Interpretability techniques: state of the art

"By far the greatest danger of artificial intelligence is that people conclude too soon that they understand it".
Eliezer Yudkowsky[40]

Concept

The scientific community[41,42] has proposed numerous definitions of model "interpretability" and "explainability", and tends to make a certain distinction between them, although in practice these concepts are often used interchangeably. Generally speaking, interpretability is linked to the ability to explain to a human being the results of a model (its cause-effect relationship), while explainability is associated with the understanding of an algorithm's internal logic, how it is designed and trained, and the steps followed in decision making to reach a particular result.

Some academic definitions in this regard are:

- Interpretability is the ability to explain or present in terms that are understandable to a human being[43].
- Interpretability is the degree to which a human being can understand the cause of a decision[44].
- The explainability of a model output is the description of how the output of the model was produced[45].
- Explainability is the extent to which the internal mechanics of a machine learning system can be explained in human terms[46].

The need for model explainability and interpretability has favored the emergence of increasingly sophisticated techniques for local and global interpretability of model results, and there is currently some level of standardization and convergence in the use of certain techniques (e.g. PDP, LIME or SHAP).

At the same time, these techniques do not completely solve the problem of interpretability and may yield contradictory or biased results under certain circumstances, which coexists with other factors that may impact model interpretability, such as:

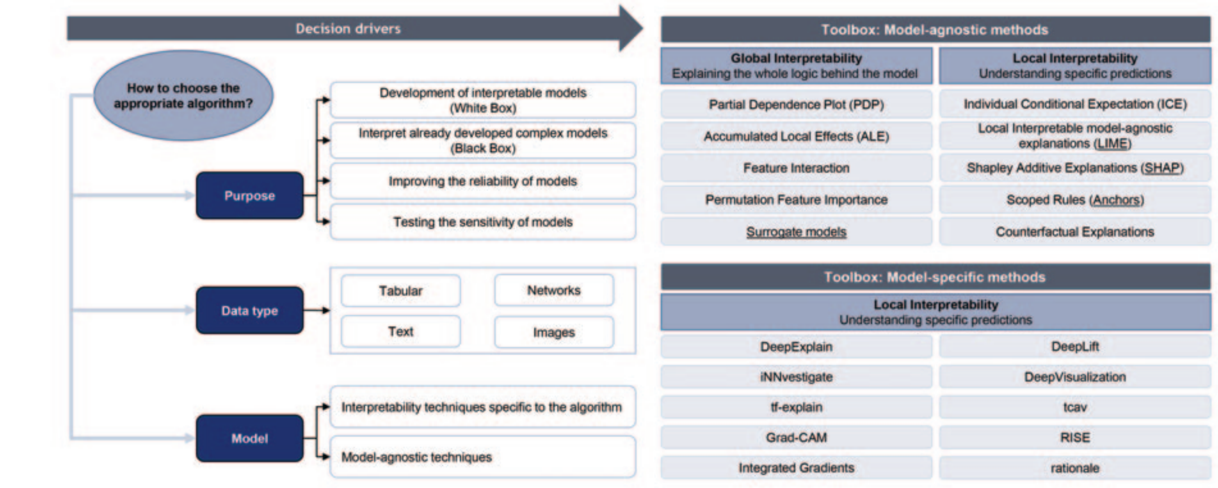
- The reproducibility of results, the model development and implementation process[47], the consistency of the model's predictions and the explanation of the most probable sequence of predictions.
- Potential bias[48] in the input data.
- Fairness[49].
- Accuracy of explanation[50].
- Conceptual soundness of the model[51].

To overcome several of these difficulties, some researchers[52] are developing alternative approaches for improving AI model interpretability, primarily focused on the development of inherently interpretable models ("white boxes").

This section describes the main interpretability techniques, considered standard in the industry, and includes the state of the art on white-box development.

[40] Eliezer Shlomo Yudkowsky (b. 1979), American researcher and writer specializing in decision theory and artificial intelligence, known for popularizing the idea of Friendly Artificial Intelligence and advocating the Singularity.
[41] Gall, R. (2018). Editor at Thoughtworks and The New Stack.
[42] Broniatowsky, D. (2021). Associate Professor, Department of Engineering Management and Systems Engineering, George Washington University.
[43] Doshi-Velez, F., et al. (2017). Professor of Computer Science at the Paulson School of Engineering and Applied Science, Harvard University.
[44] Miller, T. (2019). Lecturer in the School of Computing and Information Systems, University of Melbourne.
[45] Broniatowsky D. (2021).
[46] Gall, R. (2018).
[47] Leventi-Peetz, A.-M., et al. (2022). Scientist of the German Federal Office for Information Security.
[48] Zhou, N., et al. (2021). Senior financial analyst at Wells Fargo.
[49] Ibid.
[50] Jonathon Phillips et al. (2021). Professor of Computer Science and Engineering, National Institute of Standards and Technology (NIST).
[51] Sudjianto, A., et al. (2021).
[52] Ibid.

Figure 6. Overview of interpretability techniques.



Decision drivers
How to choose the appropriate algorithm?
Purpose
Development of interpretable models (White Box)
Interpret already developed complex models (Black Box)
Improving the reliability of models
Testing the sensitivity of models
Data type
Tabular
Networks
Text
Images
Model
Interpretability techniques specific to the algorithm
Model-agnostic techniques
Toolbox: Model-agnostic methods
Global Interpretability
Explaining the whole logic behind the model
Local Interpretability
Understanding specific predictions
Partial Dependence Plot (PDP)
Individual Conditional Expectation (ICE)
Accumulated Local Effects (ALE)
Local Interpretable model-agnostic explanations (LIME)
Feature Interaction
Shapley Additive Explanations (SHAP)
Permutation Feature Importance
Scoped Rules (Anchors)
Surrogate models
Counterfactual Explanations
Toolbox: Model-specific methods
Local Interpretability
Understanding specific predictions
DeepExplain
DeepLift
iNNvestigate
DeepVisualization
tf-explain
tcav
Grad-CAM
RISE
Integrated Gradients
rationale

Most common interpretability techniques

The most commonly used interpretability techniques can be grouped according to their approach[53]: post-hoc interpretability and inherently interpretable models. There are also complementary strategies to improve model understanding.

Post-hoc interpretability

Post-hoc interpretability or black box model interpretability techniques focus on explaining the output of already trained models, based on the information provided by the weights assigned to each input variable and the model results. These techniques are useful for understanding model results, although they do not provide information about the training process or explain the internal logic of the algorithm.

They are usually divided into global and local interpretability techniques, in reference to whether the technique explains the entire model as a whole or only the results in a subset of observations or data.

The most common post-hoc interpretability techniques are as follows (for a more comprehensive inventory, see Fig. 6):

- PDP (Partial Dependence Plots). This technique allows visualizing the influence of each individual variable on the model output, excluding the rest of the variables.
- LIME (Local Interpretable Model-agnostic Explanations). This technique allows the explanation of results at the local level, i.e. the explanation of the results of a particular specific observation, based on information from other similar cases.
- SHAP (SHapley Additive exPlanations). This technique allows the local and global explanation of a model's results, that is, the explanation of the influence of each variable on model observations, and the importance of each variable in the model's global results.
- Anchors. This involves the search for decision rules that explain the result.

Inherently interpretable models

Inherent interpretability focuses on the development of "white box" models that are interpretable by design or that can be made interpretable by construction, through a series of conditions dependent on the type of model (e.g. neural networks[54], in particular ReLu[55], and tree-based models[56], among others).

These models allow an explanation of the algorithm's internal logic and the sequence of steps taken to reach a specific result, and therefore allow a better understanding of the results, although their applicability in complex problems may be more limited, depending on the type of algorithm used.

Complementary strategies

Some strategies are used to support model interpretability, such as simplifying the model to facilitate its interpretation, using "business sense" variables, analyzing data to identify biases or lack of fairness in the inputs that may hinder explainability, or analyzing model development or model implementation reproducibility.

[53] iDanae (2022).
[54] Yang, Z., et al. (2019). Department of Statistics and Actuarial Science, University of Hong Kong.
[55] Sudjianto. A., et al. (2011).
[56] Sudjianto. A., et al. (2021).

Post-hoc interpretability

1. PDP

PDP plots[57] (Partial Dependence Plots) show how an AI model's prediction varies as a function of one or two independent variables in the prediction, i.e. the marginal effect of the predictors. Thus, they make it possible to evaluate the relationship between the independent and dependent variables.

Synthetically:

- PDPs show the average variation of the prediction graphically on a curve.
- This average variance is obtained by varying a predictor for all the observations in the dataset, and then obtaining the average impact on the prediction.
- A variant of the PDPs are the Individual Conditional Expectation (ICE) graphs, which similarly show how a prediction varies for each specific observation if one of the model's predictors is modified while keeping the rest constant.

2. LIME

LIME[59] (Local Interpretable Model-agnostic Explanations) is a local method that tests how the predictions of a model vary when the input data are perturbed. To do this, LIME applies the following steps:

- Generate synthetic data around an observation in the input data: LIME takes as a starting point a single prediction and the input data that generated it, and generates new input data by perturbing this observation, obtaining the corresponding predictions by the AI model.
- Train a simple model on synthetic data: the resulting dataset composed of the perturbed input data and the predictions generated by the model is used to train a model that is interpretable (e.g. linear models, decision trees).
- Explain the predictions of the simple model as a function of the original data: the importance of each variable in the prediction is obtained - for example, as a function of its coefficients in the regression and its corresponding sign.

Use Case: Loan origination in the banking sector. Use of PDPs

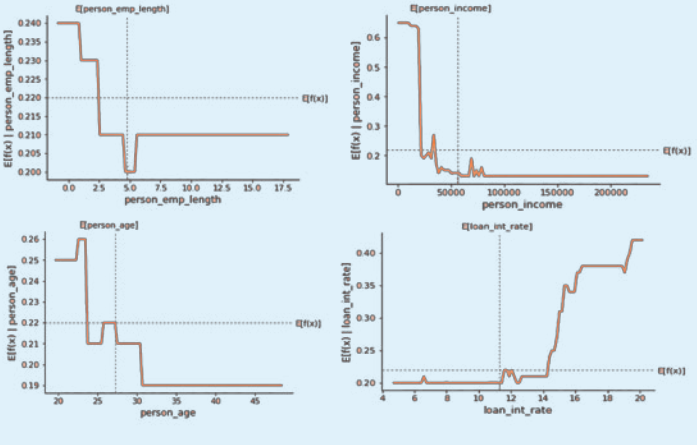
PDPs can be applied to a very common use case in the banking industry: rating customers during the lending process to determine their probability of default. This example uses an anonymized portfolio of mortgage loans with information on their performance in the first three years.

An XGBoost was used, which is a non-additive tree-based model, a feature that may make it difficult to explain. The variables employed by the model during training include the loan amount, its purpose, the borrower's ownership status, years of employment in his current job, and the interest rate, among others.

In this context, a business area may seek to understand why the model assigns a certain probability of default to a certain customer.

A PDP graph shows the explanation that would be obtained at the global level of the variables that have most participated in the result, and that would allow us to see the impact that different ranges of that variable have on the model's prediction (Fig. 7).

Figure 7. PDP for the variables "years employed" (in years), "salary" (annual EUR), "age" (years) and "interest rate" (times one). The X-axis represents the variable under study itself, and the Y-axis represents the impact that different ranges of each variable have on the model's prediction.



[57] Friedman, J. H. (2001). Professor in the Department of Statistics, Stanford University.

[58] Goldstein, A., et al. (2015). Professor in the Department of Statistics, The Wharton School, University of Pennsylvania.

[59] Ribeiro, M. T., et al. (2016). Researcher at Microsoft Research in the Adaptive Systems and Interaction group and Adjunct Professor at the University of Washington.

- Calculate the explainability: the percentage of explainability by LIME is equivalent to the linear model fit coefficient (e.g., R2). It follows that the interpretable model yields a good approximation of the predictions locally.

Formally, an explanation using local subrogated models with LIME can be defined as:

$$Explanation(X) = \arg\min_{g\in G} L(f, g, \pi_X) + \Omega(g)$$

where:

f is a black box model (e.g. a random forest), g is the model that explains f (e.g. a linear regression).

L is the loss function to be minimized in the model (e.g. mean square error), which LIME minimizes.

Ω is the model's complexity (e.g. number of variables selected) decided by the user.

G is the set of possible explanations of the model f.

$\arg$ min represents the value $g \in G$ that minimizes the function $L(f,g,\pi_X) + \Omega(g)$..

π_X represents the amplitude of the perturbations used to generate new observations decided by the user.

3. SHAP

SHAP[60] (SHapley Additive exPlanations) is a model explanation method based on Shapley's Value Theorem , which was proposed in 1952 to distribute the value of a game among the players. SHAP is used to explain the importance of each variable (measured as the average change in the model prediction when the value of the variable varies) in a particular prediction.

Specifically, SHAP uses a combination of baselines, local importance functions and Shapley's Value Theorem to calculate the importance of each variable in an individual prediction.

In this method:

- Shapley values are calculated, where the independent variables are interpreted as players who collaborate to receive the payout.
- The Shapley values correspond to the contribution of each variable to the model prediction.
- The payout is the actual prediction made by the model minus the average value of all predictions.
- Players "split" this payout according to their contribution, and this split is calculated by Shapley's values and reflects the importance of each variable.

This method also makes it possible to obtain interpretations at a global level by calculating the average of the contributions of each variable for each model prediction.

Formally, Shapley values can be defined as the contribution of each variable to the outcome of the model, weighed as a function of all possible combinations of variables used:

$$\phi_j(val) = \sum_{S \subseteq \{1,\dots,p\}/\{j\}} \frac{|S|!(p-|S|-1)!}{p!}(val(S \cup \{j\}) - val(S))$$

where val is the prediction of the model for variables included in the set S, with respect to the prediction for variables not included in S:

$$val = \int f(x_1 \dots x_p)\, dP_{x\notin S} - E_X(f(X))$$

where:

x is the vector of variables used in the model.

S is a subset of x.

p is the number of variables used in the model.

$dP_{(x\notin S)}$ represents the set of variables not included in S for which the integration is performed.

E is the expected value of the prediction of X with the f model.

Using these values, SHAP can be used to obtain a local explanation to the model as:

$$Expl(x) = E_X(f(X)) + \sum \phi_j x_j$$

Finally, SHAP is also capable of calculating global explanations through the aggregation of Shapley values in a data set.

[60] Lundberg, S. M., et al. (2017). Research Fellow at the Paul G. Allen School of Computer Science, University of Washington.
[61] Shapley, L. (1953). Professor at the University of California, Los Angeles, in the departments of Mathematics and Economics.
[62] Ribeiro, M. T.; Singh, S.; Guestrin, C. (2018). Researcher at Microsoft Research in the Adaptive Systems and Interaction group and Adjunct Professor at the University of Washington.

4. Anchor

Anchors[62] is a method that explains individual (i.e. local) predictions of black box classification models by finding decision rules called "anchors" that explain the outcome.

- As in LIME, a single prediction and the input data that generated it are taken as a starting point, and new input data are generated by perturbing this observation, obtaining the corresponding predictions by the AI model.

- The local explanation of the prediction is obtained by looking for "if-else" rules that are able to explain the outcome of the model. A rule is considered to explain the prediction if changes in other independent variables not considered in the rule do not modify it.

Formally, an anchor A is defined as:

where:

f is a black box model.

D is an arbitrary distribution used to pertub x.

x is an observation of the dataset to be explained, and z is a sample of D.

$Prec$ is the accuracy of the explanation and τ is the accuracy required..

One way to find an anchor given any given distribution D is to look for the precision to exceed a threshold with a certain probability (1 - δ), such that:

$$P(Prec\ (A) \geq \tau) \geq 1 - \delta$$

[63]Yang, Z., et al. (2019). Research Fellow in the Department of Statistics and Actuarial Science, University of Hong Kong.

Use case: Loan origination in the banking sector. Use of SHAP

If SHAP is applied on the same case for which a PDP was used, additional local information about a decision of the model is obtained for a given customer.

In this case, using SHAP on a sample of observations results in completely different Shapley values with a variable sign depending on the characteristics of the borrower. Even for clients receiving the same interest rate, the influence of this variable appears to vary due to the greater or lesser importance of the other variables in the model.

However, a "business sense" trend is observed: the higher the interest rate, the more this variable in the model contributes to a higher probability of default. Therefore, using the mean of the Shapley values for each variable to provide an overall interpretation of the model can lead to errors in the explanation if this is understood as a generalization (Fig. 8).

Shapley's values provide an explanation for particular cases such as the following, where it is observed that the probability of default of a client is determined by the mortgage loan conditions, credit history and employment conditions (e.g., salary) (Fig. 9).

Figure 8. Shapley values for the "interest rate" variable in the whole sample versus that variable. The gray bar graph shows the distribution of the variable.

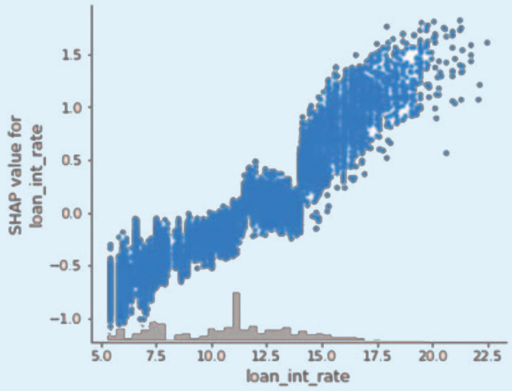
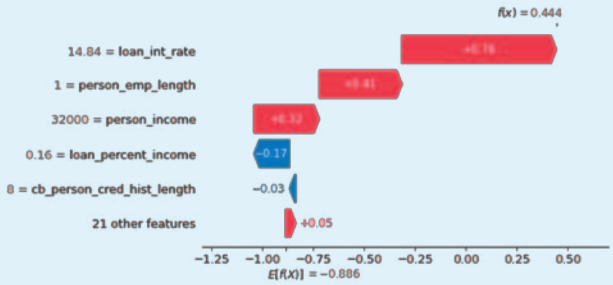
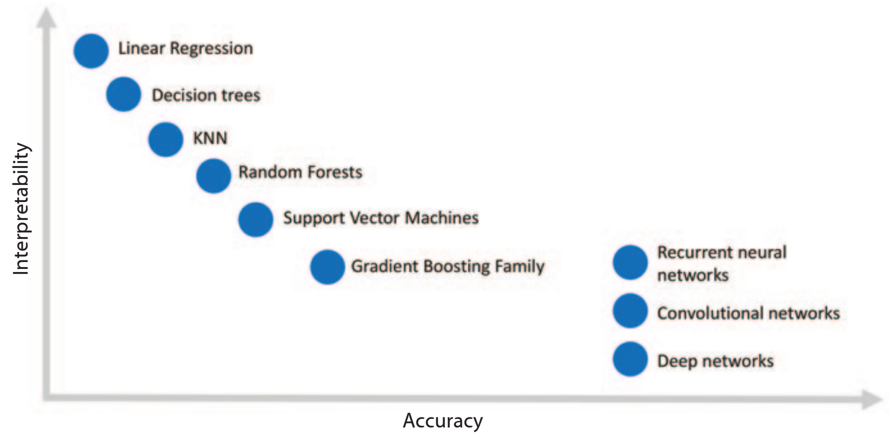


Figure 9. Shapley values influencing the prediction of a client with a denied loan[2].



[1]Scale of the graph shown in log-odds (0 corresponds to a 50% probability).
[2]Log-odds scale graph.

Figure 10. Balance between interpretability and predictive capacity by model families (including white and black boxes).



Development of inherently interpretable models (white box)

Inherently interpretable (white box) models are based on the design of algorithms that, by design, are interpretable and allow the explanation of results at both the global and local levels.

White box models are generally grouped according to the type of algorithm used:

- Linear models, such as linear or logistic regressions.
- Tree-based models, such as decision trees or random trees.
- Rule-based models, such as rule-based systems.
- Deep neural networks, with activation functions such as ReLU or the use of intermediate layers, subject to certain restrictions that make them inherently interpretable[63].

These models are usually developed with constraints on the parameters to be optimized, which allow the models to be interpretable unlike black box models, although they are less accurate (Fig. 10). These constraints include using only "business sense" variables, or restricting:

- The number of variables selected by the model for prediction.
- The number of variables explained by the model.
- The degree of complexity of the decision rules.
- The number of steps in the prediction.
- The depth of the decision trees.
- The length and depth of the neural networks.

Inherently interpretable models provide more accurate results, as they allow for a better understanding of the information, which in turn leads to better decision making. This is especially necessary in those sectors where interpretability is a critical factor in final decisions.

Two aspects relevant to the construction of inherently interpretable models are detailed below: the concept and development of interpretable supervised and unsupervised learning, and the application of other factors in the interpretability domain.

1. Interpretable supervised and unsupervised learning

Although current research is moving towards the development of inherently interpretable models, there is no mathematical formalism that fully describes the construction of these models under whatever initial conditions and algorithms used.

The state of the art is the construction of these models under initial conditions that make them more easily interpretable or equivalent to other interpretable models. One of the ways to define this interpretability condition in model training is to modify the loss[66] function that is minimized during its training, including a penalty for low interpretability, which depends on an imposed model interpretability condition f:

$$Min\left(\frac{1}{n}\sum Loss(f, z_i) + C \cdot InterpretabilityPenalty(f)\right)$$

[64] Rudin, C., et al. (2022). Professor of Computer Science, ECE, Statistics and Biostatistics and Bioinformatics at Duke University.

For example, sparsity is one of the conditions used in model development to qualify a model as more explainable with respect to the rest. This condition can be added to the loss function as:

$$Min\left(\frac{1}{n}\sum Loss(f, z_i) + \varphi(f)\right)$$

such that φ(f) is a regularization function that penalizes the loss being proportional to the sparsity of the model (e.g. if the sparsity is reduced, that term of the loss function will also be reduced).

Some authors[67] have formalized the creation of inherently interpretable models for certain families as: models based on decision trees (e.g. SIMTree or single-index model tree, which generates a single-index model tree for each terminal node), or the simplification of networks with ReLu activation function, which are shown to be equivalent to a set of local linear models.

Other authors[68] have focused on defining the characteristics that inherently interpretable models should meet, in order to optimize them during the modeling process, such as:

- Additivity of the input variables, so that their effects are aggregated in the model in a simple way.
- Sparsity, and the optimization of models to meet this condition.
- Linearity of input variables versus model output.
- Monotonicity, so that the relationship between the input variable and the outcome to be predicted is monotonic for as many ranges as possible.
- Decoupling of concepts during the neural network training, which refers to maintaining as much as possible the information about a given concept in specific network paths (i.e. in the face of information about the same concept passing through a greater number of neurons and paths dispersed in the network).
- Dimensionality reduction as a visual tool to facilitate post-hoc explanations to humans.

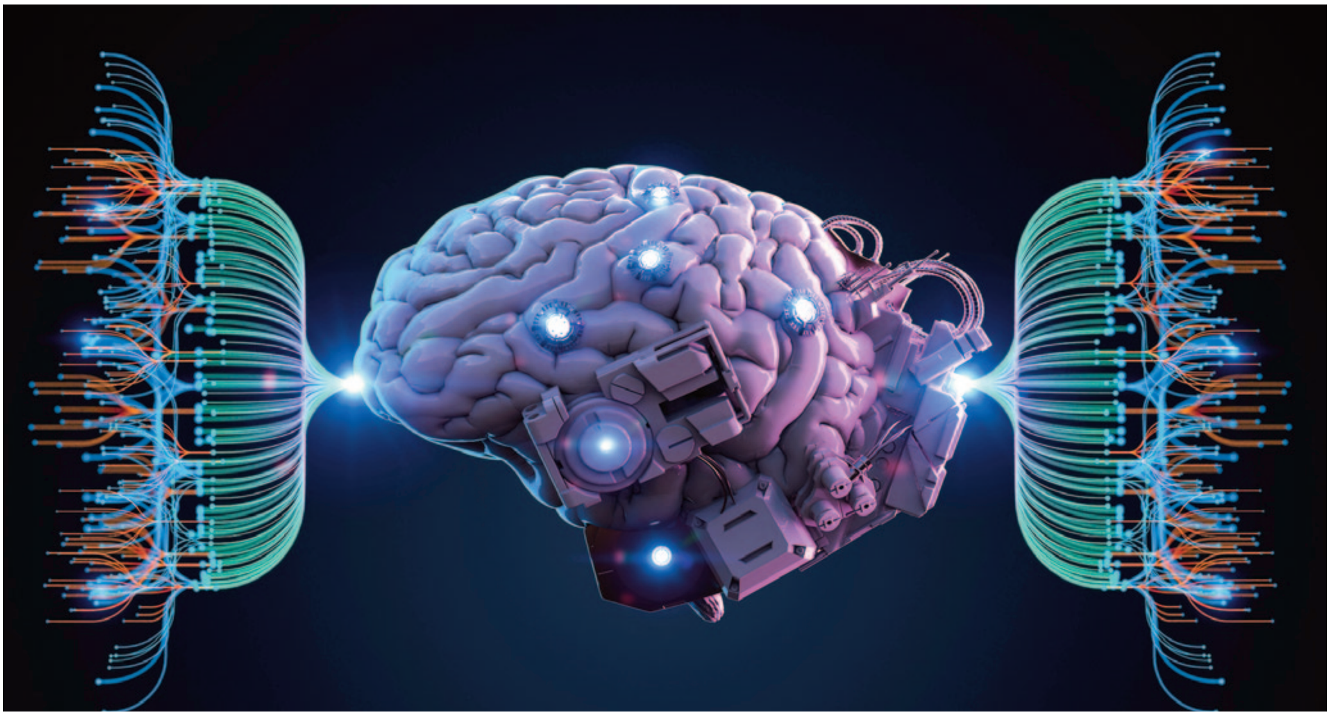
2. Other impact factors

In combination with the challenges shown in this section, there are additional key elements that can be considered to improve model interpretability, such as model fairness, absence of bias in the input data, potential expert components, or adequate performance and model control framework to avoid errors in model interpretation.

Because of their relevance, as indicated above[69], these elements have also been highlighted in the AI Act as essential requirements for high-risk AI systems.

Nowadays, there are multiple techniques and methods to evaluate model performance, and to prevent overfitting issues. There are also several ways to evaluate the error produced by models and the balance between bias and variance error.

[65] Sudjianto. A., et al. (2021).
[66] Rudin, C., et al. (2022).
[67] See section on regulation.



However, due to constraints on the use of personal data introduced by data protection regulations, one of the greatest complexities at the moment is in detecting and correcting potential biases (e.g. due to race, gender, religion, political or sexual orientation, beliefs or social position) in AI models, especially when the variables have not been stored and are therefore not available for analysis.

In this regard, several techniques for identifying unbiased input variables have been proposed by academia, such as:

- Interpretability analysis through Causal Bayesian Networks[68] as a quantification of the degree of model fairness.
- Definition[69] of fairness metrics, such as demographic parity, predictive ratio parity, false positives and equal false negatives in segments susceptible to bias.

Among these metrics, counterfactual fairness provides a measure of how similar the results of a model are to individuals (observations) with the same characteristics, but with slightly different bias-sensitive attributes.

Advantages and disadvantages of the most common interpretability techniques

As a general rule, there is no interpretability technique that can provide a single, global and intuitive explanation for any scenario. Interpretability techniques are usually combined under various use cases and scenarios to verify that they provide reproducible explanations applicable to different groups of observations.

When selecting which of these techniques to use, it is advisable to consider the advantages or disadvantages of their implementation (Fig. 11).

Latest trends and challenges

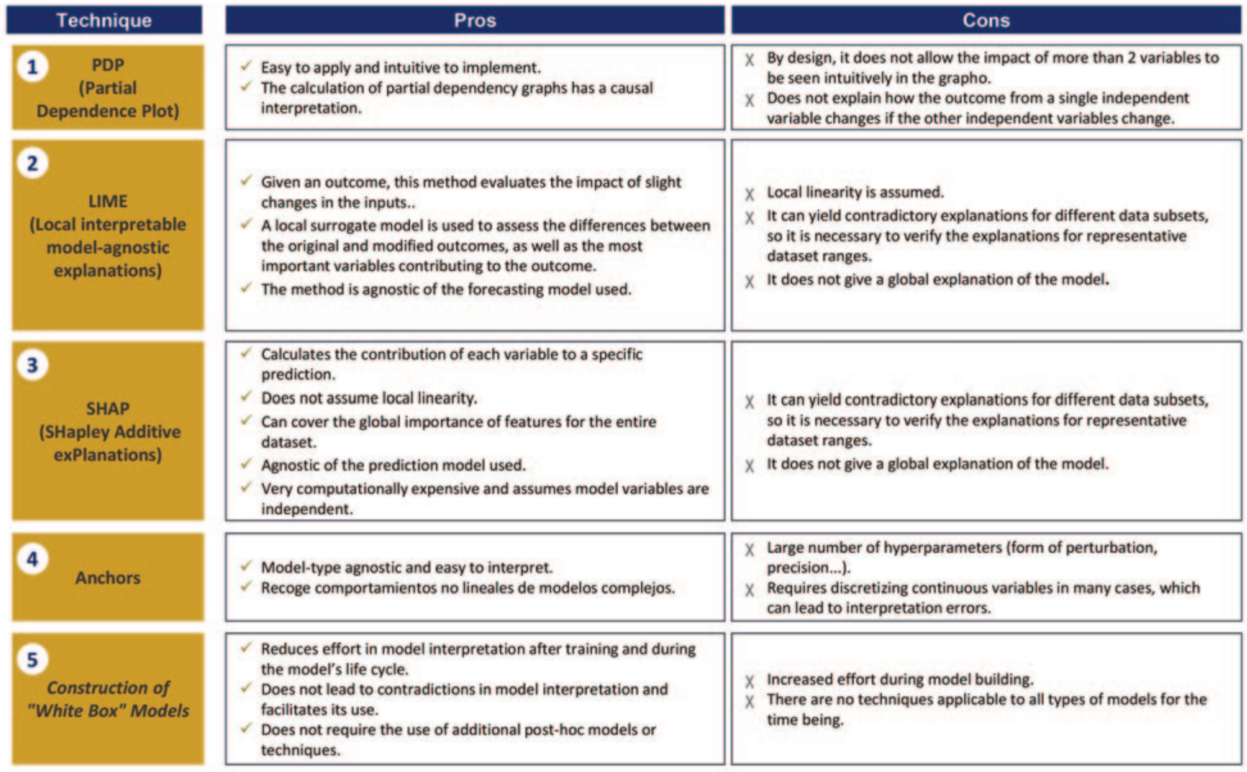
Despite advances in model interpretability, there are still challenges in explaining the results (Fig. 12).

First, model interpretability is still constrained by a number of factors such as the reproducibility of the results[70], the model training and implementation process, the consistency of model predictions, the explanation of the sequence of most likely predictions, the biases in the input data, as well as the fairness and accuracy of the explanation.

Secondly, currently available XAI techniques only allow either local explanations (i.e. for a single observation or data) or global explanations (i.e. for the whole data set). This means there is a need to develop techniques that allow midrange explanations, i.e. explaining results for groups or subsets of data in a consistent manner . In addition, without an in-depth analysis,

[68] Oneto, L.,Chiappa, S., (2020)
[69] Zhou, N., et al. (2021). Senior financial analyst at Wells Fargo.
[70] Leventi-Peetz, A.-M., et al. (2022).
[71] While SHAP is able to obtain explanations for subsets through weighted averages of Shapley values, these explanations may vary depending on the granularity of the subset data.

Figure 11. Comparison of the most common interpretability techniques.



Technique	Pros	Cons
1 PDP (Partial Dependence Plot)	✓ Easy to apply and intuitive to implement. ✓ The calculation of partial dependency graphs has a causal interpretation.	X By design, it does not allow the impact of more than 2 variables to be seen intuitively in the grapho. X Does not explain how the outcome from a single independent variable changes if the other independent variables change.
2 LIME (Local interpretable model-agnostic explanations)	✓ Given an outcome, this method evaluates the impact of slight changes in the inputs.. ✓ A local surrogate model is used to assess the differences between the original and modified outcomes, as well as the most important variables contributing to the outcome. ✓ The method is agnostic of the forecasting model used.	X Local linearity is assumed. X It can yield contradictory explanations for different data subsets, so it is necessary to verify the explanations for representative dataset ranges. X It does not give a global explanation of the model.
3 SHAP (SHapley Additive exPlanations)	✓ Calculates the contribution of each variable to a specific prediction. ✓ Does not assume local linearity. ✓ Can cover the global importance of features for the entire dataset. ✓ Agnostic of the prediction model used. ✓ Very computationally expensive and assumes model variables are independent.	X It can yield contradictory explanations for different data subsets, so it is necessary to verify the explanations for representative dataset ranges. X It does not give a global explanation of the model.
4 Anchors	✓ Model-type agnostic and easy to interpret. ✓ Recoge comportamientos no lineales de modelos complejos.	X Large number of hyperparameters (form of perturbation, precision...). X Requires discretizing continuous variables in many cases, which can lead to interpretation errors.
5 *Construction of "White Box" Models*	✓ Reduces effort in model interpretation after training and during the model's life cycle. ✓ Does not lead to contradictions in model interpretation and facilitates its use. ✓ Does not require the use of additional post-hoc models or techniques.	X Increased effort during model building. X There are no techniques applicable to all types of models for the time being.

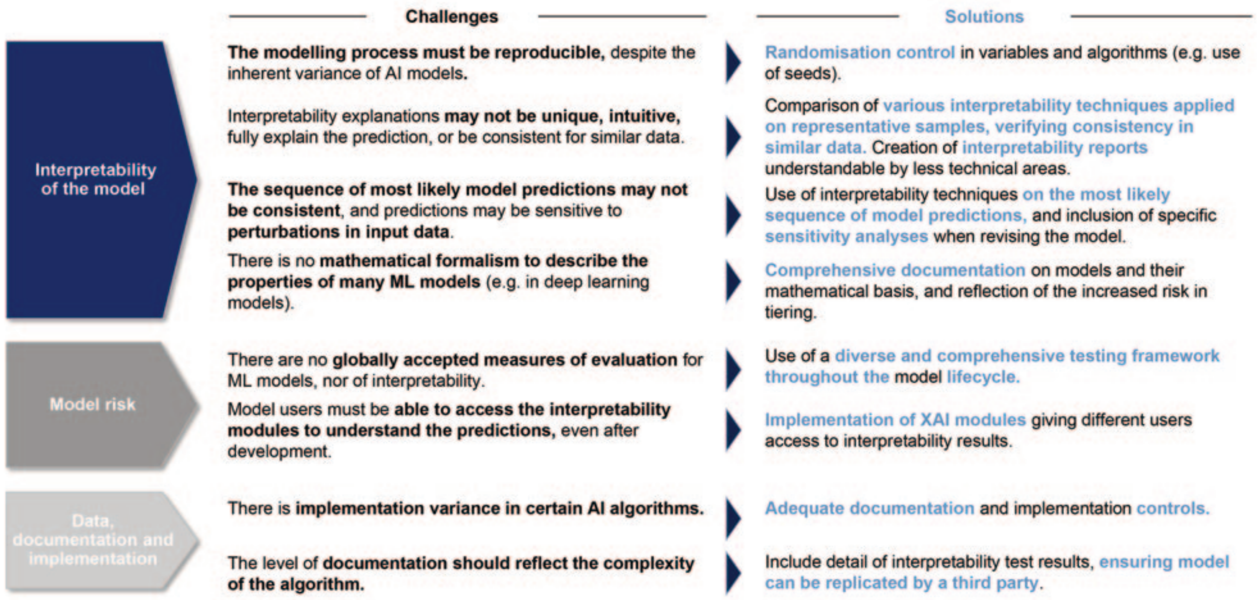
the results yielded by different interpretability techniques at different levels may initially appear contradictory (e.g. if "average" global results are compared with local results in a particular environment).

Thirdly, improvements are still needed in the development of white box models, since, despite the progress made in recent years, these models are still not able to compete in accuracy with black box models in complex problems.

Finally, the need to explain more complex models (e.g. certain types of deep neural networks) remains an unresolved challenge.

In this regard, new techniques are being developed to improve the interpretability of the models, such as the use of information from the intermediate layers of deep neural networks, the aggregation of interpretability metrics to measure the explainability of the models, the development of adversarial models to quantify the degree of explainability, the limitation of the parameters to be optimized to increase their interpretability, or the use of visualization techniques to facilitate the understanding of the results.

Figure 12. Common challenges in the interpretability of AI models.



	Challenges	Solutions
Interpretability of the model	**The modelling process must be reproducible,** despite the inherent variance of AI models.	Randomisation control in variables and algorithms (e.g. use of seeds).
	Interpretability explanations **may not be unique, intuitive,** fully explain the prediction, or be consistent for similar data.	Comparison of various interpretability techniques applied on representative samples, verifying consistency in similar data. Creation of interpretability reports understandable by less technical areas.
	The sequence of most likely model predictions may not be consistent, and predictions may be sensitive to **perturbations in input data**.	Use of interpretability techniques on the most likely sequence of model predictions, and inclusion of specific sensitivity analyses when revising the model.
	There is no **mathematical formalism to describe the properties of many ML models** (e.g. in deep learning models).	Comprehensive documentation on models and their mathematical basis, and reflection of the increased risk in tiering.
Model risk	There are no **globally accepted measures of evaluation** for ML models, nor of interpretability.	Use of a diverse and comprehensive testing framework throughout the model lifecycle.
	Model users must be **able to access the interpretability modules to understand the predictions,** even after development.	Implementation of XAI modules giving different users access to interpretability results.
Data, documentation and implementation	There is **implementation variance in certain AI algorithms.**	Adequate documentation and implementation controls.
	The level of **documentation should reflect the complexity of the algorithm.**	Include detail of interpretability test results, ensuring model can be replicated by a third party.