# Explainability

Last updated by | Mateusz Macias | Jun 15, 2020 at 12:22 PM GMT+1

## Explainability

### Definition

**Transparency and explainability** - Perhaps the greatest challenge that AI poses from a governance perspective its complexity and opacity. Not only can it be difficult to understand from a technical perspective, but early experience has already proven that it’s not always clear when an AI system has been implemented in a given context, and for what task.
The eight principles within the theme of Transparency and Explainability are a response to these challenges:

- The principle of "**transparency**” is the assertion that AI systems should be designed and implemented in such a way that oversight of their operations are possible.

> Transparency throughout an AI system’s life cycle means openness throughout the design, development, and deployment processes. While most documents treat transparency as binary — that is, an AI system is either transparent or it is not — several articulate the transparency principle as one that entities will strive for, with increased disclosure over time.

- “**Explainability**” is defined in various ways, but at its core is about the translation of technical concepts and decision outputs into intelligible, comprehensible formats suitable for evaluation.

> Many of the documents note that explainability is particularly important for systems that might “cause harm", have “a significant effect on individuals,” or impact “a person’s life, quality of life, or reputation.”

- The principle of “**open source data and algorithms**” is, as noted in the introduction to this theme, a familiar concept in technology governance, and it operates similarly in the context of AI as in other computer systems.

- “**Open government procurement,**” the requirement that governments be transparent about their use of AI systems.

> “When a government body seeks to acquire an AI system or components thereof, procurement should be done openly and transparently according to open procurement standards. This includes publication of the purpose of the system, goals, parameters, and other information to facilitate public understanding. Procurement should include a period for public comment, and states should reach out to potentially affected groups where relevant to ensure an opportunity to input.”

- The “**right to information**” concerns the entitlement of individuals to know about various aspects of the use of, and their interaction with, AI systems. This might include “information about the personal data used in the decision-making process, access to the factors, the logic, and techniques that produced the outcome” of an AI system, and generally “how automated and machine learning decision-making processes are reached.”

- The definition of the principle of “**notification when an AI system makes a decision about an individual**” is facially fairly clear: where an AI has been employed, the person to whom it was subject should know.

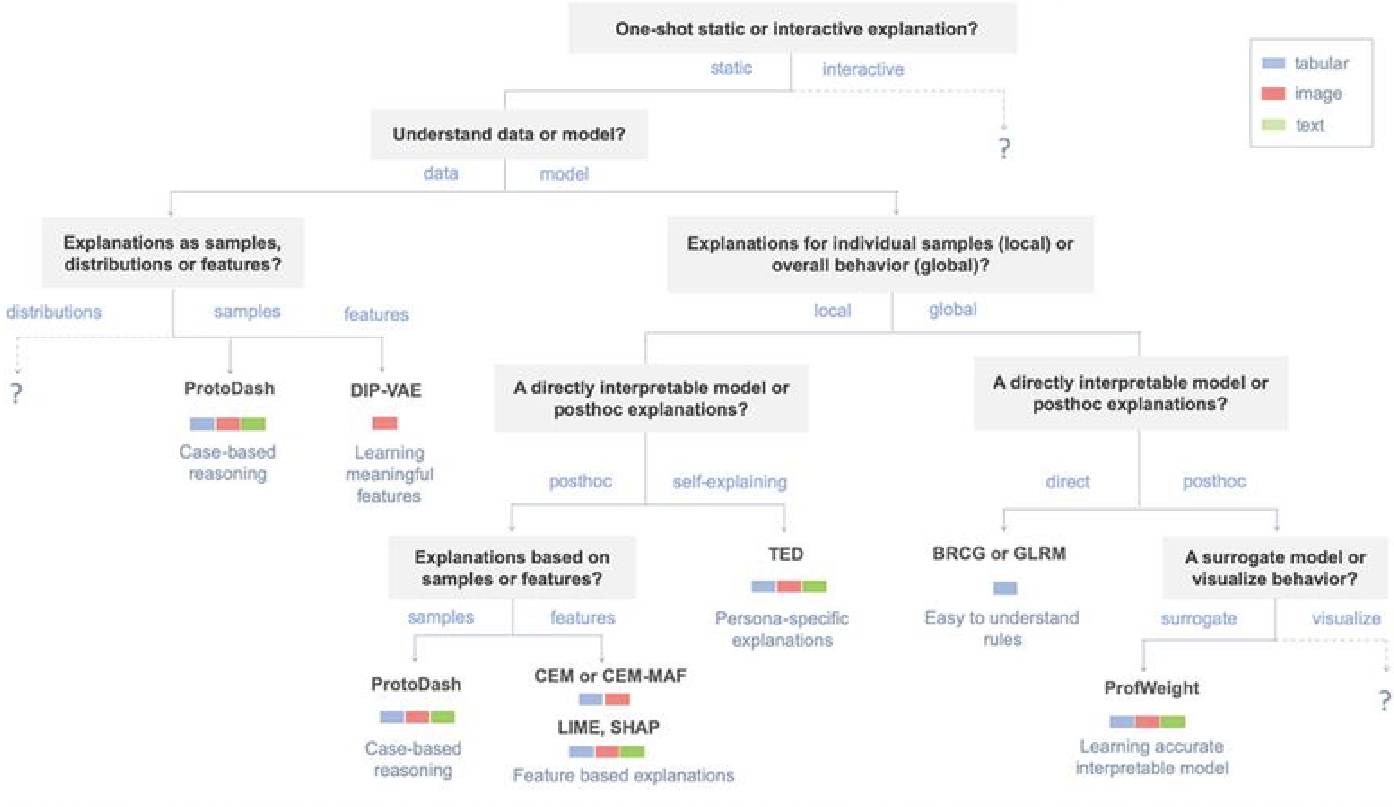
> If people don’t know when they are subject to automated decisions, they won’t have the autonomy to decide whether or not they consent, or the information to reach their own conclusions about the overall value that AI provides.

- The principle of “**notification when interacting with an AI system**” a recognition of AI’s increasing ability to pass the Turing test at least in limited applications, stands for the notion that humans should always be made aware when they are engaging with technology rather than directly with another person. Examples of when this principle is relevant include chatbot interactions, facial recognition systems, credit scoring systems, and generally “where machine learning systems are used in the public sphere.”

- “**Regular reporting**” as a principle stands for the notion that organisations that implement AI systems should systematically disclose important information about their use. This might include “how outputs are reached and what actions are taken to minimise rights-harming impacts, discovery of … operating errors, unexpected or undesirable effects, security breaches, and data leaks,” or the “evaluation of the effectiveness” of AI systems.

## Explainability methods

### Decision tree



One-shot static or interactive explanation?
static
interactive
?
tabular
image
text
Understand data or model?
data
model
Explanations as samples, distributions or features?
distributions
samples
features
?
ProtoDash
Case-based reasoning
DIP-VAE
Learning meaningful features
Explanations for individual samples (local) or overall behavior (global)?
local
global
A directly interpretable model or posthoc explanations?
posthoc
self-explaining
A directly interpretable model or posthoc explanations?
direct
posthoc
Explanations based on samples or features?
samples
features
ProtoDash
Case-based reasoning
CEM or CEM-MAF
LIME, SHAP
Feature based explanations
TED
Persona-specific explanations
BRCG or GLRM
Easy to understand rules
A surrogate model or visualize behavior?
surrogate
visualize
ProfWeight
Learning accurate interpretable model
?

### IBM explainability toolkit

#### ProtoDash

ProtoDash provides exemplar-based explanations for summarising datasets as well as explaining predictions made by an AI model. It employs a fast gradient based algorithm to find prototypes along with their (non-negative) importance weights. The algorithm minimises the maximum mean discrepancy metric and has constant factor approximation guarantees for this weakly sub modular function.

Research

Live demo

Notebook

### DIP-VAE (Disentangled Inferred Prior-VAE)

Note: can only be applied to images.

DIPVAE Explainer can be used to visualise the changes in the latent space of Disentangled Inferred Prior-VAE or DIPVAE. This model is a Variational Autoencoder that leads to a disentangled latent space. This is achieved by matching the covariance of the prior distributions with the inferred prior.

Research

Method in action applied on Fashion-MNIST dataset (image dataset)

### GLRM (Generalised Linear Rule Model)

GLRM provides access to the following directly interpretable supervised learning methods:

Linear Rule Regression
Logistic Rule Regression
Research

### BRCG (Boolean Rule Column Generation)

BooleanRuleCG is a directly interpretable supervised learning method for binary classification that learns a Boolean rule in disjunctive normal form (DNF) or conjunctive normal form (CNF) using column generation (CG). AIX360 implements a heuristic beam search version of BRCG that is less computationally intensive than the published integer programming version.

Research

### CEM (Contrastive Explanations Method)

This can be used to compute contrastive explanations for image and tabular data. This is achieved by finding what is minimally sufficient (PP - Pertinent Positive) and what should be necessarily absent (PN - Pertinent Negative) to maintain the original classification. An autoencoder can optionally be used to make the explanations more realistic.

Research

Live demo

Notebook

### TED (Teaching Explanations for Decisions)

TED is most suited for use cases where matching explanations to the mental model of the explanation consumer is the highest priority; i.e., where the explanations are similar to what would be produced by a domain expert.

To achieve this goal, the TED framework requires that the training data is augmented so that each instance contains an explanation (E). The goal is to teach the framework what are appropriate explanations in the same manner the training dataset teaches what are appropriate labels (Y). Thus, the training dataset contains the usual features (X) and labels (Y), augmented with an explanation (E) for each instance. For example, consider a loan application use case, where the features are the loan application answers, and the label is the decision to approve or reject the loan. The explanation would be the reason for the approve/reject decision.

Research ↗

Notebook ↗

### LIME (Local Interpretable Model-agnostic Explanations) ↗

LIME is model-agnostic, meaning that it can be applied to any machine learning model. The technique attempts to understand the model by perturbing the input of data samples and understanding how the predictions change.

Research ↗

Notebook ↗

Tutorials ↗

### SHAP (SHapley Additive exPlanations) ↗

SHAP is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions.

Research ↗

Notebook ↗

Tutorial ↗

Source ↗

### PyCEBox

A Python implementation of individual conditional expectation plots inspired by R's ICEbox. Individual conditional expectation plots were introduced in Peeking Inside the Black Box: Visualising Statistical Learning with Plots of Individual Conditional Expectation (arXiv:1309.6392).

Source ↗

### Skater

Skater is a unified framework to enable Model Interpretation for all forms of model to help one build an Interpretable machine learning system often needed for real world use-cases (they are actively working towards to enabling faithful interpretability for all forms models). It is an open source python library designed to demystify the learned structures of a black box model both globally(inference on the basis of a complete data set) and locally (inference about an individual prediction).

The project was started as a research idea to find ways to enable better interpretability(preferably human interpretability) to predictive "black boxes" both for researchers and practitioners. The project is still in **beta phase**.

Source

### Anchor

This repository has code for the paper High-Precision Model-Agnostic Explanations.

An anchor explanation is a rule that sufficiently “anchors” the prediction locally – such that changes to the rest of the feature values of the instance do not matter. In other words, for instances on which the anchor holds, the prediction is usually the same.
At the moment, they support explaining individual predictions for text classifiers or classifiers that act on tables (numpy arrays of numerical or categorical data).

The anchor method is able to explain any black box classifier, with two or more classes. All it requires is that the classifier implements a function that takes in raw text or a numpy array, and outputs a prediction (integer).

Source

## Resources

IBM Research Trusted AI - explainability
Explaining Explanations: An Overview of Interpretability of Machine Learning
Explainable Machine Learning for Scientific Insights and Discoveries
Explanatory Model Analysis
Interpretable Machine Learning