Project explAIn

# Explaining decisions made with AI

ico.
Information Commissioner's Office

The Alan Turing Institute

# Introduction

This co-badged guidance by the ICO and The Alan Turing Institute aims to give organisations practical advice to help explain the processes, services and decisions delivered or assisted by AI, to the individuals affected by them.

## At a glance

Increasingly, organisations are using artificial intelligence (AI) to support, or to make decisions about individuals. If this is something you do, or something you are thinking about, this guidance is for you.

The guidance consists of three parts. Depending on your level of expertise, and the make-up of your organisation, some parts may be more relevant to you than others.

### Part 1: The basics of explaining AI

Aimed at DPOs and compliance teams, part one defines the key concepts and outlines a number of different types of explanations. It will be relevant for all members of staff involved in the development of AI systems.

### Part 2: Explaining AI in practice

Aimed at technical teams, part two helps you with the practicalities of explaining these decisions and providing explanations to individuals. This will primarily be helpful for the technical teams in your organisation, however your DPO and compliance team will also find it useful.

### Part 3: What explaining AI means for your organisation

Aimed at senior management, part three goes into the various roles, policies, procedures and documentation that you can put in place to ensure your organisation is set up to provide meaningful explanations to affected individuals. This is primarily targeted at your organisation's senior management team, however your DPO and compliance team will also find it useful.

# Part 1 The basics of explaining AI

## About this guidance

### What is the purpose of this guidance?

This guidance is intended to help organisations explain decisions made by artificial intelligence systems (AI) to the people affected by them. This guidance is in three parts:

- Part 1 – The basics of explaining AI (this part)
- Part 2 – Explaining AI in practice
- Part 3 – What explaining AI means for your organisation

This part of the guidance outlines the:

- definitions;
- legal requirements for explaining AI;
- benefits and risks of explaining AI;
- explanation types;
- contextual factors; and
- principles that underpin the rest of the guidance.

There are several reasons to explain AI, including complying with the law, and realising benefits for your organisation and wider society. It clarifies how to apply data protection provisions associated with explaining AI decisions, as well as highlighting other relevant legal regimes outside the ICO's remit.

This guidance is not a statutory code of practice under the Data Protection Act 2018 (DPA 2018). Instead, we aim to provide information that will help you comply with a range of legislation, and demonstrate 'best practice'.

### How should we use this guidance?

This introductory section is for all audiences. It contains concepts and definitions that underpin the rest of the guidance.

Data Protection Officers (DPOs) and your organisation's compliance team will primarily find the legal framework section useful.

Technical teams and senior management may also need some awareness of the legal framework, as well as the benefits and risks of explaining AI systems to the individuals affected by their use.

You will also find the "at a glance" sections of this guidance in this summary document. This pulls the fundamental elements of the guidance into one place and makes it easier to find them quickly.

If you run a SME that processes personal data using AI and you have concerns, it is worth remembering that you can get additional support from the ICO's SME web hub.

## What is the status of this guidance?

This guidance is issued in response to the commitment in the Government's AI Sector Deal, but it is not a statutory code of practice under the DPA 2018, nor is it intended as comprehensive guidance on data protection compliance.

This is practical guidance that sets out good practice for explaining decisions to individuals that have been made using AI systems processing personal data.

## Why is this guidance from the ICO and The Alan Turing Institute?

The ICO is responsible for overseeing data protection in the UK, and The Alan Turing Institute (The Turing) is the UK's national institute for data science and artificial intelligence.

In October 2017, Professor Dame Wendy Hall and Jérôme Pesenti published their independent review on growing the AI industry in the UK. The second of the report's recommendations to support uptake of AI was for the ICO and The Turing to:

> "…develop a framework for explaining processes, services and decisions delivered by AI, to improve transparency and accountability."

In April 2018, the government published its AI Sector Deal. The deal tasked the ICO and The Turing to:

> "…work together to develop guidance to assist in explaining AI decisions."

The independent report and the Sector Deal are part of ongoing efforts made by national and international regulators and governments to address the wider implications of transparency and fairness in AI decisions impacting individuals, organisations, and wider society.

# Part 2: Explaining AI in practice

## About this guidance

### What is the purpose of this guidance?

This guidance helps you with the practicalities of explaining AI-assisted decisions and providing explanations to individuals. It shows you how to:

- select the appropriate explanation for your sector and use case;
- choose an appropriately explainable model; and
- use certain tools to extract explanations from less interpretable models.

### How should we use this guidance?

This guidance is primarily for technical teams, however DPOs and compliance teams will also find it useful. It covers the steps you can take to explain AI-assisted decisions to individuals. It starts with how you can choose which explanation type is most relevant for your use case, and what information you should put together for each explanation type. For most of the explanation types, you can derive this information from your organisational governance decisions and documentation.

However, given the central importance of understanding the underlying logic of the AI system for AI-assisted explanations, we provide technical teams with a comprehensive guide to choosing appropriately interpretable models. This depends on the use case. We also indicate how to use supplementary tools to extract elements of the model's workings in 'black box' systems. Finally, we show you how you can deliver your explanation, containing the relevant explanation types you have chosen, in the most useful way for the decision recipient.

### What is the status of this guidance?

This guidance is issued in response to the commitment in the Government's AI Sector Deal, but it is not a statutory code of practice under the Data Protection Act 2018 (DPA 2018) nor is it intended as comprehensive guidance on data protection compliance.

This is practical guidance that sets out good practice for explaining decisions to individuals that have been made using AI systems processing personal data.

### Why is this guidance from the ICO and The Alan Turing Institute?

The ICO is responsible for overseeing data protection in the UK, and The Alan Turing Institute (The Turing) is the UK's national institute for data science and artificial intelligence.

In October 2017, Professor Dame Wendy Hall and Jérôme Pesenti published their independent review on growing the AI industry in the UK. The second of the report's recommendations to support uptake of AI was for the ICO and The Turing to:

> “…develop a framework for explaining processes, services and decisions delivered by AI, to improve transparency and accountability.”

In April 2018, the government published its AI Sector Deal. The deal tasked the ICO and The Turing to:

> “…work together to develop guidance to assist in explaining AI decisions.”

The independent report and the Sector Deal are part of ongoing efforts made by national and international regulators and governments to address the wider implications of transparency and fairness in AI decisions impacting individuals, organisations, and wider society.

# Part 3: What explaining AI means for your organisation

## About this guidance

### What is the purpose of this guidance?

This guidance covers the various roles, policies, procedures and documentation that you can put in place to ensure your organisation is set up to provide meaningful explanations to affected individuals.

### How should we use this guidance?

This is primarily for senior executives in your organisation. It offers a broad outline of the roles that have a part to play in providing an explanation to the decision recipient, whether directly or as a part of the decision-making process.

Data protection officers (DPOs) and compliance teams as well as technical teams may also find the documentation section useful.

### What is the status of this guidance?

This guidance is issued in response to the commitment in the Government's AI Sector Deal, but it is not a statutory code of practice under the Data Protection Act 2018 (DPA 2018) nor is it intended as comprehensive guidance on data protection compliance.

This is practical guidance that sets out good practice for explaining decisions to individuals that have been made using AI systems processing personal data.

### Why is this guidance from the ICO and The Alan Turing Institute?

The ICO is responsible for overseeing data protection in the UK, and The Alan Turing Institute (The Turing) is the UK's national institute for data science and artificial intelligence.

In October 2017, Professor Dame Wendy Hall and Jérôme Pesenti published their independent review on growing the AI industry in the UK. The second of the report's recommendations to support uptake of AI was for the ICO and The Turing to:

> "…develop a framework for explaining processes, services and decisions delivered by AI, to improve transparency and accountability."

In April 2018, the government published its AI Sector Deal. The deal tasked the ICO and The Turing to:

> “…work together to develop guidance to assist in explaining AI decisions.”

The independent report and the Sector Deal are part of ongoing efforts made by national and international regulators and governments to address the wider implications of transparency and fairness in AI decisions impacting individuals, organisations, and wider society.

# Annexe 1: Example of building and presenting an explanation of a cancer diagnosis

Bringing together our guidance, the following example shows how a healthcare organisation could use the steps we have outlined to help them structure the process of building and presenting their explanation to an affected patient.

## Task 1: Select priority explanation types by considering the domain, use case and impact on the individuals

First, the healthcare organisation familiarises itself with the explanation types in this guidance. Based on the healthcare setting and the impact of the cancer diagnosis on the patient's life, the healthcare organisation selects the explanation types that it determines are a priority to provide to patients subject to its AI-assisted decisions. It documents its justification for these choices:

**Priority explanation types:**

Rationale – Justifying the reasoning behind the outcome of the AI system to maintain accountability, and useful for patients if visualisation techniques of AI explanation are available for non-experts…

Impact – Due to high impact (life/death) situation, important for patients to understand effects and next steps…

Responsibility – Non-expert audience likely to want to know who to query the AI system's output with…

Safety and performance - Given data and domain complexity, this may help reassure patients about the accuracy, safety and reliability of the AI system's output…

**Other explanation types:**

Data – Simple detail on input data as well as on original training/validation dataset and any external validation data

Fairness – Because this is likely biophysical data, as opposed to social or demographic data, fairness issues will arise in areas such as data representativeness and selection biases, so providing information about bias-mitigating efforts relevant to these may be necessary…

The healthcare organisation formalises these explanation types in the relevant part of its policy on information governance:

**Information governance policy**

Use of AI

Explaining AI decisions to patients

Types of explanations:

- Rationale
- Impact
- Responsibility
- Safety and Performance
- Data
- Fairness

## Task 2: Collect and pre-process your data in an explanation-aware manner

The data the healthcare organisation uses has a bearing on its impact and risk assessment. The healthcare organisation therefore chooses the data carefully and considers the impact of pre-processing to ensure they are able to provide an adequate explanation to the decision recipient.

### Rationale

Information on how data has been labelled and how that shows the reasons for classifying, for example, certain images as tumours.

### Responsibility

Information on who or which part of the healthcare organisation (or, if the system is procured, who or which part of the third-party vendor organisation) is responsible for collecting and pre-processing the patient's data. Being transparent about the process from end to end can help the healthcare organisation to build trust and confidence in their use of AI.

### Data

Information about the data that has been used, how it was collected and cleaned, and why it was chosen to train the model. Details about the steps taken to ensure the data was accurate, consistent, up to date, balanced and complete.

### Safety and performance

Information on the model's selected performance metrics and how, given the available data included in training the model, the healthcare organisation or third party vendor chose the accuracy-related measures they did. Also, information about the measures taken to safeguard that the preparation and pre-processing of the data ensures the system's robustness and reliability under harsh, uncertain or adversarial run-time conditions.

## Task 3: Build your system to ensure you are able to extract relevant information for a range of explanation types

The healthcare organisation, or third party vendor, decides to use an artificial neural network to sequence and extract information from radiologic images. While this model is able to predict the existence and types of tumours, the high-dimensional character of its processing makes it opaque.

The model's design team has chosen supplementary 'salience mapping' and 'class activation mapping' tools

to help them visualise the critical regions of the images that are indicative of malign tumours. These tools render the trouble-areas visible by highlighting the abnormal regions. Such mapping-enhanced images then allow technicians and radiologists to gain a clearer understanding of the clinical basis of the AI model's cancer prediction.

This enables them to ensure that the model's output is supporting evidence-based medical practice and that the model's results are being integrated into other clinical evidence that underwrites these technicians' and radiologists' professional judgment.

### Task 4: Translate the rationale of your system's results into useable and easily understandable reasons

The AI system the hospital uses to detect cancer produces a result, which is a prediction that a particular area on an MRI scan contains a cancerous growth. This prediction comes out as a probability, with a particular level of confidence, measured as a percentage. The supplementary mapping tools subsequently provide the radiologist with a visual representation of the cancerous region.

The radiologist shares this information with the oncologist and other doctors on the medical team along with other detailed information about the performance measures of the system and its certainty levels.

For the patient, the oncologist or other members of the medical team then put this into language, or another format, that the patient can understand. One way the doctors choose to do this is through visually showing the patient the scan and supplementary visualisation tools to help explain the model's result. Highlighting the areas that the AI system has flagged is an intuitive way to help the patient understand what is happening. The doctors also indicate how much confidence they have in the AI system's result based on its performance and uncertainty metrics as well as their weighing of other clinical evidence against these measures.

### Task 5: Prepare implementers to deploy your AI system

Because the technician and oncologist are both using the AI system in their work, the hospital decides they need training in how to use the system.

Implementer training covers:

- how they should interpret the results that the AI system generates, based on understanding how it has been designed and the data it has been trained on;
- how they should understand and weigh the performance and certainty limitations of the system (ie how they view and interpret confusion matrices, confidence intervals, error bars, etc);
- that they should use the result as one part of their decision-making, as a complement to their existing domain knowledge;
- that they should critically examine whether the AI system's result is based on appropriate logic and rationale; and
- that in each case they should prepare a plan for communicating the AI system's result to the patient, and the role that result has played in the doctor's judgement.

This includes any limitations in using the system.

## Task 6: Consider how to build and present your explanation

The explanation types the healthcare organisation has chosen each has a process-based and outcome-based explanation. The quality of each explanation is also influenced by how they collect and prepare the training and test data for the AI model they choose. They therefore collect the following information for each explanation type:

### Rationale

Process-based explanation: information to show that the AI system has been set up in a way that enables explanations of its underlying logic to be extracted (directly or using supplementary tools); and that these explanations are meaningful for the patients concerned.

Outcome-based explanation: information on the logic behind the model's results and on how implementers have incorporated that logic into their decision-making. This includes how the system transforms input data into outputs, how this is translated into language that is understandable to patients, and how the medical team uses the model's results in reaching a diagnosis for a particular case.

### Responsibility

Process-based explanation: information on those responsible within the healthcare organisations, or third party provider, for managing the design and use of the AI model, and how they ensured the model was responsibly managed throughout its design and use.

Outcome-based explanation: information on those responsible for using the AI system's output as evidence to support the diagnosis, for reviewing it, and for providing explanations for how the diagnosis came about (ie who the patient can go to in order to query the diagnosis).

### Safety and performance

Process-based explanation: information on the measures taken to ensure the overall safety and technical performance (security, accuracy, reliability, and robustness) of the AI model—including information about the testing, verification, and validation done to certify these.

Outcome-based explanation: Information on the safety and technical performance (security, accuracy, reliability, and robustness) of the AI model in its actual operation, eg information confirming that the model operated securely and according to its intended design in the specific patient's case. This could include the safety and performance measures used.

### Impact

Process-based explanation: measures taken across the AI model's design and use to ensure that it does not negatively impact the wellbeing of the patient.

Outcome-based explanation: information on the actual impacts of the AI system on the patient.

The healthcare organisation considers what contextual factors are likely to have an effect on what patients want to know about the AI-assisted decisions it plans to make on a cancer diagnosis. It draws up a list of the relevant factors:

### Contextual factors

Domain – regulated, safety testing…
Data – biophysical…
Urgency – if cancer, urgent…
Impact – high, safety-critical…
Audience – mostly non-expert…

The healthcare organisation develops a template for delivering their explanation of AI decisions about cancer diagnosis in a layered way:

### Layer 1

- Rationale explanation
- Impact explanation
- Responsibility explanation
- Safety and Performance explanation

Delivery – eg the clinician provides the explanation face to face with the patient, supplemented by hard copy/ email information.

### Layer 2

- Data explanation
- Fairness explanation

Delivery – eg the clinician gives the patient this additional information in hard copy/ email or via an app.

# Annexe 2: Algorithmic techniques

| Algorithm type | Basic description | Possible uses | Interpretability |
|---|---|---|---|
| **Linear regression (LR)** | Makes predictions about a target variable by summing weighted input/predictor variables. | Advantageous in highly regulated sectors like finance (eg credit scoring) and healthcare (predict disease risk given eg lifestyle and existing health conditions) because it's simpler to calculate and have oversight over. | High level of interpretability because of linearity and monotonicity. Can become less interpretable with increased number of features (ie high dimensionality). |
| **Logistic regression** | Extends linear regression to classification problems by using a logistic function to transform outputs to a probability between 0 and 1. | Like linear regression, advantageous in highly regulated and safety-critical sectors, but in use cases that are based in classification problems such as yes/no decisions on risks, credit, or disease. | Good level of interpretability but less so than LR because features are transformed through a logistic function and related to the probabilistic result logarithmically rather than as sums. |
| **Regularised regression (LASSO and Ridge)** | Extends linear regression by adding penalisation and regularisation to feature weights to increase sparsity/ reduce dimensionality. | Like linear regression, advantageous in highly regulated and safety-critical sectors that require understandable, accessible, and transparent results. | High level of interpretability due to improvements in the sparsity of the model through better feature selection procedures. |
| **Generalised linear model (GLM)** | To model relationships between features and target variables that do not follow normal (Gaussian) distributions a GLM introduces a link function that allows for the extension of LR to non-normal distributions. | This extension of LR is applicable to use cases where target variables have constraints that require the exponential family set of distributions (for instance, if a target variable involves number of people, units of time or probabilities of outcome, the result has to have a non-negative value). | Good level of interpretability that tracks the advantages of LR while also introducing more flexibility. Because of the link function, determining feature importance may be less straightforward than with the additive character of simple LR, a degree of transparency may be lost. |

| | | | |
|---|---|---|---|
| **Generalised additive model (GAM)** | To model non-linear relationships between features and target variables (not captured by LR), a GAM sums non-parametric functions of predictor variables (like splines or tree-based fitting) rather than simple weighted features. | This extension of LR is applicable to use cases where the relationship between predictor and response variables is not linear (ie where the input-output relationship changes at different rates at different times) but optimal interpretability is desired. | Good level of interpretability because, even in the presence of non-linear relationships, the GAM allows for clear graphical representation of the effects of predictor variables on response variables. |
| **Decision tree (DT)** | A model that uses inductive branching methods to split data into interrelated decision nodes which terminate in classifications or predictions. DT's moves from starting 'root' nodes to terminal 'leaf' nodes, following a logical decision path that is determined by Boolean-like 'if-then' operators that are weighted through training. | Because the step-by-step logic that produces DT outcomes is easily understandable to non-technical users (depending on number of nodes/ features), this method may be used in high-stakes and safety-critical decision-support situations that require transparency as well as many other use cases where volume of relevant features is reasonably low. | High level of interpretability if the DT is kept manageably small, so that the logic can be followed end-to-end. The advantage of DT's over LR is that the former can accommodate non-linearity and variable interaction while remaining interpretable. |
| **Rule/decision lists and sets** | Closely related to DT's, rule/decision lists and sets apply series of if-then statements to input features in order to generate predictions. Whereas decision lists are ordered and narrow down the logic behind an output by applying 'else' rules, decision sets keep individual if-then statements unordered and largely independent, while weighting them so that rule voting can occur in generating predictions. | As with DT's, because the logic that produces rule lists and sets is easily understandable to non-technical users, this method may be used in high-stakes and safety-critical decision-support situations that require transparency as well as many other use cases where the clear and fully transparent justification of outcomes is a priority. | Rule lists and sets have one of the highest degrees of interpretability of all optimally performing and non-opaque algorithmic techniques. However, they also share with DT's the same possibility that degrees of understandability are lost as the rule lists get longer or the rule sets get larger. |
| **Case-based reasoning (CBR)/** | Using exemplars drawn from prior human knowledge, CBR | CBR is applicable in any domain where experience-based | CBR is interpretable-by-design. It uses examples drawn from |

| | | | |
|---|---|---|---|
| **Prototype and criticism** | predicts cluster labels by learning prototypes and organising input features into subspaces that are representative of the clusters of relevance. This method can be extended to use maximum mean discrepancy (MMD) to identify 'criticisms' or slices of the input space where a model most misrepresents the data. A combination of prototypes and criticisms can then be used to create optimally interpretable models. | reasoning is used for decision-making. For instance, in medicine, treatments are recommended on a CBR basis when prior successes in like cases point the decision maker towards suggesting that treatment. The extension of CBR to methods of prototype and criticism has meant a better facilitation of understanding of complex data distributions, and an increase in insight, actionability, and interpretability in data mining. | human knowledge in order to syphon input features into human recognisable representations. It preserves the explainability of the model through both sparse features and familiar prototypes. |
| **Supersparse linear integer model (SLIM)** | SLIM utilises data-driven learning to generate a simple scoring system that only requires users to add, subtract, and multiply a few numbers in order to make a prediction. Because SLIM produces such a sparse and accessible model, it can be implemented quickly and efficiently by non-technical users, who need no special training to deploy the system. | SLIM has been used in medical applications that require quick and streamlined but optimally accurate clinical decision-making. A version called Risk-Calibrated SLIM (RiskSLIM) has been applied to the criminal justice sector to show that its sparse linear methods are as effective for recidivism prediction as some opaque models that are in use. | Because of its sparse and easily understandable character, SLIM offers optimal interpretability for human-centred decision-support. As a manually completed scoring system, it also ensures the active engagement of the interpreter-user, who implements it. |
| **Naïve Bayes** | Uses Bayes rule to estimate the probability that a feature belongs to a given class, assuming that features are independent of each other. To classify a feature, the Naïve Bayes classifier computes the posterior probability for the class | While this technique is called naïve for reason of the unrealistic assumption of the independence of features, it is known to be very effective. Its quick calculation time and scalability make it good for applications with high dimensional | Naïve Bayes classifiers are highly interpretable, because the class membership probability of each feature is computed independently. The assumption that the conditional probabilities of the independent variables are statistically independent, however, is |

| | | | |
|---|---|---|---|
| | membership of that feature by multiplying the prior probability of the class with the class conditional probability of the feature. | feature spaces. Common applications include spam filtering, recommender systems, and sentiment analysis. | also a weakness, because feature interactions are not considered. |
| **K-nearest neighbour (KNN)** | Used to group data into clusters for purposes of either classification or prediction, this technique identifies a neighbourhood of nearest neighbours around a data point of concern and either finds the mean outcome of them for prediction or the most common class among them for classification. | KNN is a simple, intuitive, versatile technique that has wide applications but works best with smaller datasets. Because it is non-parametric (makes no assumptions about the underlying data distribution), it is effective for non-linear data without losing interpretability. Common applications include recommender systems, image recognition, and customer rating and sorting. | KNN works off the assumption that classes or outcomes can be predicted by looking at the proximity of the data points upon which they depend to data points that yielded similar classes and outcomes. This intuition about the importance of nearness/proximity is the explanation of all KNN results. Such an explanation is more convincing when the feature space remains small, so that similarity between instances remains accessible. |
| **Support vector machines (SVM)** | Uses a special type of mapping function to build a divider between two sets of features in a high dimensional feature space. An SVM therefore sorts two classes by maximising the margin of the decision boundary between them. | SVM's are extremely versatile for complex sorting tasks. They can be used to detect the presence of objects in images (face/no face; cat/no cat), to classify text types (sports article/arts article), and to identify genes of interest in bioinformatics. | Low level of interpretability that depends on the dimensionality of the feature space. In context-determined cases, the use of SVM's should be supplemented by secondary explanation tools. |
| **Artificial neural net (ANN)** | Family of non-linear statistical techniques (including recurrent, convolutional, and deep neural nets) that build complex mapping functions to predict or classify data by employing the feedforward—and sometimes feedback—of input variables through | ANN's are best suited to complete a wide range of classification and prediction tasks for high dimensional feature spaces—ie cases where there are very large input vectors. Their uses may range from computer vision, image recognition, sales and weather forecasting, | The tendencies towards curviness (extreme non-linearity) and high-dimensionality of input variables produce very low-levels of interpretability in ANN's. They are considered to be the epitome of 'black box' techniques. Where appropriate, the use of ANN's should be |

| | | | |
|---|---|---|---|
| | trained networks of interconnected and multi-layered operations. | pharmaceutical discovery, and stock prediction to machine translation, disease diagnosis, and fraud detection. | supplemented by secondary explanation tools. |
| **Random Forest** | Builds a predictive model by combining and averaging the results from multiple (sometimes thousands) of decision trees that are trained on random subsets of shared features and training data. | Random forests are often used to effectively boost the performance of individual decisions trees, to improve their error rates, and to mitigate overfitting. They are very popular in high-dimensional problem areas like genomic medicine and have also been used extensively in computational linguistics, econometrics, and predictive risk modelling. | Very low levels of interpretability may result from the method of training these ensembles of decision trees on bagged data and randomised features, the number of trees in a given forest, and the possibility that individual trees may have hundreds or even thousands of nodes. |
| **Ensemble methods** | As their name suggests, ensemble methods are a diverse class of meta-techniques that combines different 'learner' models (of the same or different type) into one bigger model (predictive or classificatory) in order to decrease the statistical bias, lessen the variance, or improve the performance of any one of the sub-models taken separately. | Ensemble methods have a wide range of applications that tracks the potential uses of their constituent learner models (these may include DT's, KNN's, Random Forests, Naïve Bayes, etc.). | The interpretability of Ensemble Methods varies depending upon what kinds of methods are used. For instance, the rationale of a model that uses bagging techniques, which average together multiple estimates from learners trained on random subsets of data, may be difficult to explain. Explanation needs of these kinds of techniques should be thought through on a case-by-case basis. |

# Annexe 3: Supplementary models

| Supplementary explanation strategy | What is it and what is it useful for? | Limitations |
|---|---|---|
| **Surrogate models (SM)** | SM's build a simpler interpretable model (often a decision tree or rule list) from the dataset and predictions of an opaque system. The purpose of the SM is to provide an understandable proxy of the complex model that estimates that model well, while not having the same degree of opacity. They are good for assisting in processes of model diagnosis and improvement and can help to expose overfitting and bias. They can also represent some non-linearities and interactions that exist in the original model. | As approximations, SM's often fail to capture the full extent of non-linear relationships and high-dimensional interactions among features. There is a seemingly unavoidable trade-off between the need for the SM to be sufficiently simple so that it is understandable by humans, and the need for that model to be sufficiently complex so that it can represent the intricacies of how the mapping function of a 'black box' model works as a whole. That said, the R2 measurement can provide a good quantitative metric of the accuracy of the SM's approximation of the original complex model. |
| **Global/local? Internal/post-hoc?** | For the most part, SM's may be used both globally and locally. As simplified proxies, they are post-hoc. | |
| **Partial Dependence Plot (PDP)** | A PDP calculates and graphically represents the marginal effect of one or two input features on the output of an opaque model by probing the dependency relation between the input variable(s) of interest and the predicted outcome across the dataset, while averaging out the effect of all the other features in the model. This is a good visualisation tool, which allows a clear and intuitive representation of the nonlinear behaviour for complex functions (like random forests and SVM's). It is helpful, for instance, in showing that a given model of interest meets monotonicity constraints across the distribution it fits. | While PDP's allow for valuable access to non-linear relationships between predictor and response variables, and therefore also for comparisons of model behaviour with domain-informed expectations of reasonable relationships between features and outcomes, they do not account for interactions between the input variables under consideration. They may, in this way, be misleading when certain features of interest are strongly correlated with other model features.<br><br>Because PDP's average out marginal effects, they may also be misleading if features have uneven effects on the response function across different subsets of the data—ie where they have different |

| | | |
|---|---|---|
| | | associations with the output at different points. The PDP may flatten out these heterogeneities to the mean. |
| **Global/local? Internal/post-hoc?** | PDP's are global post-hoc explainers that can also allow deeper causal understandings of the behaviour of an opaque model through visualisation. These insights are, however, very partial and incomplete both because PDP's are unable to represent feature interactions and heterogenous effects, and because they are unable to graphically represent more than a couple of features at a time (human spatial thinking is limited to a few dimensions, so only two variables in 3D space are easily graspable). | |
| **Individual Conditional Expectations Plot (ICE)** | Refining and extending PDP's, ICE plots graph the functional relationship between a single feature and the predicted response for an individual instance. Holding all features constant except the feature of interest, ICE plots represent how, for each observation, a given prediction changes as the values of that feature vary. Significantly, ICE plots therefore disaggregate or break down the averaging of partial feature effects generated in a PDP by showing changes in the feature-output relationship for each specific instance, ie observation-by-observation. This means that it can both detect interactions and account for uneven associations of predictor and response variables. | When used in combination with PDP's, ICE plots can provide local information about feature behaviour that enhances the coarser global explanations offered by PDP's. Most importantly, ICE plots are able to detect the interaction effects and heterogeneity in features that remain hidden from PDP's in virtue of the way they compute the partial dependence of outputs on features of interest by averaging out the effect of the other predictor variables. Still, although ICE plots can identify interactions, they are also liable to missing significant correlations between features and become misleading in some instances.<br><br>Constructing ICE plots can also become challenging when datasets are very large. In these cases, time-saving approximation techniques such as sampling observation or binning variables can be employed (but, depending on adjustments and size of the dataset, with an unavoidable impact on explanation accuracy). |

| | | |
|---|---|---|
| **Global/local? Internal/post-hoc?** | ICE plots offer a local and post-hoc form of supplementary explanation. | |
| **Accumulated Local Effects Plots (ALE)** | As an alternative approach to PDP's, ALE plots provide a visualisation of the influence of individual features on the predictions of a 'black box' model by averaging the sum of prediction differences for instances of features of interest in localised intervals and then integrating these averaged effects across all of the intervals. By doing this, they are able to graph the accumulated local effects of the features on the response function as a whole. Because ALE plots use local differences in prediction when computing the averaged influence of the feature (instead of its marginal effect as do PDP's), it is able to better account for feature interactions and avoid statistical bias. This ability to estimate and represent feature influence in a correlation-aware manner is an advantage of ALE plots.<br><br>ALE plots are also more computationally tractable than PDP's because they are able to use techniques to compute effects in smaller intervals and chunks of observations. | A notable limitation of ALE plots has to do with the way that they carve up the data distribution into intervals that are largely chosen by the explanation designer. If there are too many intervals, the prediction differences may become too small and less stably estimate influences. If the intervals are widened too much, the graph will cease to sufficiently represent the complexity of the underlying model.<br><br>While ALE plots are good for providing global explanations that account for feature correlations, the strengths of using PDP's in combination with ICE plots should also be considered (especially when there are less interaction effects in the model being explained). All three visualisation techniques shed light on different dimensions of interest in explaining opaque systems, so the appropriateness of employing them should be weighed case-by-case. |
| **Global/local? Internal/post-hoc?** | ALE plots are a global and post-hoc form of supplementary explanation. | |
| **Global Variable Importance** | The global variable importance strategy calculates the contribution of each input feature to model output across the dataset by permuting the | While permuting variables to measure their relative importance, to some extent, accounts for interaction effects, there is still a |

| | | |
|---|---|---|
| | feature of interest and measuring changes in the prediction error: if changing the value of the permuted feature increases the model error, then that feature is considered to be important. Utilising global variable importance to understand the relative influence of features on the performance of the model can provide significant insight into the logic underlying the model's behaviour. This method also provides valuable understanding about non-linearities in the complex model that is being explained. | high degree of imprecision in the method with regard to which variables are interacting and how much these interactions are impacting the performance of the model.<br><br>A bigger picture limitation of global variable importance comes from what is known as the 'Rashomon effect'. This refers to the variety of different models that may fit the same data distribution equally well. These models may have very different sets of significant features. Because the permutation-based technique can only provide explanatory insight with regard to a single model's performance, it is unable to address this wider problem of the variety of effective explanation schemes. |
| **Global/local? Internal/post-hoc?** | Global variable importance is a form of global and post-hoc explanation. | |
| **Global Variable Interaction** | The global variable interaction strategy computes the importance of variable interactions across the dataset by measuring the variance in the model's prediction when potentially interacting variables are assumed to be independent. This is primarily done by calculating an 'H-statistic' where a no-interaction partial dependence function is subtracted from an observed partial dependence function in order to compute the variance in the prediction. This is a versatile explanation strategy, which has been employed to calculate interaction effects in many types of complex models including ANN's and Random Forests. It can be used to calculate | While the basic capacity to identify interaction effects in complex models is a positive contribution of global variable interaction as a supplementary explanatory strategy, there are a couple of potential drawbacks to which you may want to pay attention.<br><br>First, there is no established metric in this method to determine the quantitative threshold across which measured interactions become significant. The relative significance of interactions is useful information as such, but there is no way to know at which point interactions are strong enough to exercise effects. |

| | | |
|---|---|---|
| | interactions between two or more variables and also between variables and the response function as a whole. It has been effectively used, for example, in biological research to identify interaction effects among genes. | Second, the computational burden of this explanation strategy is very high, because interaction effects are being calculated combinatorially across all the data points. This means that as the number of data points increase, the number of necessary computations increase exponentially. |
| **Global/local? Internal/post-hoc?** | Global variable interaction is a form of global and post-hoc explanation. | |
| **Sensitivity Analysis and Layer-Wise Relevance Propagation (LRP)** | Sensitivity analysis and LRP are supplementary explanation tools used for artificial neural networks. Sensitivity analysis identifies the most relevant features of an input vector by calculating local gradients to determine how a data point has to be moved to change the output label. Here, an output's sensitivity to such changes in input values identifies the most relevant features. LRP is another method to identify feature relevance that is downstream from sensitivity analysis. It uses a strategy of moving backward through the layers of a neural net graph to map patterns of high activation in the nodes and ultimately generates interpretable groupings of salient input variables that can be visually represented in a heat or pixel attribution map. | Both sensitivity analysis and LRP identify important variables in the vastly large feature spaces of neural nets. These explanatory techniques find visually informative patterns by mathematically piecing together the values of individual nodes in the network. As a consequence of this piecemeal approach, they offer very little by way of an account of the reasoning or logic behind the results of an ANNs' data processing.<br><br>Recently, more and more research has focused on attention-based methods of identifying the higher-order representations that are guiding the mapping functions of these kinds of models as well as on interpretable CBR methods that are integrated into ANN architectures and that analyse images by identifying prototypical parts and combining them into a representational wholes. These newer techniques are showing that some significant progress is being made in uncovering the underlying logic of some ANN's. |

| | | |
|---|---|---|
| **Global/local? Internal/post-hoc?** | Sensitivity analysis and salience mapping are forms of local and post-hoc explanation, although the recent incorporation of CBR techniques is moving neural net explanations toward a more internal basis of interpretation. | |
| **Local Interpretable Model-Agnostic Explanation (LIME) and anchors** | LIME works by fitting an interpretable model to a specific prediction or classification produced by an opaque system. It does this by sampling data points at random around the target prediction or classification and then using them to build a local approximation of the decision boundary that can account for the features which figure prominently in the specific prediction or classification under scrutiny.<br><br>LIME does this by generating a simple linear regression model by weighting the values of the data points, which were produced by randomly perturbing the opaque model, according to their proximity to the original prediction or classification. The closest of these values to the instance being explained are weighted the heaviest, so that the supplemental model can produce an explanation of feature importance that is locally faithful to that instance. Note that other interpretive models like decision trees may be used as well. | While LIME appears to be a step in the right direction, in its versatility and in the availability of many iterations in very useable software, a host of issues that present challenges to the approach remains unresolved.<br><br>For instance, the crucial aspect of how to properly define the proximity measure for the 'neighbourhood' or 'local region' where the explanation applies remains unclear, and small changes in the scale of the chosen measure can lead to greatly diverging explanations. Likewise, the explanation produced by the supplemental linear model can quickly become unreliable, even with small and virtually unnoticeable perturbations of the system it is attempting to approximate. This challenges the basic assumption that there is always some simplified interpretable model that successfully approximates the underlying model reasonably well near any given data point.<br><br>LIME's creators have largely acknowledged these shortcomings and have recently offered a new explanatory approach that they call 'anchors'. These 'high precision rules' incorporate into their formal structures 'reasonable patterns' that are operating within the underlying model (such as the implicit linguistic conventions that are at work in a |

| | | |
|---|---|---|
| | | sentiment prediction model), so that they can establish suitable and faithful boundaries of their explanatory coverage of its predictions or classifications. |
| **Global/local? Internal/post-hoc?** | LIME offers a local and post-hoc form of supplementary explanation. | |
| **Shapley Additive ExPlanations (SHAP)** | SHAP uses concepts from cooperative game theory to define a 'Shapley value' for a feature of concern that provides a measurement of its influence on the underlying model's prediction.<br><br>Broadly, this value is calculated by averaging the feature's marginal contribution to every possible prediction for the instance under consideration. The way SHAP computes marginal contributions is by constructing two instances: the first instance includes the feature being measured, while the second leaves it out by substituting a randomly selected stand-in variable for it. After calculating the prediction for each of these instances by plugging their values into the original model, the result of the second is subtracted from that of the first to determine the marginal contribution of the feature. This procedure is then repeated for all possible combinations of features so that the weighted average of all of the marginal contributions of the feature of concern can be | Of the several drawbacks of SHAP, the most practical one is that such a procedure is computationally burdensome and becomes intractable beyond a certain threshold.<br><br>Note, though, some later SHAP versions do offer methods of approximation such as Kernel SHAP and Shapley Sampling Values to avoid this excessive computational expense. These methods do, however, affect the overall accuracy of the method.<br><br>Another significant limitation of SHAP is that its method of sampling values in order to measure marginal variable contributions assumes feature independence (ie that values sampled are not correlated in ways that might significantly affect the output for a particular calculation). As a consequence, the interaction effects engendered by and between the stand-in variables that are used as substitutes for left-out features are necessarily unaccounted for when conditional contributions are approximated. The result is the introduction of uncertainty into the explanation that is produced, because the |

| | | |
|---|---|---|
| | computed.<br><br>This method then allows SHAP, by extension, to estimate the Shapley values for all input features in the set to produce the complete distribution of the prediction for the instance. While computationally intensive, this means that for the calculation of the specific instance, SHAP can axiomatically guarantee the consistency and accuracy of its reckoning of the marginal effect of the feature. This computational robustness has made SHAP attractive as an explainer for a wide variety of complex models, because it can provide a more comprehensive picture of relative feature influence for a given instance than any other post-hoc explanation tool. | complexity of multivariate interactions in the underlying model may not be sufficiently captured by the simplicity of this supplemental interpretability technique. This drawback in sampling (as well as a certain degree of arbitrariness in domain definition) can cause SHAP to become unreliable even with minimal perturbations of the model it is approximating.<br><br>There are currently efforts being made to account for feature dependencies in the SHAP calculations. The original creators of the technique have introduced Tree SHAP to, at least partially, include feature interactions. Others have recently introduced extensions of Kernel SHAP. |
| **Global/local? Internal/post-hoc?** | SHAP offers a local and post-hoc form of supplementary explanation. | |
| **Counterfactual Explanation** | Counterfactual explanations offer information about how specific factors that influenced an algorithmic decision can be changed so that better alternatives can be realised by the recipient of a particular decision or outcome.<br><br>Incorporating counterfactual explanations into a model at its point of delivery allows stakeholders to see what input variables of the model can be modified, so that the outcome could be altered to their benefit. For AI systems that assist | While counterfactual explanation offers a useful way to contrastively explore how feature importance may influence an outcome, it has limitations that originate in the variety of possible features that may be included when considering alternative outcomes. In certain cases, the sheer number of potentially significant features that could be at play in counterfactual explanations of a given result can make a clear and direct explanation difficult to obtain and selected sets of possible explanations seem potentially arbitrary. |

| | | |
|---|---|---|
| | decisions about changeable human actions (like loan decisions or credit scoring), incorporating counterfactual explanation into the development and testing phases of model development may allow the incorporation of actionable variables, ie input variables that will afford decision subjects with concise options for making practical changes that would improve their chances of obtaining the desired outcome.<br><br>In this way, counterfactual explanatory strategies can be used as way to incorporate reasonableness and the encouragement of agency into the design and implementation of AI systems. | Moreover, there are as yet limitations on the types of datasets and functions to which these kinds of explanations are applicable.<br><br>Finally, because this kind of explanation concedes the opacity of the algorithmic model outright, it is less able to address concerns about potentially harmful feature interactions and questionable covariate relationships that may be buried deep within the model's architecture. It is a good idea to use counterfactual explanations in concert with other supplementary explanation strategies—that is, as one component of a more comprehensive explanation portfolio. |
| **Global/local? Internal/post-hoc?** | Counterfactual explanations are a local and post-hoc form of supplementary explanation strategy. | |
| **Self-Explaining and Attention-Based Systems** | Self-explaining and attention-based systems actually integrate secondary explanation tools into the opaque systems so that they can offer runtime explanations of their own behaviours. For instance, an image recognition system could have a primary component, like a convolutional neural net, that extracts features from its inputs and classifies them while a secondary component, like a built-in recurrent neural net with an 'attention-directing' mechanism translates the extracted features into a natural language representation that | Automating explanations through self-explaining systems is a promising approach for applications where users benefit from gaining real-time insights about the rationale of the complex systems they are operating. However, regardless of their practical utility, these kinds of secondary tools will only work as well as the explanatory infrastructure that is actually unpacking their underlying logics. This explanatory layer must remain accessible to human evaluators and be understandable to affected individuals. Self-explaining systems, in other words, should themselves remain optimally interpretable. The |

| | | |
|---|---|---|
| | produces a sentence-long explanation of the result to the user.<br><br>Research into integrating 'attention-based' interfaces is continuing to advance toward potentially making their implementations more sensitive to user needs, explanation-forward, and humanly understandable. Moreover, the incorporation of domain knowledge and logic-based or convention-based structures into the architectures of complex models are increasingly allowing for better and more user-friendly representations and prototypes to be built into them. | task of formulating a primary strategy of supplementary explanation is still part of the process of building out a system with self-explaining capacity.<br><br>Another potential pitfall to consider for self-explaining systems is their ability to mislead or to provide false reassurance to users, especially when humanlike qualities are incorporated into their delivery method. This can be avoided by not designing anthropomorphic qualities into their user interface and by making uncertainty and error metrics explicit in the explanation as it is delivered. |
| **Global/local? Internal/post-hoc?** | Because self-explaining and attention-based systems are secondary tools that can utilise many different methods of explanation, they may be global or local, internal or post-hoc, or a combination of any of them. | |

# Annexe 4: Further reading

## PROV provenance standard

Moreau, L. & Missier, P. (2013). *PROV-DM: The PROV Data Model.* W3C Recommendation. URL: https://www.w3.org/TR/2013/REC-prov-dm-20130430/

Huynh, T. D., Stalla-Bourdillon, S. & Moreau, L. (2019). Provenance-based Explanations for Automated Decisions : Final IAA Project Report. URL: https://kclpure.kcl.ac.uk/portal/en/publications/provenancebased-explanations-forautomated-decisions(5b1426ce-d253-49fa-8390-4bb3abe65f54).html

## Resources for exploring algorithm types

### General

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer, 27(2)*, 83-85. http://thuvien.thanglong.edu.vn:8081/dspace/bitstream/DHTL_123456789/4053/1/%5BSpringer%20Series%20in%20Statistics-1.pdf

Molnar, C. (2019). Interpretable machine learning: A guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1(5)*, 206. https://www.nature.com/articles/s42256-019-0048-x

### Regularised regression (LASSO and Ridge)

Gaines, B. R., & Zhou, H. (2016). Algorithms for fitting the constrained lasso. *Journal of Computational and Graphical Statistics*, 27(4), 861-871. https://arxiv.org/pdf/1611.01511.pdf

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the *Royal Statistical Society: Series B (Methodological), 58*(1), 267-288. http://beehive.cs.princeton.edu/course/read/tibshirani-jrssb-1996.pdf

### Generalised linear model (GLM)

https://CRAN.R-project.org/package=glmnet

Friedman, J., Hastie, T., & Tibshirani, R. (2010). *Regularization paths for generalized linear models via coordinate descent*. Journal of Statistical Software, 33(1), 1-22. http://www.jstatsoft.org/v33/i01/

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). *Regularization paths for Cox's proportional hazards model via coordinate descent*. Journal of Statistical Software, 39(5), 1-13. URL http://www.jstatsoft.org/v39/i05/

### Generalised additive model (GAM)

https://CRAN.R-project.org/package=gam

Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 150-158). ACM. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.433.8241&rep=rep1&type=pdf

Wood, S. N. (2006). *Generalized additive models: An introduction with R*. CRC Press.

### Decision tree (DT)

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. CRC Press.

### Rule/decision lists and sets

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning certifiably optimal rule lists for categorical data. *The Journal of Machine Learning Research, 18*(1), 8753-8830. http://www.jmlr.org/papers/volume18/17-716/17-716.pdf

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016, August). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1675-1684). ACM. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5108651/

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics, 9*(3), 1350-1371. https://projecteuclid.org/download/pdfview_1/euclid.aoas/1446488742

Wang, F., & Rudin, C. (2015). Falling rule lists. In *Artificial Intelligence and Statistics* (pp. 1013-1022). http://proceedings.mlr.press/v38/wang15a.pdf

### Case-based reasoning (CBR)/ Prototype and criticism

Aamodt, A. (1991). A knowledge-intensive, integrated approach to problem solving and sustained learning. *Knowledge Engineering and Image Processing Group. University of Trondheim*, 27-85. http://www.dphu.org/uploads/attachements/books/books_4200_0.pdf

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications, 7*(1), 39-59. https://www.idi.ntnu.no/emner/tdt4171/papers/AamodtPlaza94.pdf

Bichindaritz, I., & Marling, C. (2006). Case-based reasoning in the health sciences: What's next?. *Artificial intelligence in medicine, 36*(2), 127-135. http://cs.oswego.edu/~bichinda/isc471-hci571/AIM2006.pdf

Bien, J., & Tibshirani, R. (2011). Prototype selection for interpretable classification. *The Annals of Applied Statistics, 5*(4), 2403-2424.

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems* (pp. 2280-2288).

http://papers.nips.cc/paper/6300-examples-are-not-enough-learn-to-criticize-criticism-for-interpretability.pdf

MMD-critic in python: https://github.com/BeenKim/MMD-critic

Kim, B., Rudin, C., & Shah, J. A. (2014). The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in Neural Information Processing Systems* (pp. 1952-1960). http://papers.nips.cc/paper/5313-the-bayesian-case-model-a-generative-approach-for-case-based-reasoning-and-prototype-classification.pdf

### Supersparse linear integer model (SLIM)

Jung, J., Concannon, C., Shroff, R., Goel, S., & Goldstein, D. G. (2017). Simple rules for complex decisions. *Available at SSRN 2919024*. https://arxiv.org/pdf/1702.04690.pdf

Rudin, C., & Ustun, B. (2018). Optimized scoring systems: toward trust in machine learning for healthcare and criminal justice. *Interfaces, 48*(5), 449-466. https://pdfs.semanticscholar.org/b3d8/8871ae5432c84b76bf53f7316cf5f95a3938.pdf

Ustun, B., & Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning, 102*(3), 349-391. https://link.springer.com/article/10.1007/s10994-015-5528-6

Optimized scoring systems for classification problems in python: https://github.com/ustunb/slim-python

Simple customizable risk scores in python: https://github.com/ustunb/risk-slim

## Resources for exploring supplementary explanation strategies

### Surrogate models (SM)

Bastani, O., Kim, C., & Bastani, H. (2017). Interpretability via model extraction. *arXiv preprint arXiv:1706.09773*. https://obastani.github.io/docs/fatml17.pdf

Craven, M., & Shavlik, J. W. (1996). Extracting tree-structured representations of trained networks. In *Advances in neural information processing systems* (pp. 24-30). http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf

Van Assche, A., & Blockeel, H. (2007). Seeing the forest through the trees: Learning a comprehensible model from an ensemble. In *European Conference on Machine Learning* (pp. 418-429). Springer, Berlin, Heidelberg. https://link.springer.com/content/pdf/10.1007/978-3-540-74958-5_39.pdf

Valdes, G., Luna, J. M., Eaton, E., Simone II, C. B., Ungar, L. H., & Solberg, T. D. (2016). MediBoost: a patient stratification tool for interpretable decision making in the era of precision medicine. *Scientific reports, 6*, 37854. https://www.nature.com/articles/srep37854

### Partial Dependence Plot (PDP)

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232. https://projecteuclid.org/download/pdf_1/euclid.aos/1013203451

Greenwell, B. M. (2017). pdp: an R Package for constructing partial dependence plots. *The R Journal, 9*(1), 421-436. https://pdfs.semanticscholar.org/cdfb/164f55e74d7b116ac63fc6c1c9e9cfd01cd8.pdf

For the software in R: https://cran.r-project.org/web/packages/pdp/index.html

### Individual Conditional Expectations Plot (ICE)

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics, 24*(1), 44-65. https://arxiv.org/pdf/1309.6392.pdf

For the software in R see:
https://cran.r-project.org/web/packages/ICEbox/index.html
https://cran.r-project.org/web/packages/ICEbox/ICEbox.pdf

### Accumulated Local Effects Plots (ALE)

Apley, D. W., & Zhu, J. (2019). Visualizing the effects of predictor variables in black box supervised learning models. *arXiv preprint arXiv:1612.08468*. https://arxiv.org/pdf/1612.08468;Visualizing

https://cran.r-project.org/web/packages/ALEPlot/index.html

### Global variable importance

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Casalicchio, G., Molnar, C., & Bischl, B. (2018, September). Visualizing the feature importance for black box models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 655-670). Springer, Cham. https://arxiv.org/pdf/1804.06620.pdf

Fisher, A., Rudin, C., & Dominici, F. (2018). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. arXiv:1801.01489

Fisher, A., Rudin, C., & Dominici, F. (2018). Model class reliance: Variable importance measures for any machine learning model class, from the "Rashomon" perspective. *arXiv preprint arXiv:1801.01489*. https://arxiv.org/abs/1801.01489v2

Hooker, G., & Mentch, L. (2019). Please Stop Permuting Features: An Explanation and Alternatives. *arXiv preprint arXiv:1905.03151*. https://arxiv.org/pdf/1905.03151.pdf

Zhou, Z., & Hooker, G. (2019). Unbiased Measurement of Feature Importance in Tree-Based Methods. *arXiv preprint arXiv:1903.05179*. https://arxiv.org/pdf/1903.05179.pdf

### Global variable interaction

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics, 2*(3), 916-954. https://projecteuclid.org/download/pdfview_1/euclid.aoas/1223908046

Greenwell, B. M., Boehmke, B. C., & McCarthy, A. J. (2018). A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*. https://arxiv.org/pdf/1805.04755.pdf

Hooker, G. (2004, August). Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 575-580). ACM. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.7500&rep=rep1&type=pdf

### Local Interpretable Model-Agnostic Explanation (LIME)

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). ACM. https://arxiv.org/pdf/1602.04938.pdf?mod=article_inline

LIME in python: https://github.com/marcotcr/lime
LIME experiments in python: https://github.com/marcotcr/lime-experiments

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.7500&rep=rep1&type=pdf

Anchors in python: https://github.com/marcotcr/anchor
Anchors experiments in python: https://github.com/marcotcr/anchor-experiments

### Shapley Additive ExPlanations (SHAP)

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765-4774). http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Software for SHAP and its extensions in python: https://github.com/slundberg/shap
R wrapper for SHAP: https://modeloriented.github.io/shapper/

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games, 2*(28), 307-317. http://www.library.fa.ru/files/Roth2.pdf#page=39

### Counterfactual explanation

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech., 31, 841. https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf

Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*(pp. 10-19). ACM. https://arxiv.org/pdf/1809.06514.pdf

Evaluate recourse in linear classification models in python: https://github.com/ustunb/actionable-recourse

### Secondary explainers and attention-based systems

Li, O., Liu, H., Chen, C., & Rudin, C. (2018). Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Thirty-Second AAAI Conference on Artificial Intelligence*. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewFile/17082/16552

Park, D. H., Hendricks, L. A., Akata, Z., Schiele, B., Darrell, T., & Rohrbach, M. (2016). Attentive explanations: Justifying decisions and pointing to the evidence. *arXiv preprint arXiv:1612.04757*. https://arxiv.org/pdf/1612.04757

### Other resources for supplementary explanation

IBM's Explainability 360: http://aix360.mybluemix.net

Biecek, B., & Burzykowski, T. (2019). *Predictive Models: Explore, Explain, and Debug, Human-Centered Interpretable Machine Learning*. Retrieved from https://pbiecek.github.io/PM_VEE/

Accompanying software, Dalex, Descriptive mAchine Learning Explanations: https://github.com/ModelOriented/DALEX

Przemysław Biecek, *Interesting resources related to XAI*: https://github.com/pbiecek/xai_resources

Christoph Molnar, iml: Interpretable machine learning https://cran.r-project.org/web/packages/iml/index.html

# Annexe 5: Argument-based assurance cases

An assurance case is a set of structured claims, arguments, and evidence which gives confidence that an AI system will possess the particular qualities or properties that need to be assured. Take, for example, a 'safety and performance' assurance case. This would involve providing an argument, supported by evidence, that a system possesses the properties that will allow it to function safely, securely, reliably, etc given the challenges of its operational context.

Though argument-based assurance cases have historically arisen in safety-critical domains (as 'safety cases' for software-based technologies), the methodology is widely used. This is because it is a reasonable way to structure and document an anticipatory, goal-based, and procedural approach to innovation governance. This stands in contrast to the older, more reactive and prescriptive methods that are now increasingly challenged by the complex and rapidly evolving character of emerging AI technologies.

This older prescriptive approach stressed the application of one-size-fits-all general standards that specified how systems should be built and often treated governance as a retrospective check-box exercise. However, argument-based assurance takes a different tack. It starts at the inception of an AI project and plays an active role at all stages of the design and use lifecycle. It begins with high-level normative goals derived from context-anchored impact and risk-based assessment for each specific AI application and then sets out structured arguments demonstrating:

1. how such normative requirements address the impacts and risks associated with the system's use in its specified operating environment;
2. how activities undertaken across the design and deployment workflow assure the properties of the system that are needed for the realisation of these goals; and
3. how appropriate monitoring and re-assessment measures have been set up to ensure the effectiveness of the implemented controls.

We have included a list of background reading and resources to help you in the area of governance and standards that relate to argument-based assurance cases. This includes consolidated standards for system and software assurance from the International Standards Organisation, the International Electrotechnical Commission, and the Institute of Electrical and Electronics Engineers (ISO/IEC/IEEE 15026 series) as well as the Object Management Group's Structured Assurance Case Metamodel (SACM). This also includes references for several of the main assurance platforms like Goal Structuring Notation (GSN), the Claims, Arguments and Evidence Notation (CAE), NOR-STA Argumentation, and Dynamic Safety Cases (DSC).

## Main components of argument-based assurance cases

While it is beyond the scope of this guidance to cover details of the different methods of building assurance cases, it may be useful to provide a broad view of how the main components of any comprehensive assurance case fit together.

### Top-level normative goals

These are the high-level aims or goals of the system that address the risks and potential harms that may be caused by the use of that system in its defined operating environment and are therefore in need of assurance. In the context of process-based explanation, these include fairness and bias-mitigation,

responsibility, safety and optimal performance, and beneficial and non-harmful impact. Starting from each of these normative goals, building an assurance case then involves identifying the properties and qualities that a given system has to possess to ensure that it achieves the specified goal in light of the risks and challenges it faces in its operational context.

### Claims

These are the properties, qualities, traits, or attributes that need to be assured in order for the top-level normative goals to be realised. For instance, in a fairness and bias-mitigation assurance case, the property, 'target variables or their measurable proxies do not reflect underlying structural biases or discrimination,' is one of several such claims. As a central component of structured argumentation, it needs to be backed both by appropriate supporting arguments about how the relevant activities behind the system's design and development process ensured that structural biases were, in fact, not incorporated into target variables and by corresponding evidence that documented these activities.

In some methods of structured argumentation like GSN, claims are qualified by **context components**, which:

- clarify the scope of a given claim;
- provide definitions and background information;
- make relevant assumptions about the system and environment explicit; and
- spell out risks and risk-mitigation needs associated with the claim across the system's design and operation lifecycle.

Claims may also be qualified by **justification components**, that is, clarifications of:

- why the claims have been chosen; and
- how they provide a solution or means of realisation for the specified normative goal.

In general, the addition of context and justification components reinforces the accuracy and completeness of claims, allowing them to support the top-level goals of the system. This focus on precision, clarification, and thoroughness is crucial to establishing confidence through the development of an effective assurance case.

### Arguments

These support claims by linking them with evidence and other supporting claims through reasoning. Arguments provide warrants for claims by establishing an inferential relationship that connects the proposed property with a body of evidence and argumentative backing sufficient to establish its rational acceptability or truth. For example, in a safety and performance assurance case, which had 'system is sufficiently robust' as one of its claims, a possible argument might be 'training processes included an augmentation element where adversarial examples and perturbations were employed to model harsh real-world conditions'. Such a claim would then be backed by evidentiary support that this actually happened during the design and development of the AI model.

While justified arguments are always backed by some body of evidence, they may also be supported by **subordinate claims** and **assumptions** (ie claims without further backing that are taken as self-evident or true). Subordinate claims underwrite arguments (and the higher-level claims they support) based on their own arguments and evidence. In structured argument, there will often be multiple levels subordinate

claims, which work together to provide justification for the rational acceptability or truth of top-level claims.

### Evidence

This is the collection of artefacts and documentation that provide evidential support for the claims made in the assurance case. A body of evidence is formed by objective, demonstrable, and repeatable information recorded during production and use of a system. This underpins the arguments justifying the assurance claims. In some instances, a body of evidence may be organised in an **evidence repository** (SACM) where that primary information can be accessed along with secondary information about evidence management, interpretation of evidence, and clarification of evidentiary support underlying the claims of the assurance case.

## Advantages of approaching process-based explanation through argument-based assurance cases

There are several advantages to using argument-based assurance to organise the documentation of your innovation practices for process-based explanations:

- Assurance cases demand proactive and end-to-end understanding of the impacts and risks that come from each specific AI application. Their effective execution is anchored in building practical controls which show that these impacts and risks have been appropriately managed. Because of this, assurance cases encourage the planned integration of good process-based governance controls. This, in turn, ensures that the goals governing the development of AI systems have been met, with a deliberate and argument-based method of documented assurance that demonstrates this. In argument-based assurance, anticipatory and goal-driven governance and documentation processes work hand-in-glove, mutually strengthening best practices and improving the quality of the products and services they support.
- Argument-based assurance involves a method of reasoning-based governance practice rather than a task- or technology-specific set of instructions. This allows AI designers and developers to tackle a diverse range of governance activities with a single method of using structured argument to assure properties that meet standard requirements and mitigate risks. This also means that procedures for building assurance cases are uniform and that their documentation can be more readily standardised and automated (as seen, for instance, in various assurance platforms like GSN, SACM, CAE, and NOR-STA Argumentation).
- Argument-based assurance can enable effective multi-stakeholder communication. It can generate confidence that a given AI application possesses desired properties on the basis of explicit, well-reasoned, and evidence-backed grounds. When done effectively, assurance cases clearly and precisely convey information to various stakeholder groups through structured arguments. These demonstrate that specified goals have been achieved and risks have been mitigated. They do this by providing documentary evidence that the properties of the system needed to meet these goals have been assured by solid arguments.
- Using the argument-based assurance methodology can enable assurance cases to be customised and tailored to the relevant audiences . These assurance cases are built on structured arguments (claims, justifications, and evidence) in natural language, so they are more readily understood by those who are not technical specialists. While detailed technical arguments and evidence may support assurance cases, the basis of these cases in everyday reasoning makes them especially amenable to non-technical and understandable summary. A summary of an assurance case provided to a decision recipient can then be backed by a more detailed version which includes extended structural arguments better tailored to experts, independent assessors, and auditors. Likewise, the evidence used for an assurance case can be

organised to fit the audience and context of explanation. In this way, any potentially commercially sensitive or privacy impinging information, which comprises a part of the body of evidence, may be held in an evidence repository. This can be made accessible to a more limited audience of internal or external overseers, assessors, and auditors.

## Governing procurement practices and managing stakeholder expectations through tailored assurance

For both vendors and customers (ie developers and users/procurers), argument-based assurance may provide a reasonable way to govern procurement practices and to mutually manage expectations. If a vendor has a deliberate and anticipatory approach to the explanation-aware AI system design demanded by argument-based assurance, they will be better able to assure (through justification, evidence, and documentation) crucial properties of their models to those interested in acquiring them.

By offering an evidence-backed assurance portfolio, in advance, a vendor will be able to demonstrate that their products have been designed with appropriate normative goals in mind. They will also be able to assure potential customers that these goals have been realised across development processes. This will then also allow users/procurers to pass on this part of the process-based explanation to decision-recipients and affected parties. It would also allow procurers to more effectively assess whether the assurance portfolio meets the normative criteria and AI innovation standards they are looking for based on their organisational culture, domain context, and application interests.

Using assurance cases will also enable standards-based independent assessment and third-party audit. The tasks of information management and sharing can be undertaken efficiently between developers, users, assessors, and auditors. This can provide a common and consolidated platform for process-based explanation, which organises both the presentation of the details of assurance cases and the accessibility of the information which supports them. This will streamline communication processes across all affected stakeholders, while preserving the trust-generating aspects of procedural and organisational transparency all the way down.

**Further reading**

**Resources for exploring documentation and argument-based assurance**

**General readings on documentation for responsible AI design and implementation**

Fact sheets

Datasheets for datasets

Model cards for model reporting

AI auditing framework blog

Understanding Artificial Intelligence ethics and safety

**Relevant standards and regulations on argument-based assurance and safety cases**

ISO/IEC/IEEE 15026-1:2019, *Systems and software engineering — Systems and software assurance — Part 1: Concepts and vocabulary*.

ISO/IEC 15026-2:2011, *Systems and software engineering — Systems and software assurance — Part 2: Assurance case.*

ISO/IEC 15026-3:2015, *Systems and software engineering — Systems and software assurance — Part 3: System integrity levels*.

ISO/IEC 15026-4:2012, *Systems and software engineering — Systems and software assurance — Part 4: Assurance in the life cycle*.

Object Management Group, *Structured Assurance Case Metamodel* (SACM), Version 2.1 beta, March 2019.

Ministry of Defence. Defence Standard 00-42 Issue 2, Reliability and Maintainability (R&M) Assurance Guidance. Part 3, R&M Case, 6 June 2003.

Ministry Of Defence. Defence Standard 00-55 (PART 1)/Issue 4, *Requirements for Safety Related Software in Defence Equipment Part 1: Requirements*, December 2004.

Ministry of Defence. Defence Standard 00-55 (PART 2)/Issue 2, *Requirements for Safety Related Software in Defence Equipment Part 2: Guidance*, 21 August 1997.

Ministry of Defence. Defence Standard 00-56. *Safety Management Requirements for Defence Systems. Part 1. Requirements Issue 4*, 01 June 2007

Ministry of Defence. Defence Standard 00-56. *Safety Management Requirements for Defence Systems. Part 2 : Guidance on Establishing a Means of Complying with Part 1 Issue 4*, 01 June 2007.

UK CAA CAP 760 *Guidance on the Conduct of Hazard Identification, Risk Assessment and the Production of Safety Cases For Aerodrome Operators and Air Traffic Service Providers*, 13 January 2006.

*The Offshore Installations (Safety Case) Regulations* 2005 No. 3117. http://www.legislation.gov.uk/uksi/2005/3117/contents/made

*The Control of Major Accident Hazards (Amendment) Regulations* 2005 No.1088. http://www.legislation.gov.uk/uksi/2005/1088/contents/made

Health and Safety Executive. *Safety Assessment Principles for Nuclear Facilities*. HSE; 2006.

*The Railways and Other Guided Transport Systems (Safety) Regulations* 2006. UK Statutory Instrument 2006 No.599. http://www.legislation.gov.uk/uksi/2006/599/contents/made

EC Directive 91/440/EEC. *On the development of the community's railways*. 29 July 1991.

**Background readings on methods of argument-based assurance**

Ankrum, T. S., & Kromholz, A. H. (2005). Structured assurance cases: Three common standards. In

*Ninth IEEE International Symposium on High-Assurance Systems Engineering (HASE'05)* (pp. 99-108).

Ashmore, R., Calinescu, R., & Paterson, C. (2019). Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *arXiv preprint arXiv:1905.04223*.

Barry, M. R. (2011). CertWare: A workbench for safety case production and analysis. In *2011 Aerospace conference* (pp. 1-10). IEEE.

Bloomfield, R., & Netkachova, K. (2014). Building blocks for assurance cases. In *2014 IEEE International Symposium on Software Reliability Engineering Workshops*(pp. 186-191). IEEE.

Bloomfield, R., & Bishop, P. (2010). Safety and assurance cases: Past, present and possible future–an Adelard perspective. In *Making Systems Safer* (pp. 51-67). Springer, London.

Cârlan, C., Barner, S., Diewald, A., Tsalidis, A., & Voss, S. (2017). ExplicitCase: integrated model-based development of system and safety cases. *International Conference on Computer Safety, Reliability, and Security* (pp. 52-63). Springer.

Denney, E., & Pai, G. (2013). A formal basis for safety case patterns. In *International Conference on Computer Safety, Reliability, and Security* (pp. 21-32). Springer.

Denney, E., Pai, G., & Habli, I. (2015). Dynamic safety cases for through-life safety assurance. *IEEE/ACM 37th IEEE International Conference on Software Engineering* (Vol. 2, pp. 587-590). IEEE.

Denney, E., & Pai, G. (2018). Tool support for assurance case development. *Automated Software Engineering*, 25(3), 435-499.

Despotou, G. (2004). Extending the safety case concept to address dependability. *Proceedings of the 22nd International System Safety Conference*.

Gacek, A., Backes, J., Cofer, D., Slind, K., & Whalen, M. (2014). Resolute: an assurance case language for architecture models. *ACM SIGAda Ada Letters*, 34(3), 19-28.

Ge, X., Rijo, R., Paige, R. F., Kelly, T. P., & McDermid, J. A. (2012). Introducing goal structuring notation to explain decisions in clinical practice. *Procedia Technology*, 5, 686-695.

Gleirscher, M., & Kugele, S. (2019). Assurance of System Safety: A Survey of Design and Argument Patterns. *arXiv preprint arXiv:1902.05537*.

Górski, J., Jarzębowicz, A., Miler, J., Witkowicz, M., Czyżnikiewicz, J., & Jar, P. (2012). Supporting assurance by evidence-based argument services. *International Conference on Computer Safety, Reliability, and Security* (pp. 417-426). Springer, Berlin, Heidelberg.

Habli, I., & Kelly, T. (2014, July). Balancing the formal and informal in safety case arguments. In *VeriSure: Verification and Assurance Workshop, colocated with Computer-Aided Verification (CAV)*.

Hawkins, R., Habli, I., Kolovos, D., Paige, R., & Kelly, T. (2015). Weaving an assurance case from design: a model-based approach. *2015 IEEE 16th International Symposium on High Assurance Systems Engineering* (pp. 110-117). IEEE.

Health Foundation (2012). Evidence: Using safety cases in industry and healthcare.

Kelly, T. (1998) *Arguing Safety: A Systematic Approach to Managing Safety Cases. Doctoral Thesis*. University of York: Department of Computer Science.

Kelly, T., & McDermid, J. (1998). Safety case patterns-reusing successful arguments. *IEEE Colloquium on Understanding Patterns and Their Application to System Engineering, London*.

Kelly, T. (2003). *A Systematic Approach to Safety Case Management*. SAE International.

Kelly, T., & Weaver, R. (2004). The goal structuring notation–a safety argument notation. In *Proceedings of the dependable systems and networks 2004 workshop on assurance cases*.

Maksimov, M., Fung, N. L., Kokaly, S., & Chechik, M. (2018). Two decades of assurance case tools: a survey. *International Conference on Computer Safety, Reliability, and Security* (pp. 49-59). Springer.

Nemouchi, Y., Foster, S., Gleirscher, M., & Kelly, T. (2019). Mechanised assurance cases with integrated formal methods in Isabelle. *arXiv preprint arXiv:1905.06192*.

Netkachova, K., Netkachov, O., & Bloomfield, R. (2014). Tool support for assurance case building blocks. In *International Conference on Computer Safety, Reliability, and Security* (pp. 62-71). Springer.

Picardi, C., Hawkins, R., Paterson, C., & Habli, I. (2019, September). A pattern for arguing the assurance of machine learning in medical diagnosis systems. In *International Conference on Computer Safety, Reliability, and Security* (pp. 165-179). Springer.

Picardi, C., Paterson, C., Hawkins, R. D., Calinescu, R., & Habli, I. (2020). Assurance Argument Patterns and Processes for Machine Learning in Safety-Related Systems. *Proceedings of the Workshop on Artificial Intelligence Safety* (pp. 23-30). CEUR Workshop Proceedings.

Rushby, J. (2015). The interpretation and evaluation of assurance cases. Comp. Science Laboratory, SRI International, Tech. Rep. SRI-CSL-15-01.

Strunk, E. A., & Knight, J. C. (2008). The essential synthesis of problem frames and assurance cases. *Expert Systems*, 25(1), 9-27.