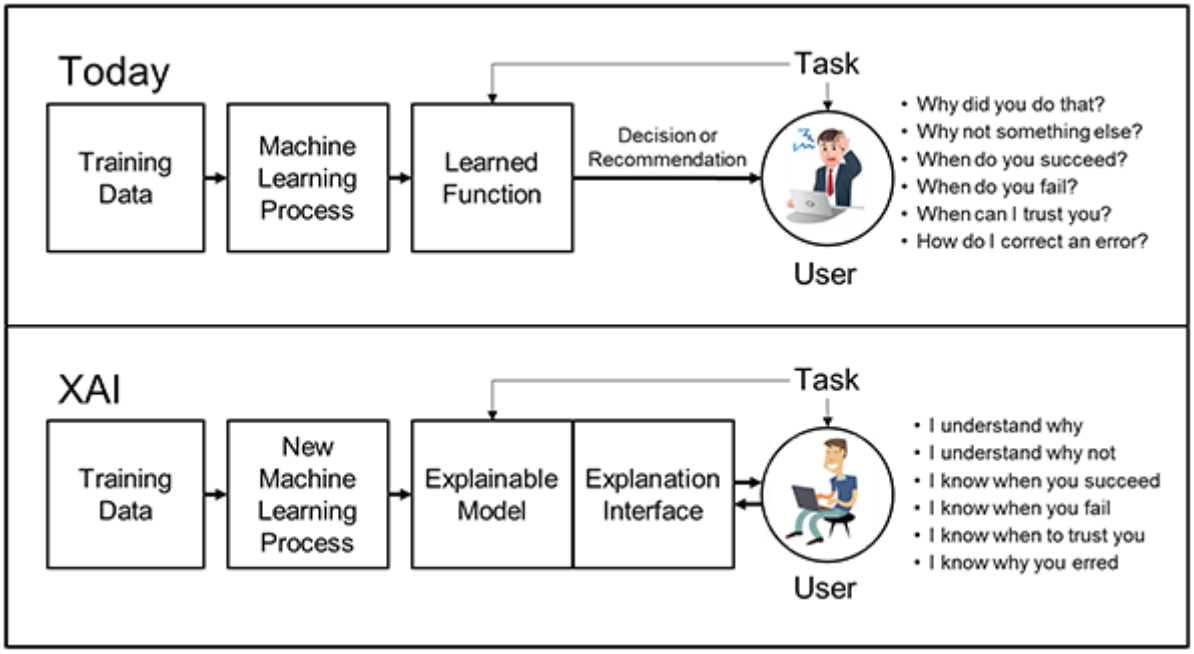
What is Explainable AI?

Artificial intelligence is becoming more complex and increasingly implemented across society, which makes explainability even more crucial.

IBM provides a simple but effective definition for XAI:

> ***"Explainable artificial intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms"***

XAI helps describe an AI model, its expected impact and potential biases. All of this leads to better model accuracy, fairness, transparency and outcomes when AI is used for data-driven decision making.

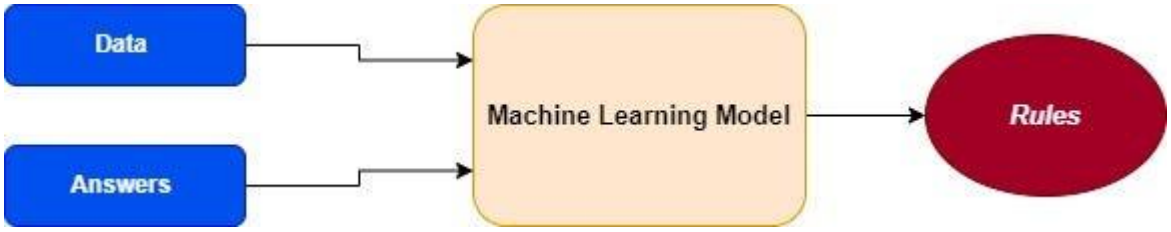


Explainability is critical as AI algorithms take control of more applications and sectors, which brings along the risk of bias, faulty algorithms, and various other issues. By ensuring transparency for your company through explainability, you can truly leverage the power of AI.

Explainable AI is not just one single tool but rather a set of tools and frameworks that help you, your company and the public understand and interpret predictions made by machine learning models.

AI models based decisions are still not mathematically completely explainable, there is still not enough explicit declarative knowledge,

for example Neural Networks model treats high dimensional vectors making them unintelligible to humans. Supervised machine learning algorithm takes an input X and Output Y, decisions are made through input output mapping. These models come up with the rules which are used for future prediction. Take a look at at them



Machine Learning

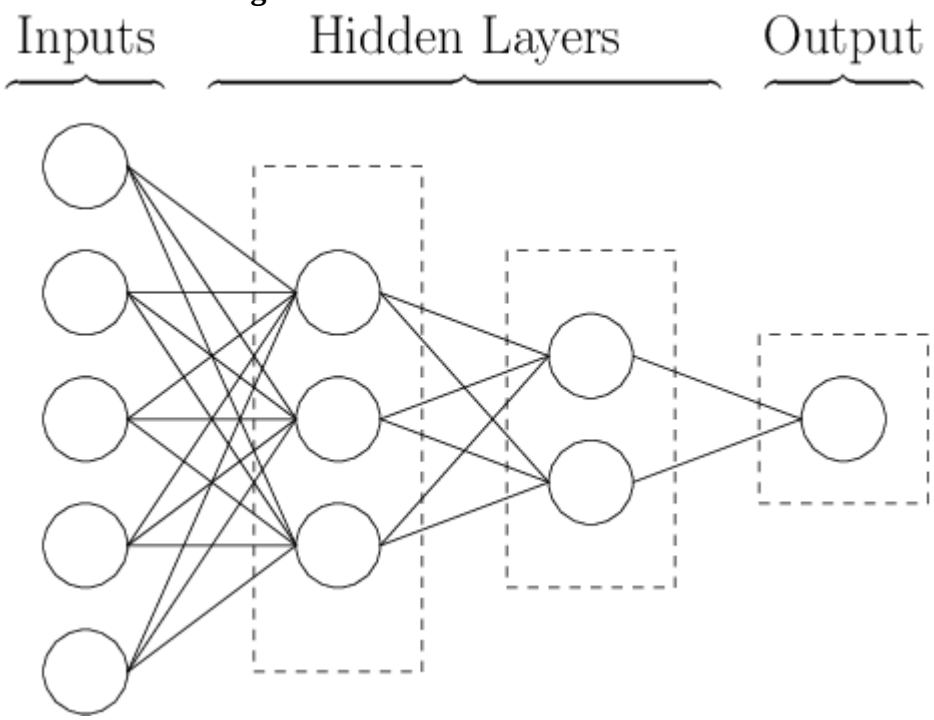


Fig. Simple neural Network

In deep neural networks there are several hidden nodes, layers, and weights, there weights are continuously changing until high performance score is achieved by comparing the model generated outputs with the original outputs.

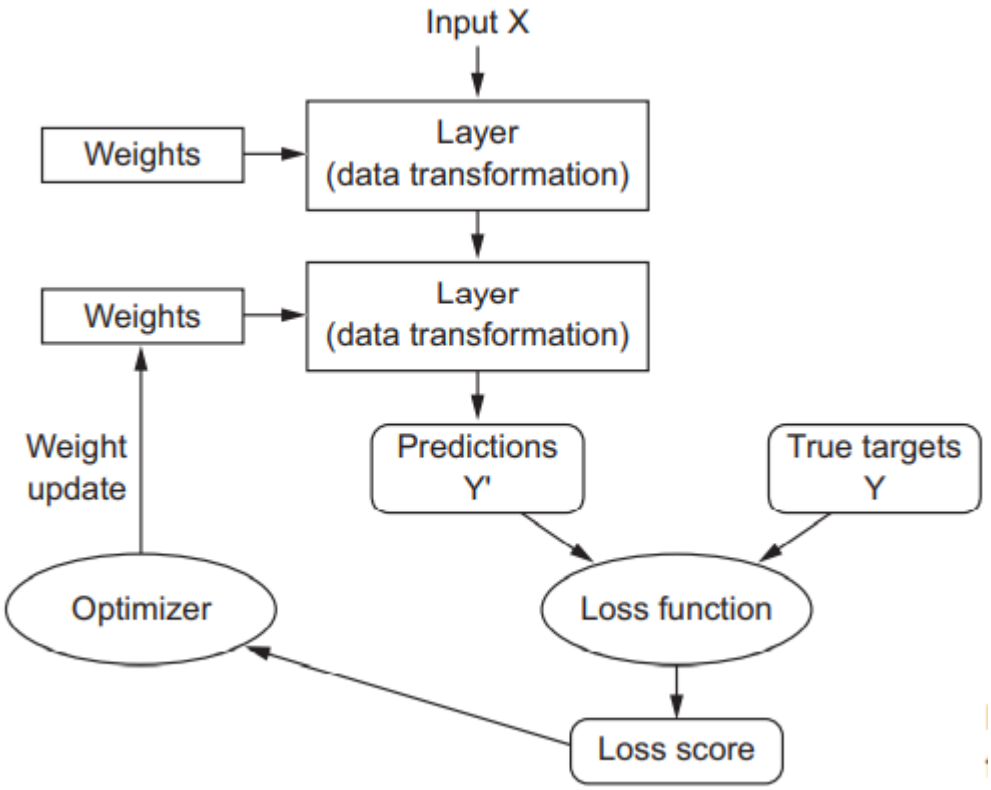


Fig : Neural network : Loss Score is used as feedback Signal to adjust the weights

These machine learning and deep learning models show lack of transparency in their decision making specially on high dimensional inputs, the problem is that there is no explainability of the decision made, Such black box models cannot tells why they made that decision , its upon user whether to take it or leave it. But in certain cases user or more curios about the reason why this decision is made. for example , where a decision can save or cost someone's life, here explainability will play a role, it will be more reliable decision if model can generate satisfactory explanations about the decision made, unless it will be hard to trust AI in such case.

An explanation is the **answer to a why-question** (Miller 2017)

- Why did not the treatment work on the patient?
- Why was my loan rejected?
- Why this email is mark as spammed?

How explainable AI works

With explainable AI – as well as interpretable machine learning – organizations can gain access to AI technology's underlying decision-making and are empowered to make adjustments. Explainable AI can improve the user experience of a product or service by helping the end user trust that the AI is making good decisions. When do AI systems give enough confidence in the decision that you can trust it, and how can the AI system correct errors that arise?[4]

As AI becomes more advanced, ML processes still need to be understood and controlled to ensure AI model results are accurate. Let's look at the difference between AI and XAI, the methods and techniques used to turn AI to XAI, and the difference between interpreting and explaining AI processes

Comparing AI and XAI

What exactly is the difference between "regular" AI and explainable AI? XAI implements specific techniques and methods to ensure that each decision made during the ML process can be traced and explained. AI, on the other hand, often arrives at a result using an ML algorithm, but the architects of the AI systems do not fully understand how the algorithm reached that result. This makes it hard to check for accuracy and leads to loss of control, accountability and auditability.

NEED OF XAI

- **For the Sake of Social Responsibility, Fairness and Risk Avoidance.**
  Especially, within healthcare, clinical and justice work, risks and responsibility are a major concern, as they are potentially dealing with human lives and not merely cost-benefit analyses . Risk avoidance occurs as responsibility is assigned to the individual professional. Hence, developing mental models for expert (e.g. clinical) reasoning to develop better understanding of the reasoning behind deep neural networks and opaque models].

- **Generate Accountable, Reliable and Sound Models for Justification.**
  A theme that has caused great attraction towards xAI is the possibility to ensure fairness and unbiased models by auditing them or create proof of their rightfulness. this approach and argue that xAI provide the required results for auditing the algorithms and generates a provable way for defending algorithmic decisions as being fair and ethical. Hence, generating algorithms that are not only fair and socially responsible, but also accountable and able to justify their output is another aspect motivating the need for xAI.

- **Minimize Biases and Misinterpretation in Model Performance and Interpretation**. Biases in models and their performance have shown to be an important driver for xAI, as media coverage of models performing sub-par to humans in e.g. filtering out appropriate candidates in hiring processes or failing at recognizing people of color. Especially when dealing with neural network learning patterns from training data, biased training data becomes an issue that impacts the validity of the model output .

Defining Interpretability, Explainability

Interpretability:

We consider a model *intrinsically interpretable*, if a human can understand the internal workings of the model, either the entire model at once or at least the parts of the model relevant for a given prediction. This may include understanding decision rules and cutoffs and the ability to manually derive the outputs of the model. For example, the scorecard for the recidivism model can be considered interpretable, as it is compact and simple enough to be fully understood. Ideally, we even understand the learning algorithm well enough to understand how the model's decision boundaries were derived from the training data — that is, we may not only understand a model's rules, but also why the model has these rules.

Explainability:

We consider a *model* explainable if we find a mechanism to provide (partial) information about the workings of the model, such as identifying influential features. We consider a model's *prediction* explainable if a mechanism can provide (partial) information about the prediction, such as identifying which parts of an input were most important for the resulting prediction or which changes to an input would result in a different prediction. For example, for the proprietary COMPAS model for recidivism prediction, an explanation may indicate that the model heavily relies on the age, but not the gender of the accused; for a single prediction made to assess the recidivism risk of a person, an explanation may indicate that the large number of prior arrests are the main reason behind the high risk score. Explanations can come in many different forms, as text, as visualizations, or as examples. Explanations are usually easy to derive from intrinsically interpretable models, but can be provided also for models of which humans may not understand the internals. Explanations are usually partial in nature and often approximated. The explanations may be divorced from the actual internals used to make a decision; they are often called *post-hoc explanations.*

How to think about explaining machine learning models

Before looking at specific techniques for explaining ML models, it will be helpful to build a vocabulary that will help us think about what we can explain and how we may go about it. This helps us consider what type of explanations we want and what methods are compatible with the model we have trained:

- **Intrinsically explainable** - An intrinsically explainable model is designed to be simple and transparent enough that we can get a sense for how it works by looking at its structure, e.g. simple regression models and small decision trees. These models are **directly interpretable**.
- **Post-hoc explainable** - For more complicated, already trained models, we can use explainability tools (often called interpretability tools) to obtain post-hoc explanations. Explanations of sufficiently complex models such as deep neural networks are always post-hoc explanations as they are not directly interpretable.

The types of 'explanations' *typically* fall into one of two categories:

- **Global explanations** - A global explanation of a ML model details what features are important to the model overall. This can be measured by looking at effect sizes or determining which features have the biggest impact on model accuracy. Global explanations are helpful for guiding policy or finding evidence for, or rejecting a hypothesis that a particular feature is important. Figure (4) shows a visualisation of a global explanation for a wine classification task.

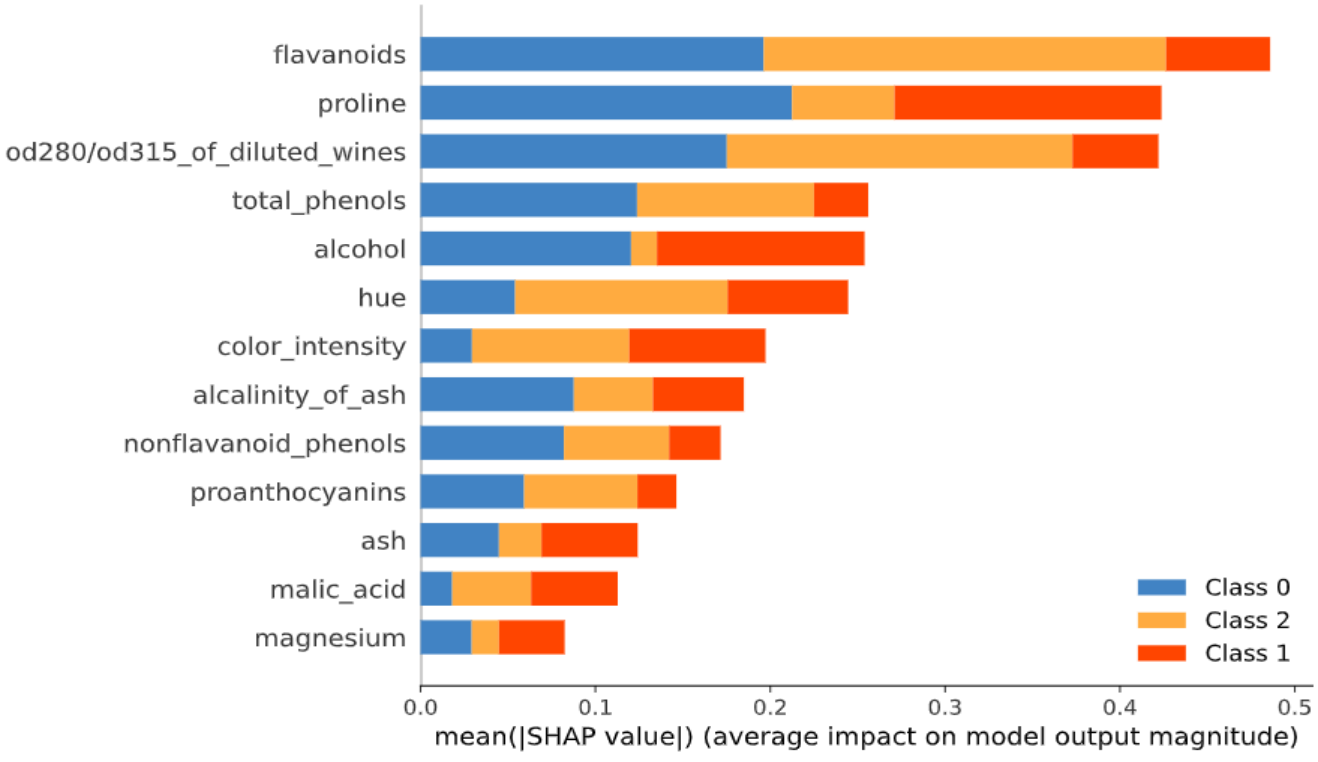


Figure 4. An example of a global explanation for a multi-class classification problem. The size of the horizontal bars indicate how much each feature (on average) influences the classification of a wine into one of three classes.

- **Local explanations** - A local explanation details how a ML model arrived at a specific prediction. For tabular data, it could be a list of features with their impact on the prediction. For a computer vision task, it might be a subset of pixels that had the biggest impact on the classification. Figure (5) shows an example of some local explanations for model predictions for three unique instances. Local explanations are useful for deep-dive insights or diagnosing issues and can provide answers to questions like:
    - Why did the model return this output for this input?
    - What if this feature had a different value?

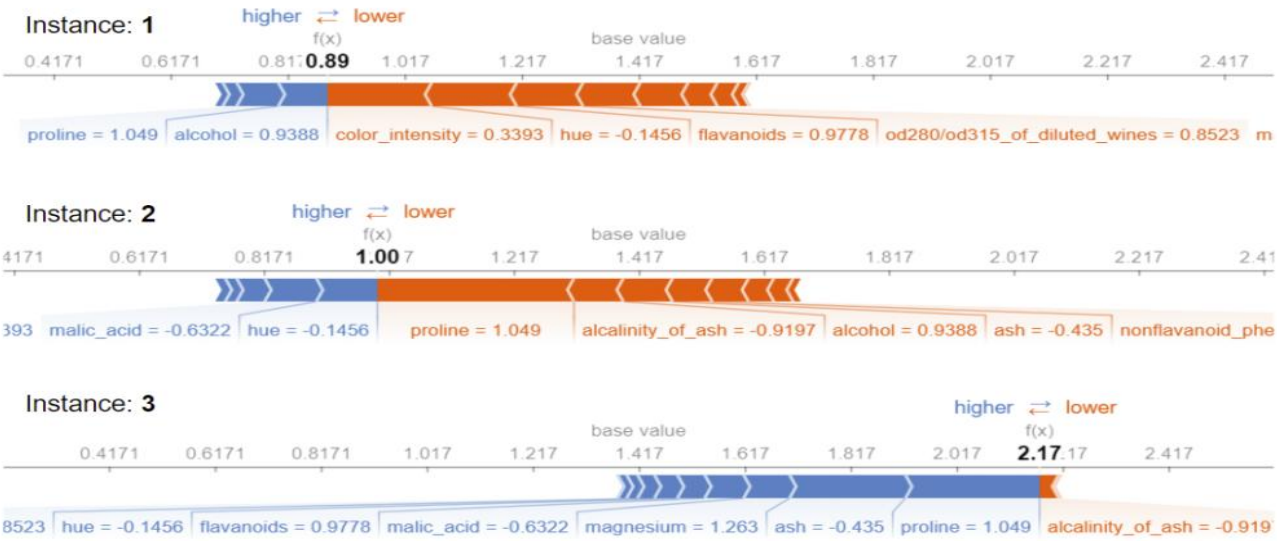


Figure 5. Local explanations of the predictions for three different instances in a wine classification task. The local explanations give the direction and magnitude each feature has on the model output relative to the baseline.

Some methods are a mix of both, providing detailed explanations of how a single feature or interaction of two features impacts a set of predictions. These are **modular global explanations** because they can only be used to inspect the impact of one or two features at a time.

The methods used to provide explanations are either model-specific or model-agnostic:

- **Model-specific** - Model-specific methods work by inspecting or having access to the model internals. Interpreting regression coefficient weights or P-values in a linear model or counting the number of times a feature is used in an ensemble tree model are examples of model-specific methods.
- **Model-agnostic** - Model-agnostic methods work by investigating the relationship between input-output pairs of trained models. They do not depend on the in-

ternal structure of the model. These methods are very useful for when we have no theory or other mechanism to interpret what is happening inside the model.

Properties of Explanations

We want to explain the predictions of a machine learning model. To achieve this, we rely on some explanation method, which is an algorithm that generates explanations. **An explanation usually relates the feature values of an instance to its model prediction in a humanly understandable way.**

Properties of Explanation Methods

- **Expressive Power** is the "language" or structure of the explanations the method is able to generate. An explanation method could generate IF-THEN rules, decision trees, a weighted sum, natural language or something else.
- **Translucency** describes how much the explanation method relies on looking into the machine learning model, like its parameters. For example, explanation methods relying on intrinsically interpretable models like the linear regression model (model-specific) are highly translucent. Methods only relying on manipulating inputs and observing the predictions have zero translucency. Depending on the scenario, different levels of translucency might be desirable. The advantage of high translucency is that the method can rely on more information to generate explanations. The advantage of low translucency is that the explanation method is more portable.
- **Portability** describes the range of machine learning models with which the explanation method can be used. Methods with a low translucency have a higher portability because they treat the machine learning model as a black box. Surrogate models might be the explanation method with the highest portability. Methods that only work for e.g. recurrent neural networks have low portability.
- **Algorithmic Complexity** describes the computational complexity of the method that generates the explanation. This property is important to consider when computation time is a bottleneck in generating explanations.

Properties of Individual Explanations

- **Accuracy**: How well does an explanation predict unseen data? High accuracy is especially important if the explanation is used for predictions in place of the machine learning model. Low accuracy can be fine if the accuracy of the machine learning model is also low, and if the goal is to explain what the black box model does. In this case, only fidelity is important.
- **Fidelity**: How well does the explanation approximate the prediction of the black box model? High fidelity is one of the most important properties of an explanation, because an explanation with low fidelity is useless to explain the machine learning model. Accuracy and fidelity are closely related. If the black box model has high accuracy and the explanation has high fidelity, the explanation also has high accuracy. Some explanations offer only local fidelity, meaning the explanation only ap-

proximates well to the model prediction for a subset of the data (e.g. local surrogate models) or even for only an individual data instance (e.g. Shapley Values).

- **Consistency**: How much does an explanation differ between models that have been trained on the same task and that produce similar predictions? For example, I train a support vector machine and a linear regression model on the same task and both produce very similar predictions. I compute explanations using a method of my choice and analyze how different the explanations are. If the explanations are very similar, the explanations are highly consistent. I find this property somewhat tricky, since the two models could use different features, but get similar predictions (also called "Rashomon Effect"). In this case a high consistency is not desirable because the explanations have to be very different. High consistency is desirable if the models really rely on similar relationships.
- **Stability**: How similar are the explanations for similar instances? While consistency compares explanations between models, stability compares explanations between similar instances for a fixed model. High stability means that slight variations in the features of an instance do not substantially change the explanation (unless these slight variations also strongly change the prediction). A lack of stability can be the result of a high variance of the explanation method. In other words, the explanation method is strongly affected by slight changes of the feature values of the instance to be explained. A lack of stability can also be caused by non-deterministic components of the explanation method, such as a data sampling step, like the local surrogate method uses. High stability is always desirable.
- **Comprehensibility**: How well do humans understand the explanations? This looks just like one more property among many, but it is the elephant in the room. Difficult to define and measure, but extremely important to get right. Many people agree that comprehensibility depends on the audience. Ideas for measuring comprehensibility include measuring the size of the explanation (number of features with a non-zero weight in a linear model, number of decision rules, …) or testing how well people can predict the behavior of the machine learning model from the explanations. The comprehensibility of the features used in the explanation should also be considered. A complex transformation of features might be less comprehensible than the original features.
- **Certainty**: Does the explanation reflect the certainty of the machine learning model? Many machine learning models only give predictions without a statement about the models confidence that the prediction is correct. If the model predicts a 4% probability of cancer for one patient, is it as certain as the 4% probability that another patient, with different feature values, received? An explanation that includes the model's certainty is very useful.
- **Degree of Importance**: How well does the explanation reflect the importance of features or parts of the explanation? For example, if a decision rule is generated as an explanation for an individual prediction, is it clear which of the conditions of the rule was the most important?
- **Novelty**: Does the explanation reflect whether a data instance to be explained comes from a "new" region far removed from the distribution of training data? In such cases, the model may be inaccurate and the explanation may be useless. The

concept of novelty is related to the concept of certainty. The higher the novelty, the more likely it is that the model will have low certainty due to lack of data.

- **Representativeness**: How many instances does an explanation cover? Explanations can cover the entire model (e.g. interpretation of weights in a linear regression model) or represent only an individual prediction (e.g. Shapley Values).