisation

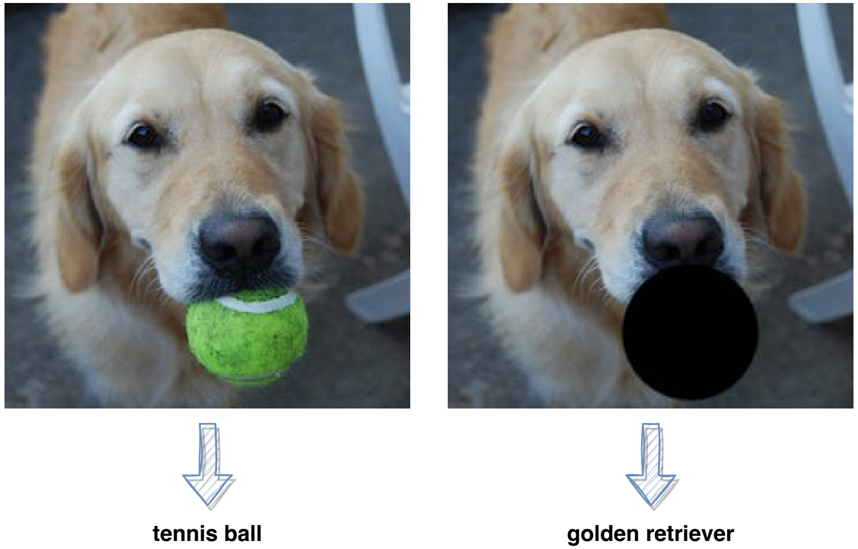
- Work Sheets: design & development
- Fact Sheets: assessment & comparison
- Checklist: inspection, compliance, impact & certification

Applicability

- Explainability Approaches (*theory*)
- Algorithms (*design*)
- Implementations (*code*)

RUNNING EXAMPLE: COUNTERFACTUAL EXPLANATIONS

> Had you been **10 years younger**, your loan application would be **accepted**.



(F) FUNCTIONAL REQUIREMENTS

- **F1** Problem Supervision Level
- **F2** Problem Type
- **F3** Explanation Target
- **F4** Explanation Breadth/Scope
- **F5** Computational Complexity
- **F6** Applicable Model Class
- **F7** Relation to the Predictive System
- **F8** Compatible Feature Types
- **F9** Caveats and Assumptions

F1 Problem Supervision Level	• unsupervised • semi-supervised • supervised • reinforcement

F2 Problem Type	• classification
	  ▪ probabilistic / non-probabilistic
	  ▪ binary / multi-class
	  ▪ multi-label
	• regression
	• clustering

F6 Applicable Model Class	• model-agnostic • model class-specific • model-specific
F7 Relation to the Predictive System	• **ante-hoc** (based on endogenous information) • post-hoc (based on exogenous information)

F5 Computational Complexity	• off-line explanations • real-time explanations
F8 Compatible Feature Types	• numerical • categorical (one-hot encoding)
F9 Caveats and Assumptions	• any underlying assumptions, e.g., black box linearity

F3 Explanation Target	• data (both raw data and features) • models • **predictions**
F4 Explanation Breadth/Scope	• **local** – data point / prediction • cohort – subgroup / subspace • global

(U) USABILITY REQUIREMENTS

- **U1** Soundness
- **U2** Completeness
- **U3** Contextfullness
- **U4** Interactiveness
- **U5** Actionability
- **U6** Chronology
- **U7** Coherence
- **U8** Novelty
- **U9** Complexity
- **U10** Personalisation
- **U11** Parsimony

U1 Soundness	How truthful it is with respect to the black box?	(✔)
U2 Completeness	How well does it generalise?	(✗)
U3 Contextfullness	"It only holds for people older than 25."	
U11 Parsimony	How short is it?	(✔)

U6 Chronology	More recent events first.
U7 Coherence	Comply with the natural laws (mental model).
U8 Novelty	Avoid stating obvious / being a truism.
U9 Complexity	Appropriate for the audience.

U5 Actionability	Actionable foil.	(✔)
U4 Interactiveness	User-defined foil.	(✔)
U10 Personalisation	User-defined foil.	(✔)

(O) OPERATIONAL REQUIREMENTS

- **O1** Explanation Family
- **O2** Explanatory Medium
- **O3** System Interaction
- **O4** Explanation Domain
- **O5** Data and Model Transparency
- **O6** Explanation Audience
- **O7** Function of the Explanation
- **O8** Causality vs. Actionability
- **O9** Trust vs. Performance
- **O10** Provenance

O1 Explanation Family	• associations between antecedent and consequent • **contrasts and differences** • causal mechanisms
O2 Explanatory Medium	• (statistical / numerical) summarisation • **visualisation** • **textualisation** • formal argumentation
O3 System Interaction	• **static** – one-directional • **interactive** – bi-directional

O4 Explanation Domain	• **original domain** (exemplars, model parameters) • **transformed domain** (interpretable representation)
O5 Data and Model Transparency	• **transparent/opaque data** • transparent/opaque model
O6 Explanation Audience	• **domain experts** • **lay audience**

O7 Function of the Explanation	• **interpretability** • **fairness** (disparate impact) • **accountability** (model robustness / adversarial examples)
O8 Causality vs. Actionability	• **look like causal insights but aren't**
O9 Trust and Performance	• **truthful** to the black-box (perfect fidelity) • predictive performance is **not affected**

- **predictive model**
- data set

O10 Provenance

- predictive model and data set (explainability trace)

(S) SAFETY REQUIREMENTS

- **S1** Information Leakage
- **S2** Explanation Misuse
- **S3** Explanation Invariance
- **S4** Explanation Quality

S1 Information Leakage	Contrastive explanation **leak** precise values.
S2 Explanation Misuse	Can be used to **reverse-engineer** the black box.
S3 Explanation Invariance	Does it always output the same explanation (stochasticity / stability)?
S4 Explanation Quality	Is it from the data distribution? How far from a decision boundary (confidence)?

(V) VALIDATION REQUIREMENTS

- **V1** User Studies
- **V2** Synthetic Experiments

V1 User Studies

- Technical correctness
- Human biases
- Unfounded generalisation
- Usefulness

V2 Synthetic Experiments

EXAMPLES

👩‍🔬 RESEARCHER'S 🎩

- 🔍 only works with predictive models that **output numbers** (**F2** *Problem Type*)
    - Is 🔍 intended for regressors?
    - Can 🔍 be used with probabilistic classifiers?

- 🔍 only works with **numerical features** (**F8** *Compatible Feature Types*)
  - If data have categorical features, is applying one-hot encoding suitable?

- 🔍 is **model agnostic** (**F6** *Applicable Model Class*)
  - Can 🔍 be used with any predictive model?

- 🔍 has nice **theoretical properties** (**F9** *Caveats and Assumptions*)

  > The explanation is always **[insert your favourite claim here]**.

  - This claim may not hold for **every black-box** model (model agnostic explainer)
  - The implementation **does not adhere** to the claim

👨‍💻 ENGINEER'S 🎩

- 🔍 explains **song recommendations** (**O7** *Function of the Explanation*)
- 🔍 explains how users' **listening habits** and **interactions** with the service influence the recommendations (**O10** *Provenance* & **U5** *Actionability*)

- How does 🔍 scale? (**F5** *Computational Complexity*)
  - Required to serve explanations in **real time**
  - Will the computational complexity of the algorithm introduce any **lags**?

- **Music listeners** are the recipients of the explanations (**O6** *Explanation Audience*)
  - They are not expected to have any ML experience or background (**U9** *Complexity*)
- They should be familiar with **general music concepts** (genre, pace, etc.) to appreciate the explanations (**O4** *Explanation Domain*)

- The explanations will be delivered as **snippets of text** (**O2** *Explanatory Medium*)
- They will include a single **piece of information** (**U11** *Parsimony*)
- They are **one-directional** communication (**O3** *System Interaction* & **U4** *Interactiveness*)

🕵️‍♀️ AUDITOR'S 🎩

- Are the explanations **sound** (**U1**) and **complete** (**U2**)?
  - Do they agree with the predictive model?
  - Are they coherent with the overall behaviour of the model?
- Are the explanations placed in a **context**? (**U3** *Contextfullness*)
  - "This explanation only applies to songs of this particular band."

- Will I get the **same explanation** tomorrow? (**S3** *Explanation Invariance*)
  - Confidence of the predictive model
  - Random effects within the 🔍 algorithm

- Does the explainer **leak any sensitive information**? (**S1** *Information Leakage*)
  - →*explanation*←
    "Had you been older than 30, your loan application would have been approved."
  - →*context*←
    "This age threshold applies to people whose annual income is upwards of £25,000."
- Why don't **I "round up"** my income the next time? (**S2** *Explanation Misuse*)

- Was 🔍 **validated** for the problem class that it is being deployed on? (**V2** *Synthetic Validation*)
- Does 🔍 **improve users' understanding**? (**V1** *User Studies*)

Local Interpretable Model-agnostic Explanations

This is an *Explainability Fact Sheet* for Local Interpretable Model-agnostic Explanations (LIME). It is distributed as a supplementary material of the "Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches" paper (Kacper Sokol and Peter Flach, 2020) published in Conference on Fairness, Accountability, and Transparency (FAT* 2020).

Approach Characteristic

Description

Local Interpretable Model-agnostic Explanations (LIME) is a surrogate explainability method that aims to approximate a local decision boundary with a sparse linear model to interpret individual predictions. It was introduced by this paper and

CHALLENGES

- The desiderata list is neither **exhaustive** nor **prescriptive**
- Some properties are **incompatible** or **competing** – choose wisely and justify your choices
  - Should I focus more on property F42 or F44?
  - For O13, should I go for X or Y?
- Other properties cannot be answered **uniquely**
  - E.g., coherence with the user's mental model
- The taxonomy **does not define explainability**

WHAT IS EXPLAINABILITY?

(You know it when you see it!)

LACK OF A UNIVERSALLY ACCEPTED DEFINITION

- **Simulatability**
  (*Lipton, 2018. The mythos of model interpretability*)
- **The Chinese Room Theorem**
  (*Searle, 1980. Minds, brains, and programs*)
- **Mental Models**
  (*Kulesza et al., 2013. Too much, too little, or just right? Ways explanations impact end users' mental models*)
  - **Functional** – operationalisation without understanding
  - **Structural** – appreciation of the underlying mechanism

DEFINING EXPLAINABILITY

$$\texttt{Explainability} = \underbrace{\texttt{Reasoning}\left(\texttt{Transparency} \mid \texttt{Background Knowledge}\right)}_{understanding}$$

- *Transparency* – **insight** (of arbitrary complexity) into operation of a system
- *Background Knowledge* – implicit or explicit **exogenous information**
- *Reasoning* – **algorithmic** or **mental processing** of information

Sokol and Flach, 2021. Explainability Is in the Mind of the Beholder: Establishing the Foundations of Explainable Artificial Intelligence

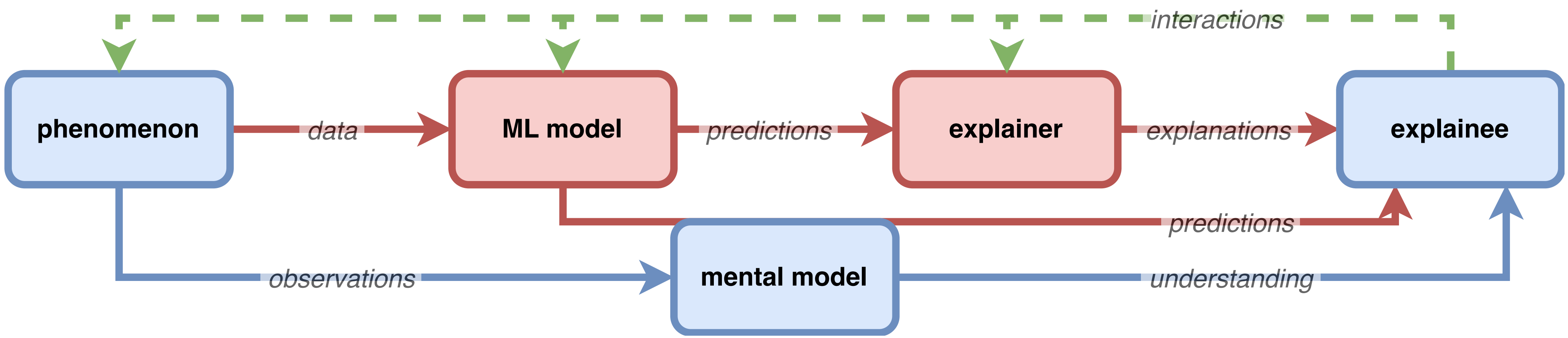
Explainability → **explainee** walking away with **understanding**

UNDERSTANDING, EXPLAINABILITY & TRANSPARENCY

A **continuous spectrum** rather than a binary property

EVALUATING EXPLAINABILITY

AUTOMATED DECISION-MAKING



NAÏVE VIEW

Does the explanation work?

No → **Yes**

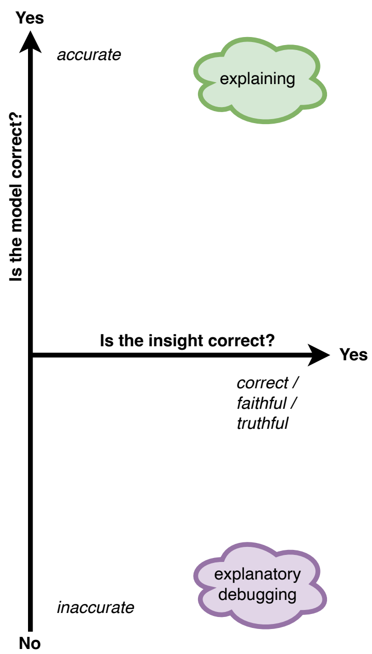
EVALUATION TIERS

	Humans	Task
Application-grounded Evaluation	Real Humans	Real Tasks
Human-grounded Evaluation	Real Humans	Simple Tasks
Functionally-grounded Evaluation	No Real Humans	Proxy Tasks

Kim and Doshi-Velez, 2017. Towards A Rigorous Science of Interpretable Machine Learning

EXPLANATORY INSIGHT & PRESENTATION MEDIUM

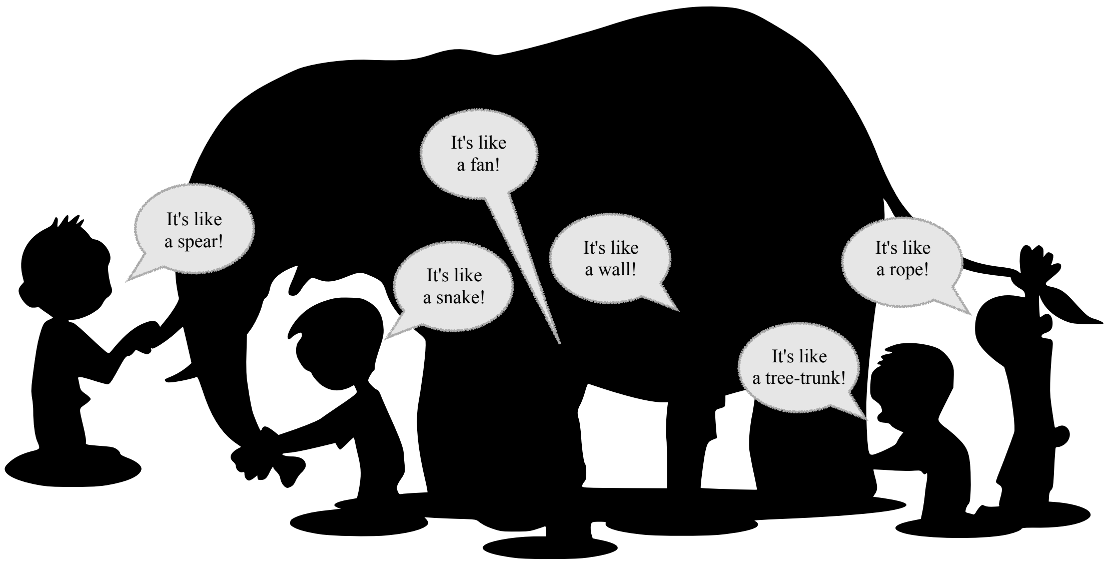
PHENOMENON & EXPLANATION



TAKE-HOME MESSAGES

Each (real-life) explainability scenario is **unique** and requires a **bespoke solution**

> Explainers are **socio-technical** constructs, hence we should strive for **seamless integration with humans** as well as **technical correctness and soundness**



(The Blind Men and the Elephant)

USEFUL RESOURCES

📖 BOOKS

- Survey of machine learning interpretability in form of an online book
- Overview of explanatory model analysis published as an online book
- Hands-on machine learning explainability online book (*URL to follow*)

PAPERS

- General introduction to interpretability
- Introduction to human-centred explainability
- Critique of post-hoc explainability
- Survey of interpretability techniques
- Taxonomy of explainability approaches

SOFTWARE

- LIME (Python, R)
- SHAP (Python, R)
- Microsoft's Interpret
- Oracle's Skater
- IBM's Explainability 360
- FAT Forensics