Project ExplAIn

Explaining decisions made with AI

Draft guidance for consultation

Part 2:

Explaining AI in practice

ico.
Information Commissioner's Office

The Alan Turing Institute

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 go about selecting the appropriate explanation for your sector and use case, how to choose an appropriately explainable model, and which tools you can use 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 goes through 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.

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.

Summary of the steps to take

We have set out a number of steps to help you provide explanations of your AI decisions. These offer a systematic approach to selecting, extracting and delivering explanations and they should help in navigating the detailed technical recommendations in this part. However, we recognise that in practice some steps may be concurrent rather than consecutive, and organisations may wish develop their own plan for doing this.

1. **Select priority explanations by considering the domain, use case and impact on the individual**

Start by getting to know the different types of explanation in Part 1 of this guidance. This should help you to separate out the different aspects of an AI-assisted decision that people may want you to explain. While we have identified what we think are the key types of explanation that people will need, there may be additional relevant explanations in the context of your

organisation, and the way you do (or plan to) use AI to make decisions about people. Or perhaps some of the explanations we identify are not of particular relevance to your organisation and the people you make decisions about.

That's absolutely fine. The explanations we identify are intended to underline the fact that there are many different aspects to explanations, and to get you thinking about what those aspects are, and whether or not they are relevant to your customers. You may think the list we have created works for your organisation or you might want to create your own.

Either way, we recommend that your approach to explaining AI-assisted decisions should be informed by the importance of putting the principles of transparency and accountability into practice, and of paying close attention to context and impact.

Next, think about the specifics of the context within which you are deploying your AI decision-support system. The domain you work in, the particular use case and the impact on the person will further help you choose the relevant explanations. In most cases, it will be useful for you to include rationale and responsibility in your priority explanations.

It is likely that you will identify multiple explanations to prioritise for the AI-assisted decisions you make. Make a list of these and document the justification for your choices.

While you have identified the explanations that are most important in the context of your AI decision-support system, this does not mean that the remaining explanations should be discarded.

Choosing what to prioritise is not an exact science, and while your choices may reflect what the majority of the people you make decisions about want to know, it's likely that other individuals will still want and benefit from the explanations you have not prioritised. These will probably also be useful for your own accountability or auditing purposes.

It therefore makes sense that all the explanations you have identified as relevant are made available to the people subject to your AI-assisted decisions. You should consider how to prioritise the remaining explanations based on the contextual factors you identified, and how useful they might be for people.

Speak with colleagues involved in the design/procurement, testing and deployment of AI decision-systems to get their views. If possible, speak with your customers.

2. **Collect the information you need for each explanation type**

For each explanation, it will be useful for you to gather the information you need for its process-based and outcome-based explanation. The process-based explanation will help you explain how your general decision-making is structured for each explanation, while the outcome-based explanation helps you explain what happened in the case of a specific decision. Create your own list of the types of explanations you have determined are relevant and formalise this in a policy or procedure.

3. **Build your rationale explanation to provide meaningful information about the underlying logic of your AI system**

It will be useful to understand the inner workings of your AI system, particularly to be able to comply with certain parts of the GDPR. The model you choose should be at the right level of interpretability for your use case and the impact it will have on the decision recipient. If you use a 'black box' model, make sure the supplementary explanation techniques you use provide a reliable and accurate representation of the system's behaviour.

4. **Translate the rationale of your system's results into useable and easily understandable reasons**

You should determine how you are going to convey your model's statistical results to users and decision recipients as understandable reasons.

A central part of delivering an explanation is communicating how the statistical inferences, which were the basis for your model's output, played a part in your thinking. This involves translating the mathematical rationale of the explanation extraction tools into plain, easily understandable language to justify the outcome.

For example, say your extracted rationale explanation provides you with:

- information about the relative importance of features that influence your model's results; and
- a more global understanding of how this specific decision fits with the model's linear and monotonic constraints.

These factors should then be translated into simple, everyday language that can be understood by non-technical stakeholders. Transforming your model's logic from quantitative rationale into intuitive reasons should lead you to present information as clearly and meaningfully as possible. You could do this through textual clarification, visualisation media, graphical representations, summary tables, or any combination of these.

The main thing is to make sure that there is a simple way to describe or explain the result to an individual. If the decision is fully automated, you may use software to do this. Otherwise this will be through a person who is responsible for translating the result (the implementer – see below).

5. Prepare implementers to deploy your AI system

When human decision-makers are meaningfully involved in an AI-assisted outcome they must be appropriately trained and prepared to use your model's results responsibly and fairly.

Training should include conveying basic knowledge about the nature of machine learning, and about the limitations of AI and automated decision-support technologies. It should also encourage users (the implementers) to view the benefits and risks of deploying these systems in terms of their role in helping humans to come to judgements, rather than replacing that judgement.

If the system is wholly automated and provides a result directly to the decision recipient, it should be set up to provide understandable explanations to them.

6. Consider contextual factors when you deliver your explanation

Consider contextual factors (domain, impact, data, urgency, audience) to help you determine how you should deliver the explanation to the individual.

Again, you may feel that some of the factors we identify (or aspects of them) are simply not relevant to what you do, or that there are additional issues to consider that are unique to the circumstances of your AI model and the decisions it helps you make.

What's important is that you give thought to all the different things that may have an effect on what people will find useful to know about the AI-assisted

decisions you make, and what they might want to do with that knowledge. As a result of this, draw up a list of the relevant factors.

7. Consider how to present your explanation

Finally, you should think about how you will present your explanation of an AI-assisted decision to an individual, whether you are doing this via a website or app, in writing or in person.

A layered approach can be helpful because it presents people with the most relevant information about the decision, while making further explanations easily accessible if they are required. The explanations you have identified as priorities can go in the first layer, while the others can go into a second layer.

You should also think about what information to provide in advance of a decision, and what information to provide to individuals about a decision in their particular case.

Step 1: Select priority explanations by considering the domain, use case and impact on the individual

At a glance

- Getting to know the different types of explanation will help you identify the dimensions of an explanation that decision recipients will find useful.
- In most cases, explaining AI-assisted decisions involves identifying what is happening in your AI system and who is responsible. That means you should prioritise the rationale and responsibility explanation types.
- The setting and sector you are working in is important in figuring out what kinds of explanation you should be able to provide. You should therefore consider domain context and use case.
- In addition, consider the potential impacts of your use of AI to determine which other types of explanation you should provide.
- This will also help you think about how much information is required, and how comprehensive it should be.
- Choosing what to prioritise is not an exact science, and while your choices may reflect what the majority of the people you make decisions about want to know, it's likely that other individuals will still benefit from the explanations you have not prioritised. These will probably also be useful for your own accountability or auditing purposes.

Checklist

☐ We have prioritised rationale and responsibility explanations. We have therefore put in place and documented processes that optimise the end-to-end transparency and accountability of our AI model.

☐ We have considered the setting and sector in which our AI model will be used, and how this affects the types of explanation we provide.

☐ We have considered the potential impacts of our system, and how these affect the scope and depth of the explanation we provide.

In more detail

Introduction

You should consider what types of explanation you need before you start the design process for your AI system, or procurement of a system if you are outsourcing it. You can think of this as 'explanation-by-design'. It involves operationalising the principles we set out in 'The basics of explaining AI'. The following considerations will help you to decide which explanation types you should choose.

Familiarise yourself with the different types of explanation

We introduced the different types of explanation in Part 1 of this guidance, 'The basics of explaining AI'. Making sure you are aware of the range of explanations will provide you with the foundations for considering the different dimensions of an explanation that decision recipients will find useful.

Prioritise rationale and responsibility explanation

It is likely that most explanations of AI-assisted decisions will involve knowing both what your system is doing and who is responsible. In other words, they are likely to involve both rationale and responsibility explanations.

To set up your AI use case to cover these explanations, it is important to consider how you are going to put in place and document processes that:

- optimise the end-to-end transparency and accountability of your AI model. This means making sure your organisation's policies, protocols and procedures are lined up to ensure that when you design and deploy your AI system, you do this in a way that makes it possible to provide clear and accessible process-based explanations; and

- ensure that the intelligibility and interpretability of your AI model is prioritised from the outset. This also means that the explanation you offer to affected individuals appropriately covers the other types of explanation, given the use case and possible impacts of your system.

Consider domain or sector context and use case

When you are trying to work out what kinds of explanation you provide, a good starting point is to consider the setting and sector in which it will be used.

In certain safety-critical/high-stakes and highly regulated domains, sector-specific standards for explanations may largely dictate the sort of information you need to provide to affected individuals.

For instance, AI applications that are employed in safety-critical domains like medicine will have to be set up to provide the safety and performance explanation in line with the established standards and expectations of that sector. Likewise, in a high-stakes setting like criminal justice, where biased decision-making is a significant concern, the fairness explanation will play an important and necessary role.

Understanding your AI application's domain context and setting may also give you useful information about public expectations regarding the content and scope of explanations that have been previously offered in relevant decisions. Doing due diligence and researching these sorts of sector-specific expectations will help you to draw on background knowledge as you weigh which types of AI explanation to include as part of your model's design and implementation processes.

Consider potential impacts

Paying attention to the setting in which your model will be deployed will also put you in a good position to consider its potential impacts. This will be especially useful for selecting your explanations, because it will key you in to

the relevance of impact-specific explanations that should be included as part of your more general explanation of your AI system.

Assessing the potential impact of your AI model on the basis of its use case will help you to determine the extent to which you need to include fairness, safety and performance and more general impact explanations, together with the scope and depth of these types of explanation.

Assessing your AI model's potential impact will also help you understand how comprehensive your explanation needs to be. This includes the risks of deploying the system, and the risks for the person receiving the AI-assisted decision. It will allow you to make sure that the scope and depth of the explanations you are going to be able to offer line up with the real-world impacts of the specific case. For example, an AI system that triages customer service complainers in a luxury goods retailer will have a different (and much lower) explanatory burden than one that triages patients in a hospital critical care unit.

Once you have worked through these considerations, you should choose the most appropriate explanations for your use case (in addition to the rationale and responsibility explanations you have already prioritised). You should document these choices and why you made them.

Prioritise remaining explanations

Once you have identified the other explanations that are relevant to your use case, you should make these available to the people subject to your AI-assisted decisions. You should also document why you made these choices.

See more on the types of explanation in the link below for '[The basics of explaining AI]' and the information you need to put together for each one in Step 2.

> [The basics of explaining AI]

Examples for choosing suitable explanation types

AI-assisted recruitment

An AI system is deployed as a job application filtering tool for a company that is looking to hire someone for a vacancy. This system classifies decision recipients (who receive either a rejection or an invitation to interview) by processing social or demographic data related to individual human attributes and social patterns that are implied in the CVs that have been submitted. A resulting concern might be that bias is 'baked into' the dataset, and that discriminatory features or their proxies might have been used in the model's training and processing. For example, the strong correlation in a dataset between 'all-male' secondary schools attended and successful executive placement in higher paying positions might lead a model trained on this data to discriminate against non-male applicants when it renders recommendations about granting job interviews related to positions of a certain higher paying and executive-level profile.
Which explanation types should you choose in this case?

- **Prioritise rationale and responsibility explanations**: it is highly likely that you will need to include the responsibility and rationale explanations, to tell the individual affected by the AI-assisted hiring decision who is responsible for the decision, and why the decision was reached.
- **Consider domain or sector context and use case**: the recruitment and human resources domain context suggests that bias should be a primary concern in this case.
- **Consider potential impacts**: considering the impact of the AI system on the applicant relates to whether they think the decision was justified, and whether they were treated fairly. Your explanation should be comprehensive enough for the applicant to understand the risks involved in your use of the AI system, and how you have mitigated these risks.
- **Prioritise other explanation types**: This example demonstrates how understanding the specific area of use (the domain) and the particular nature of the data is important for knowing which type of explanation is required for the decision recipient. In this case, a fairness explanation is required because the decision recipient wants to know that they have not been discriminated against. This discrimination could be due to the legacies of discrimination and historical patterns of inequity that may have influenced an AI system trained on biased social and demographic data. In addition, the individual may want an impact explanation to understand how

the recruiter thought about the AI tool's impact on the individual whose data it was processing. A data explanation might also be helpful to understand what data was used to determine whether the candidate would be invited to interview.

AI-assisted medical diagnosis

An AI system utilises image recognition algorithms to support a radiologist to identify cancer in scans. It is trained on a dataset containing millions of images from patient MRI scans and learns by processing billions of corresponding pixels. It is possible that the system may fail unexpectedly when confronted with unfamiliar data patterns or unforeseen environmental anomalies (objects it does not recognise). Such a system failure might lead to catastrophic physical harm being done to an affected patient.

Which explanation types should you choose in this case?

- **Prioritise rationale and responsibility explanations**: it is highly likely that you will need to include the responsibility and rationale explanations, to tell the individual affected by the AI-assisted diagnostic decision who is responsible for the decision, and why the decision was reached.
- **Consider domain or sector context and use case**: the medical domain context suggests that demonstrating the safety and optimum performance of the AI system should be a primary concern in this case.
- **Consider potential impacts**: the impact of the AI system on the patient is high if the system makes an incorrect diagnosis. Your explanation should be comprehensive enough for the patient to understand the risks involved in your use of the AI system, and how you have mitigated these risks.
- **Prioritise other explanation types**: The safety and performance explanation provides justification, when possible, that an AI system is sufficiently robust, accurate, secure and reliable, and that codified procedures of testing and validation have been able to certify these attributes.

Step 2: Collect the information you need for each explanation type

At a glance

- For each type of explanation you should provide:
    - process-based explanations which give you information on the governance of your AI system across its design and deployment; and
    - outcome-based explanations which tell you what happened in the case of a particular decision.
- Your rationale explanation should cover how the system performed and turned inputs into outputs, as well as how the outputs are translated into understandable reasons.
- Your responsibility explanations should identify who is responsible at each stage of the design and deployment of your AI system.
- The data explanation should outline what data was used and why, as well as where it came from.
- Your fairness explanation should reflect that:
    - you made sure the AI system was trained and tested on representative, relevant, accurate and generalisable datasets;
    - you can justify how you built the model architecture;
    - the system does not have a discriminatory effect on those affected by the decision; and
    - the system is deployed by users who are trained to implement it responsibly.
- Safety and performance explanations should cover how you have guaranteed the accuracy, reliability, security and robustness of your system.
- Finally, impact explanations should show how you have considered the impact your AI system has on the individuals affected, as well as wider society.
- The data that you collect and pre-process before inputting it into your system also has an important role to play in the ability to derive each explanation type.

Checklist

☐ We have identified the people within our organisation that are responsible for providing explanations and what exactly they are responsible for.

☐ Our policies, protocols and procedures make it possible to provide clear and accessible process-based explanations when we design and deploy our AI system.

☐ We have considered the setting and sector in which our AI system will be used.

☐ We have considered the potential impacts of our AI system.

☐ We have thought about which other explanation types to include, as well as the depth of the information we will provide in the explanation.

☐ We have documented the information required for process-based and outcome-based explanations for each explanation type.

How collecting and pre-processing the data impacts explanation:

☐ Our data is representative of those we will make decisions about, reliable, relevant and up-to-date.

☐ We have checked with a domain expert to ensure that the data we are using is appropriate and adequate.

☐ We know where the data has come from, the purpose it was originally collected for, and how it was collected.

☐ Where we are using synthetic data, we know how it was created and what properties it has.

☐ We know what the risks are of using the data we have chosen to use, as well as the risks to data subjects of having their data included.

☐ We have labelled the data we are using in our AI system with information including what it is, where it is from, and the reasons why we have included it.

☐ Where we are using unstructured or high-dimensional data, we are clear about why we are doing this and the impact of this on explainability.

☐ We have ensured as far as possible that the data does not reflect past discrimination, whether based explicitly on protected characteristics or possible proxies.

☐ We have mitigated possible bias through pre-processing techniques such as re-weighting, up-weighting, masking, or excluding features and their proxies.

☐ It is clear who within our organisation is responsible for data collection and pre-processing.

In more detail

Building the different explanations

The main aim of explaining fully automated or AI-assisted decisions is justifying a particular result to the individual whose interests are affected by it. In this part, that means making the reasoning behind the outcome of that decision clear, and demonstrating how you were responsible when you chose the processes to design and deploy the system that led to the decision.

We have therefore divided each type of explanation into the subcategories of 'process' and 'outcome':

- **Process-based explanations** of AI systems are about demonstrating that you have followed good governance processes and best practices throughout your design and use.

  For example, if you are trying to explain the fairness and safety of a particular AI-assisted decision, one component of your explanation will involve establishing that you have taken adequate measures across the system's production and deployment to ensure that its outcome is fair and safe.
- **Outcome-based explanations** of AI systems are about clarifying the results of a specific decision. They involve explaining the reasoning behind a particular algorithmically-generated result in plain, easily understandable, and everyday language.

  If there is meaningful human involvement in the decision-making process, you also have to make clear to the affected individual how and why a human judgement that is assisted by an AI output was reached.

  In addition, you may also need to confirm that the actual outcome of an AI decision meets the criteria that you established in your design process to ensure that the AI system is being used in a fair, safe, and ethical way.

The list of explanations below helps you put together the information you will need to be able to build the different explanations.

While we include the rationale explanation here, due to its central importance in AI explanations, we go into further detail in the following sections about how to derive it from a technical perspective. The rationale explanation helps you understand the underlying logic of your AI system, and helps you comply with Articles 13, 14 and 15 of the GDPR.

Rationale explanation

What you need to show

- How the system performed and behaved to get to that decision outcome.
- How the different components in the AI system led it to transform inputs into outputs in a particular way, so you can communicate which features, interactions, and parameters were most significant.

- How these technical components of the logic underlying the result can provide supporting evidence for the decision reached.
- How this underlying logic can be conveyed as easily understandable reasons to decision recipients.
- How you have thought about their impacts on the lives of affected individuals and society.

What information goes into this explanation

- Process-based explanation:
    - Explain how the procedures you have set up help you provide meaningful explanations of the underlying logic of your AI model's results.
    - Ensure that these are appropriate given the model's particular domain context and its possible impacts on the affected decision recipients and wider society.
    - Demonstrate that you have thought about how you are going to set up your AI system and its data collection and pre-processing, model selection, explanation extraction, and explanation delivery procedures so that your system is appropriately interpretable and explainable.

This explanation might answer:

    - Have we selected an algorithmic model, or set of models, that will provide a degree of interpretability that corresponds with its impact on affected individuals?
    - Are the supplementary explanation tools that we are using to help make our complex system explainable good enough to provide meaningful and accurate information about its underlying logic?

- Outcome-based explanation:

    - Explain the formal and logical rationale of the AI system – how the AI system is verified against its formal specifications, so you can verify that the AI system will operate reliably and behave correctly.
    - Explain the technical rationale of the AI system or its output – how the AI model's components (its variables, rules and procedures) transform inputs into outputs, so you know what role these components play in producing the AI system's

output. By understanding the roles and functions of the individual components of the AI system, it is possible to identify the features and parameters that most influence a particular output/decision.

- Explain the translation of the AI system's workings – transforming its input and output variables, parameters and so on into accessible everyday language, so that it becomes clear what role these factors play in reasoning about the real-world problem that the model is trying to address or solve.
- Explain the application of the statistical result to the individual concerned – an application of the reasoning behind the result which takes into account the uniqueness of the specific circumstances, background and personal qualities of affected individuals.
- Explain the justification of the impacts of the use of the AI system, so that you remain accountable both to the individuals about whom you make decisions and to their communities.

The GDPR also makes reference to providing meaningful information about the logic involved in automated decision-making under Articles 13, 14 and 15.

In order to be able to derive your rationale explanation, you need to know how your algorithm works. See Step 3 for more detail about how to do this.

Responsibility explanation

What you need to show

- Identify who is accountable at each stage of the AI system's design and deployment, from defining outcomes for the system at its initial phase of design, through to providing the explanation to the affected individual at the end.
- Define the mechanisms by which each of these people will be held accountable, as well as how you have made the design and implementation processes of your AI system traceable and auditable.

What information goes into this explanation

- Process-based explanation:
  - Detail the roles and functions across your organisation that are involved in the various stages of developing and implementing

your AI decision system, including any human involvement in the decision-making.
    - Explain broadly what the roles do, why they are important, and where overall responsibility lies for management of the AI model – who is ultimately accountable.
    - Explain who is responsible at each step from the design of an AI system through to its implementation to make sure that there is effective accountability throughout.

- Outcome-based explanation:

    - Cover information on how to request a human review of an AI-enabled decision or object to the use of AI, including details on who to contact, and what the next steps will be (eg how long it will take, what the human reviewer will take into account, how they will present their own decision and explanation).
    - Provide a way for individuals to directly contact the role or team responsible for the review. You do not need to identify a specific person in your organisation. One person involved in this should be someone who implemented the decision, and who used the statistical results of a decision-support system to come to a determination about an individual.

Data explanation

What you need to show

- What data was used when you trained your AI system.
- What data you used in a particular decision.

What information goes into this explanation

- Process-based explanation:
    - Detail the source of the training/ test data.
    - Explain how you boost 'explainability' (eg labelling), assess and improve its quality.
    - Explain how you ensure the data is representative.
    - Explain how you ensure bias and discrimination have been mitigated.

- Outcome-based explanation:

    - Clarify the input data used for a specific decision, and the sources of that data.
    - Document the handling and preparation of training and test data, so that a clear and meaningful picture of data handling and use can be provided to affected individuals and other relevant parties.

Fairness explanation

What you need to show

- Dataset fairness: It is trained and tested on properly representative, relevant, accurate, and generalisable datasets.

How?

- Make sure your data sample is representative of all those affected.
- Ensure your data is sufficient in terms of its quantity and quality, so it represents the underlying population and the phenomenon you are modelling.
- Ensure your data is assessed and recorded through suitable, reliable and impartial sources of measurement and has been sourced through sound collection methods.
- Make sure your data is up-to-date and accurately reflects the characteristics of individuals, populations and the phenomena you are trying to model.
- Make sure your data is relevant by calling on domain experts to help you understand, assess and use the most appropriate sources and types of data to serve your objectives.

- Design fairness: It has model architectures that do not include target variables, features, processes, or analytical structures (correlations, interactions, and inferences) which are unreasonable or unjustifiable.

How?

- When defining the problem at the start of the AI project, identify how structural biases can play a factor in translating your objectives into target variables and measurable proxies. These biases could also influence what system designers expect

target variables to measure and what they statistically represent.

- In data pre-processing, take into account the sector or organisational context in which you are operating, as you may introduce bias into your classification process. When this process is automated or outsourced, review what has been done, maintain oversight, and use certification. You should also attach information on the context and metadata to the datasets, so that those coming to the pre-processed data later on have access to the relevant properties when they undertake bias mitigation.
- When you determine which features are relevant as input variables for your model, be aware that the choices you make about grouping or separating and including or excluding features, as well as more general judgements about the comprehensiveness or coarseness of the total set of features, may have consequences for protected groups of people.
- Bias can come in when tuning parameters and setting metrics at the modelling, testing and evaluation stages – ie into the trained model. Your AI development team should iterate the model and peer review it to help ensure that how they choose to adjust the dials and metrics of the model are in line with your objectives of mitigating bias.
- Look out for hidden proxies for discriminatory features in your trained model, as these may act as influences on your model's output. Designers should also look into whether the significant correlations and inferences determined by the model's learning mechanisms are justifiable.

- Outcome fairness: It does not have discriminatory or inequitable impacts on the lives of the people they affect.

How?

- This depends on the definitions of fairness you choose. For example, data scientists can apply different formalised fairness criteria to choose how specific groups in a selected set will receive benefits in comparison to others in the same set, or how the accuracy or precision of the model will be distributed among subgroups. This can be done by reweighting model parameters; embedding trade-offs in a classification procedure; or re-tooling algorithmic results to adjust for outcome preferences.

- Implementation fairness: It is deployed by users sufficiently trained to implement it responsibly and without bias.

How?

  - To avoid automation bias (over-relying on the outputs of AI systems) or automation-distrust bias (under-relying on AI system outputs because of a lack of trust in them) you should train implementers of AI system outputs on how to use them in the specific context in which they are being used. That is, they should understand the individual circumstances of the individual to which that output is being applied.

What information goes into this explanation

This explanation is about providing people with appropriately simplified and concise information on the considerations, measures and testing you carry out. Fairness considerations come into play through the whole lifecycle of an AI model, from inception to deployment, monitoring and review.

- Process-based explanation:
  - Detail your chosen measures to mitigate risks of bias and discrimination at the data collection, preparation, model design and testing stages.
  - Detail the results of your initial (and ongoing) fairness testing and external validation – proving that your chosen fairness measures are working in practice. You could do this by showing that different groups of people receive similar outcomes, or that protected characteristics have not played a factor in the results.

- Outcome-based explanation:
  - Explain how your organisation has decided to define fairness by the criteria it has selected in its formal model(s). It should then be possible to explain how these fairness criteria were implemented in the case of a particular decision or output.
  - Include the relevant fairness metrics and performance measurements in the delivery interface of your model.
  - Explain how others similar to the individual were treated, ie whether they received the same decision outcome as the individual. For example, you could use information generated from counter-factual scenarios to show whether or not someone with similar characteristics, but of a different ethnicity or

gender, would receive the same decision outcome as the individual.

Safety and performance explanation

What you need to show

- Accuracy: the proportion of examples for which your model generates a correct output. This component may also include other related performance measures such as precision, sensitivity (true positives), and specificity (true negatives). Individuals may want to understand how accurate, precise, and sensitive the output was in their particular case.
- Reliability: how dependably the AI system does what it was intended to do. If it did not do what it was programmed to carry out, individuals may want to know why, and whether this happened in the process of producing the decision that affected them.
- Security: the system is able to protect its architecture from unauthorised modification or damage of any of its component parts; the system remains continuously functional and accessible to its authorised users and keeps confidential and private information secure, even under hostile or adversarial conditions.
- Robustness: the system functions reliably and accurately under harsh conditions. Individuals may want to know how well the system works when things go wrong, how this has been anticipated and tested, and how the system has been immunised from adversarial attacks.

What information goes into this explanation

- Process-based explanation:
    - Accuracy: how you measure it and why you chose those measures, eg maximising precision to reduce the risk of false negatives; what you did at the data collection stage to ensure your training data was up-to-date and reflective of the characteristics of the people you are now making AI-assisted decisions about; what kinds of external validation you have undertaken to test and confirm your model's accuracy; what the overall accuracy rate of the system was at testing stage, and what you do to monitor this (eg measuring for concept drift over time).
    - Reliability: how you measure it and why you chose those measures, which helps the individual to understand how

confident you are in the system's consistency and therefore its safety.

- Security: how you measure it and why you chose those measures, eg who is able to access the AI system; how you manage the security of confidential and private information.
- Robustness: how you measure it and why you chose those measures, eg how you've stress-tested the system to understand how it responds to adversarial intervention, implementer error, or skewed goal-execution by an automated learner (in reinforcement learning applications).
- Summarise the type of AI model(s) used (eg decision tree, random forest, neural network), the AI software, software development kit, or programme used (eg TensorFlow, scikit-learn, H2O), and the technical approach you use to extract rationale explanations from your model (eg sensitivity analysis, SHAP, LIME).

- Outcome-based explanation:

It is unlikely (or even, in some cases, impossible) for you to be able to conduct testing on the accuracy of your AI model's predictions or classifications at the individual level (eg for particular decisions).

- In the case of accuracy and the other performance metrics, however, you should include in your model's delivery interface the results of your cross-validation (training/testing splits) and any external validation carried out.
- You may also include relevant information related to your system's confusion matrix (the table that provides the range of performance metrics) and ROC curve (receiver operating characteristics)/AUC (area under the curve) – with guidance for users and affected individuals that makes the meaning of these measurement methods, and specifically the ones you have chosen to use, easily accessible and understandable. This should also include a clear representation of the uncertainty of the results (eg confidence intervals and error bars).
- Provide information for components other than accuracy that confirms that the AI system operated securely, reliably, and in accordance with its intended design in the case of a specific decision.

Impact explanation

What you need to show

- Demonstrate that you have thought about how your AI system will potentially affect individuals and wider society, and make the process you have gone through to determine these possible impacts plain to affected individuals.

What information goes into this explanation

- Process-based explanation:
    - Summarise the considerations you made, how you made them, and the measures and steps you took to mitigate possible negative effects on society, and to amplify the positive effects.
    - Include information about how you plan to monitor and re-assess impacts while your system is deployed.
- Outcome-based explanation:
    - Explain the intention and purpose behind the AI model – you should say what the system is being used to help make decisions about and what you were optimising for when designing and developing it.
    - Explain the consequences for the individual of the different possible decision outcomes, eg if the decision has favourable and unfavourable outcomes, what will this mean for the individual, and what happens next for them?
    - Explain the impacts on the wellbeing of wider society – be explicit about the considerations you have made regarding the effect of your system on both communities in which it is being deployed and society as whole.

How collecting and pre-processing data impacts the explanation

How you collect and pre-process the data you use in your chosen model has a bearing on the quality of the explanation you can offer to decision recipients. Below we set out some of the things you should think about, and how this can contribute to the information you provide to individuals for each explanation type.

Rationale

Understanding the logic of an AI model, or of a specific AI-assisted decision, is much simpler when the features (the input variables from which the model draws inferences and that influence a decision) are already interpretable by humans, for example, someone's age or location. Limit your pre-processing of that data so that it isn't transformed through extensive feature engineering into more abstract features that are difficult for humans to understand.

Careful, transparent, and well-informed data labelling practices will set up your AI model to be maximally interpretable. If you are using data that is not already naturally labelled, there will be a stage at which you will have humans labelling the data with relevant information. At this stage you should ensure that the information recorded is as rich and meaningful as possible. Ask those charged with labelling data to not only tag and annotate what a piece of data is, but also the reasons for that tag. For example, rather than 'this x-ray contains a tumour', say 'this x-ray contains a tumour because…'. Then, when your AI system classifies new x-ray images as tumours, you will be able to look back to the labelling of the most similar examples from the training data to contribute towards your explanation of the decision rationale.

Of course, all of the above isn't always possible. The domain in which you wish to use AI systems may require the collection and use of unstructured, high-dimensional data (where there are countless different input variables interacting with each other in complex ways).

In these cases, you should justify and document the need to use such data. You should also use the guidance in the next step to assess how best to obtain an explanation of the rationale through appropriate model selection and approaches to explanation extraction.

Responsibility

Responsibility explanations are about telling people who, or which part of your organisation, is responsible for overall management of the AI model. This is primarily to make your organisation more accountable to the individuals it makes AI-assisted decisions about.

But you may also want to use this as an opportunity to be more transparent with people about which parts of your organisation are responsible for each stage of the development and deployment process, including data collection and preparation.

Of course, it may not be feasible for your customers to have direct contact with these parts of your organisation (depending on your organisation's size and how you interact with customers). But informing people about the different business functions involved will make them more informed about the process. This may increase their trust and confidence in your use of AI-assisted decisions because you are being open and informative about the whole process.

If you are adopting a layered approach to the delivery of explanations, it is likely that this information will sit more comfortably in the second or third layer – where interested individuals can access it, without overloading others with too much information. See Step 7 for more on layering explanations.

Data

The data explanation is, in part, a catch-all for giving people information about the data used to train your AI model.

There is a lot of overlap therefore with information you may already have included about data collection and preparation in your rationale, fairness and safety and performance explanations.

However, there are other aspects of the data collection and preparation stage, which you could also include. For example:

- the source of the training data;
- how it was collected;
- assessments about its quality; and
- steps taken to address quality issues, such as completing or removing data.

While these may be more procedural aspects (less directly linked to key areas of interest such as fairness and accuracy) there remains value in providing this information to people. As with the responsibility explanation, the more insight individuals have on the AI model that makes decisions about them, the more confident they are likely to be in interacting with these systems and trusting your use of them.

Fairness

Fairness explanations are about giving people information on the steps taken to mitigate risks of discrimination both in the production and implementation of your AI system and in the results it generates. They shed light on how individuals have been treated in comparison to others. Some of the most

important steps to mitigate discrimination and bias arise at the data collection stage.

For example, when you collect data, you should have a domain expert to assess whether it is sufficiently representative of the people you will make AI-assisted decisions about.

You should also consider where the data came from, and assess to what extent it reflects past discrimination, whether based explicitly on protected characteristics such as race, or on possible proxies such as post code. You may need to modify the data to avoid your AI model learning and entrenching this bias in its decisions. Pre-processing techniques such as re-weighting, up-weighting, masking, or even excluding features may be used to mitigate implicit discrimination in the dataset and to prevent bias from entering into the training process. If you exclude features, you should also ensure that you exclude proxies or related features.

Considerations and actions such as these, that you take at the data collection and preparation stages, should feed directly into the fairness explanations you give to individuals. Ensure that you appropriately document what you do at these early stages so you can reflect this in your explanation.

Safety and performance

The safety and performance explanation is concerned with the actions and measures you take to ensure that your AI system is accurate, secure, reliable and robust.

The accuracy component of this type of explanation is mainly concerned with the actions and measures you take at the modelling, testing, and monitoring stages of developing an AI model. It involves providing people with information about the accuracy rate of a model, and about the various accuracy related measures you used.

Impact

The impact explanation involves telling people about how an AI model, and the decisions it makes, may impact them as individuals, communities, and members of wider society. It involves making decision recipients aware of what the possible positive and negative effects of an AI model’s outcomes are for people taken singly and as a whole. It also involves demonstrating that your organisation has put appropriate forethought into mitigating any potential harm and pursuing any potential societal benefits.

Information on this will come from considerations you made as part of your impact or risk assessment (eg a data protection impact assessment). But it will also come from the practical measures you took throughout the development and deployment of the AI model to act on the outcome of the impact assessment.

This includes what you do at the data collection and preparation stage to mitigate risks of negative impact and amplify the possibility of positive impact on society.

While you may have covered such steps in your fairness and accuracy explanations (eg ensuring the collection of representative and up-to-date datasets), the impact explanation type is a good opportunity to clarify in simple terms how this affects the impact on society (eg by reducing the likelihood of systematic disadvantaging of minority groups, or improving the consistency of decision-making for all groups).

> For an introduction to the explanation types, see 'The basics of explaining AI'. For further details on how to take measures to ensure these kinds of fairness in practice and across your AI system's design and deployment, see the fairness section of Understanding Artificial Intelligence Ethics and Safety, a guidance produced by the Office for AI, the Government Digital Service, and The Alan Turing Institute.

Step 3: Build your rationale explanation to provide meaningful information about the underlying logic of your AI system

At a glance

- Deriving the rationale explanation is key to understanding your AI system and helps you comply with parts of the GDPR. It requires looking 'under the hood' and helps you gather information you need for some of the other explanations, such as safety and performance and fairness. However, this is a complex task that requires you to know when to use more and less interpretable models and how to understand their outputs.
- To choose the right AI model for your explanation needs, you should think about the domain you are working in, and the potential impact of the deployment of your system on individuals and society.
- Following this, you should consider:
    - the costs and benefits of replacing your current system with a newer and potentially less explainable AI model;
    - whether the data you use requires a more or less explainable system;
    - whether your use case and domain context encourage choosing an inherently interpretable system;
    - if your processing needs lead you to select a 'black box' model; and
    - whether your supplementary interpretability tools are appropriate in your context.
- To extract explanations from inherently interpretable models, look at the logic of the model's mapping function by exploring it and its results directly.
- To extract explanations from 'black box' systems, there are many techniques you can use. Make sure that they provide a reliable and accurate representation of the system's behaviour.

Checklist

Selecting an appropriately explainable model:

☐ We know what the interpretability/transparency expectations and requirements are in our sector or domain.

☐ In choosing our AI model, we have taken into account the specific type of application and the impact of the model on decision recipients.

☐ We have considered the costs and benefits of replacing the existing technology we use with an AI system.

☐ Where we are using social or demographic data, we have considered the need to choose a more interpretable model.

☐ Where we are using biophysical data, for example in a healthcare setting, we have weighed the benefits and risks of using opaque or less interpretable models.

☐ Where we are using a 'black box' system, we have considered the risks and potential impacts of using them.

☐ Where we are using a 'black box' system we have also determined that the case we will use it for and our organisational capacity both support the responsible design and implementation of these systems.

☐ Where we are using a 'black box' system we have considered which supplementary interpretability tools are appropriate for our use case.

☐ Where we are using 'challenger' models alongside more interpretable models, we have established that we are using them lawfully and responsibly, and we have justified why we are using them.

☐ We have considered how to measure the performance of the model and how best to communicate those measures to implementers, and decision recipients.

☐ We have mitigated any bias we have found in the model.

☐ We have made it clear how the model has been tested, including which parts of the data have been used to train the model, and which have been used to test it, and which have formed the holdout data.

☐ We have a record of each time the model is updated, how each version has changed, and how this affects the model's outputs.

☐ It is clear who within our organisation is responsible for validating the explainability of our AI system.

Tools for extracting a rationale explanation:

All the explanation extraction tools we use:

☐ Convey the model's results reliably and clearly.

☐ Help implementers of AI-assisted decisions to exercise better-informed judgements.

☐ Offer affected individuals plausible and easily understandable accounts of the logic behind the model's output.

For interpretable AI models:

☐ We are confident in our ability to extract easily understandable explanations from models such as regression-based and decision/rule-based systems, Naïve Bayes, and K nearest neighbour.

For supplementary explanation tools to interpret 'black box' AI models:

☐ We are confident that they are suitable for our application.

☐ We recognise that they will not give us a full picture of the opaque model.

☐ In selecting the supplementary tool, we have prioritised the need for it to provide a reliable, accurate and close approximation of the logic behind our AI system's behaviour, for both local and global

explanations.

Combining supplementary explanation tools to produce meaningful information about your AI system's results:

☐ We have included a visualisation of how the model works.

☐ We have included an explanation of variable importance and interaction effects, both global and local.

☐ We have included counterfactual tools to explore alternative possibilities and actionable recourse.

In more detail

Introduction

The rationale explanation is key to understanding your AI system and helps you comply with parts of the GDPR. It requires detailed consideration because it is about how the AI system works, and can help you obtain an explanation for the underlying logic of the AI model you decide to use.

Selecting an appropriately explainable model

Where do we start?

Before you consider the technical factors, you should consider:

Domain: Consider the specific standards, conventions, and requirements of the domain in which your AI system will be applied.

For example, in the financial services sector, rigorous justification standards for credit and loan decisions largely dictate the need to use fully transparent and easily understandable AI decision-support systems. Likewise, in the medical sector, rigorous safety standards largely dictate the extensive levels of performance testing, validation and assurance that are demanded of

treatments and decision-support tools. Such domain specific factors should actively inform the choices you make about model complexity and interpretability.

Impact: Think about the type of application you are building and its potential impacts on affected individuals.

For example, there is a big difference between a computer vision system that sorts handwritten employee feedback forms and one that sorts safety risks at a security checkpoint. Likewise, there is also a difference between a complex random forest model that triages applicants at a licensing agency and one that triages sick patients in an accident and emergency department.

Higher-stakes or safety-critical applications will require you to be more thorough in how you consider whether prospective models can appropriately ensure outcomes that are non-discriminatory, safe, and supportive of individual and societal wellbeing.

Low-stakes AI models that are not safety-critical, do not directly impact the lives of people, and do not process potentially sensitive social and demographic data are likely to mean there is less need for you to dedicate extensive resources to developing an optimally performing but highly interpretable system.

Draw on the appropriate domain knowledge, policy expertise and managerial vision in your organisation. You need to consider these when your team is looking for the best-performing model.

What are the technical factors that we should consider when selecting a model?

You should also discuss a set of more technical considerations with your team before you select a model.

Existing technologies: consider the costs and benefits of replacing current data analysis systems with newer systems that are possibly more resource-intensive and less explainable.

One of the purposes of using an AI system might be to replace an existing algorithmic technology that may not offer the same performance level as the more advanced machine learning techniques that you are planning to deploy.

In this case, you may want to carry out an assessment of the performance and interpretability levels of your existing technology. This will provide you with a baseline against which you can compare the trade-offs of using a more advanced AI system. This could also help you weigh the costs and benefits of building or using a more complex system that requires more support for it to be interpretable, in comparison to the costs and benefits of using a simpler model.

It might also be helpful to look into which AI systems are being used in your application area and domain. This should help you to understand the resource demands that building a complex but appropriately interpretable system will place on your organisation.

> For more information on the trade-offs involved in using AI systems, see the ICO's AI Auditability Framework blogpost on trade-offs.

Data: integrate a comprehensive understanding of what kinds of data you are processing into considerations about the viability of algorithmic techniques.

To select an appropriately explainable model, you need to consider what kind of data you are processing and what you are processing it for.

There are two groups of data that it is helpful to consider:

i. Data that refers to demographic characteristics, measurements of human behaviour, social and cultural characteristics of people.
ii. Biological or physical data, such as biomedical data used for research and diagnostics (ie data that does not refer to demographic characteristics or measurements of human behaviour);

With these in mind, there are certain things to consider:

- In cases where social or demographic data (group i. above) is being processed you may come across issues of bias and discrimination. Here, you should prioritise selecting an optimally interpretable model, and avoid 'black box' systems.
- More complex systems may be appropriate in cases where biological or physical data (group ii. above) is being processed, only for the purposes of gaining scientific insight (eg radiological diagnostics), or

operational functionality (eg computer vision for vehicle navigation). However, where the application is high impact or safety-critical, you should weigh the safety and performance (accuracy, security, reliability and robustness) of the AI system heavily in selecting the model. Note, though, that bias and discrimination issues may arise in processing biological and physical data, for example in the representativeness of the datasets on which these models are trained and tested.

- In cases where both these groups of data are being processed and the processing directly affects individuals, you should consider concerns about both bias and safety and performance when you are selecting your model.

Another distinction you should consider is between conventional data (eg a person's payment history or length of employment at a given job) and unconventional data (eg sensor data – whether raw or interlinked with other data to generate inferences – collected from a mobile phone's gyroscope, accelerometer, battery monitor, or geolocation device or text data collected from social media activity).

In cases where unconventional data is being used to support decisions that affect individuals, you should bear the following in mind:

- This data can be considered to be of the same type as group i. data above, and treated the same way (as it gives rise to the same issues).
- You should select transparent and explainable AI systems that yield interpretable results, rather than black box models.
- You can justify its use by indicating what attribute the unconventional data represents in its metadata, and how such an attribute might be a factor in evidence-based reasoning.

  For example, if GPS location data is included in a system that analyses credit risk, the metadata must indicate what interpretively significant feature such data is supposed to indicate about the individual whose data is being processed.

Interpretable algorithms: when possible and application-appropriate, draw on standard and maximally interpretable algorithmic techniques.

In high impact, safety-critical or other potentially sensitive environments, you are likely to need an AI system that maximises accountability and

transparency. In some cases, this will mean you prioritise choosing standard but sophisticated non-opaque techniques.

These techniques (some of which are outlined in the table below) may include decision trees/rule lists, linear regression and its extensions like generalised additive models, case-based reasoning, or logistic regression. In many cases, reaching for the 'black box' model first may not be appropriate and may even lead to inefficiencies in project development. This is because more interpretable models are also available, which perform very well but do not require supplemental tools and techniques for facilitating interpretable outcomes.

Careful data pre-processing and iterative model development can hone the accuracy of interpretable systems in ways that may make the advantages gained by the combination of their performance and transparency outweigh the benefits of less transparent approaches.

'Black box' AI systems: when considering the use of opaque algorithmic techniques, make sure that the supplementary interpretability tools that will be used to explain the model are appropriate to meet the domain-specific risks and explanatory needs that may arise from deploying it.

For certain data processing activities it may not be feasible to use straightforwardly interpretable AI systems. For example, the most effective machine learning approaches will likely be opaque when AI applications are sought for classifying images, recognising speech, or detecting anomalies in video footage. The feature spaces of these kinds of AI systems grow exponentially to hundreds of thousands or even millions of dimensions. At this scale of complexity, conventional methods of interpretation no longer apply.

For clarity, we define a 'black box' model as any AI system whose inner workings and rationale are opaque or inaccessible to human understanding. These systems may include neural networks (including recurrent and convolutional neural nets), ensemble methods (an algorithmic technique such as the random forest method that strengthens an overall prediction by combining and aggregating the results of several or many different base models), and support vector machines (a classifier that uses a special type of mapping function to build a divider between two sets of features in a high dimensional space). The main kinds of opaque models are described in more detail below.

You should only use 'black box' models if you have thoroughly considered their potential impacts and risks in advance, and the members of your team have determined that your use case and your organisational capacities/resources support the responsible design and implementation of these systems.

Likewise, you should only use them if supplemental interpretability tools provide your system with a domain-appropriate level of explainability that is reasonably sufficient to mitigate its potential risks and to provide a solid basis for providing affected decision recipients with meaningful information about the rationale of any given outcome. A range of the supplementary techniques and tools that assist in providing some access to the underlying logic of 'black box' models is explored below.

As part of the process-based aspect of the rationale explanation of your AI system, you should document and keep a record of any deliberations that go into your organisation's selection of a 'black box' model.

Hybrid methods – use of 'challenger' models: in cases where an interpretable model is selected to ensure explainable data processing, you should carry out parallel use of opaque 'challenger' models for purposes of feature engineering/selection, insight, or comparison in a transparent, responsible, and lawful manner.

Our research has shown that, while some organisations in highly regulated areas like banking and insurance are continuing to select interpretable models in their customer-facing AI decision-support applications, they are increasingly using more opaque 'challenger' models alongside these, for the purposes of feature engineering/selection, comparison, and insight.

'Black box' challenger models are trained on the same data that trains transparent production models and are used both to benchmark the latter, and in feature engineering and selection.

When challenger models are employed to craft the feature space, ie to reduce the number of variables (feature selection) or to transform/combine/bucket variables (feature engineering), they can potentially reduce dimensionality and show additional relationships between features. They can therefore increase the interpretability of the production model.

If you use challenger models for this purpose, the process should be made explicit and documented. Moreover, any highly engineered features that are

drawn from challenger models and used in production models must be properly justified and annotated in the metadata to indicate what attribute the combined feature represents and how such an attribute might be a factor in evidence-based reasoning.

When challenger models are used to process the data of affected decision recipients – even for benchmarking purposes – they should be properly recorded and documented. If the insights from this challenger model's processing are incorporated into any dimension of actual decision-making (for instance, the comparative benchmarking results are shared with implementers/users, who are making decisions), you should treat them as core production models, document them, and hold them to the same explainability standards.

What types of models are we choosing between?

To help you get a better picture of the spectrum of algorithmic techniques, the following table lays out some of the basic properties, potential uses, and interpretability characteristics of the most widely used algorithms at present.

The first eleven techniques listed are considered to be largely interpretable, although for some of them, like the regression-based and tree-based algorithms, this depends on the number of input features that are being processed. The final four techniques are more or less considered to be 'black box' algorithms.

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	High level of interpretability because of linearity and monotonicity. Can become less interpretable with increased number of features (ie high dimensionality).

		simpler to calculate and have oversight over.	
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	This extension of LR is applicable to use cases where target variables have constraints that require the exponential family	Good level of interpretability that tracks the advantages of LR while also introducing more flexibility.

	introduces a link function that allows for the extension of LR to non-normal distributions.	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).	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 (i.e 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	Because the step-by-step logic that produces DT outcomes is easily understandable to non-technical users (depending on number of	High level of interpretability if the DT is kept manageably small, so that the logic can be followed end-to-end. The

	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.	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.	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	Using exemplars drawn from prior	CBR is applicable in any domain	CBR is interpretable-by-

(CBR)/ Prototype and criticism	human knowledge, CBR 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.	where experience-based 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.	design. It uses examples drawn from 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	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	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

	accessible model, it can be implemented quickly and efficiently by non-technical users, who need no special training to deploy the system.	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.	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 membership of that feature by multiplying the prior probability of the class with the class conditional probability of the feature.	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 feature spaces. Common applications include spam filtering, recommender systems, and sentiment analysis.	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 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	KNN is a simple, intuitive, versatile technique that has wide applications but works best with smaller	KNN works off the assumption that classes or outcomes can be predicted by looking at the

	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.	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.	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 trained networks of interconnected and multi-layered operations.	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, pharmaceutical discovery, and stock prediction to machine translation, disease diagnosis, and fraud detection.	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 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	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

		and have also been used extensively in computational linguistics, econometrics, and predictive risk modelling.	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.

> Further reading on algorithm types

Tools for extracting explanations

Extracting and delivering meaningful explanations about the underlying logic of your AI model's results involves both technical and non-technical components.

At the technical level, to be able to offer an explanation of how your model reached its results, you need to:

- become familiar with how AI explanations are extracted from intrinsically interpretable models;
- get to know the supplementary explanation tools that may be used to shed light on the logic behind the results and behaviours of 'black box' systems; and
- learn how to integrate these different supplementary techniques in a way that will enable you to provide meaningful information about your system to its users and decision recipients.

At the non-technical level, extracting and delivering meaningful explanations involves establishing how conveying your model's results can reliably and clearly enable users and implementers to:

- exercise better-informed judgements; and
- offer plausible and easily understandable accounts of the logic behind its output to affected individuals and concerned parties.

Technical dimensions of AI interpretability

Before going into detail about how to set up a strategy for explaining your AI model, you need to be aware of a couple of commonly used distinctions that will help you and your team to think about what is possible and desirable for an AI explanation.

- Local vs global explanation

The distinction between the explanation of single instances of a model's results and an explanation of how it works across all of its outputs is often characterised as the difference between local explanation and global

explanation. Both types of explanation offer potentially helpful support for providing significant information about the rationale behind an AI system's output.

A **local explanation** aims to interpret individual predictions or classifications. This may involve identifying the specific input variables or regions in the input space that had the most influence in generating a particular prediction or classification.

Providing a **global explanation** entails offering a wide-angled view that captures the inner-workings and logic of that model's behaviour in sum and across predictions or classifications. This kind of explanation can capture the overall significance of features and variable interactions for model outputs and significant changes in the relationship of predictor and response variables across instances. It can also provide insights into dataset-level and population-level patterns, which are crucial for both big picture and case-focused decision-making.

- Internal/model intrinsic vs. external/post-hoc explanation

Providing an **internal or model intrinsic explanation** of an AI model involves making the way its components and relationships function intelligible. It is therefore closely related to, and overlaps to some degree with, global explanation - but it is not the same. An internal explanation makes insights available about the parts and operations of an AI system **from the inside**. These insight can help your team understand why the trained model does what it does, and how to improve it.

Similarly, when this type of internal explanation is applied to a 'black box model', it can shed light on that opaque model's operation by breaking it down into more understandable, analysable, and digestible parts. For example, in the case of an artificial neural network (ANN) it can break it down into interpretable characteristics of its vectors, features, interactions, layers, parameters etc. This is often referred to as 'peeking into the black box'.

Whereas internal explanations can be drawn from both interpretable and opaque AI systems, **external or post-hoc explanations** are more applicable to 'black box' systems where it is not possible to fully access the internal underlying rationale due to the model's complexity and high dimensionality.

Post-hoc explanations attempt to capture essential attributes of the observable behaviour of a 'black box' system by subjecting it to a number of different techniques that reverse-engineer explanatory insights. Post-hoc approaches can do a number of different things:

- test the sensitivity of the outputs of an opaque model to perturbations in its inputs;
- allow for the interactive probing of its behavioural characteristics; or
- build proxy-based models that utilise simplified interpretable techniques to gain a better understanding of particular instances of its predictions and classifications, or of system behaviour as a whole.

Getting familiar with AI explanations through interpretable models

For AI models that are basically interpretable (such as regression-based and decision/rule-based systems, Naïve Bayes, and K nearest neighbour), the technical aspect of extracting a meaningful explanation is relatively straightforward - draw on the intrinsic logic of the model's mapping function by looking directly at it and at its results.

For instance, in decision trees or decision/rule lists, the logic behind an output will depend on the interpretable relationships of weighted conditional (if-then) statements. In other words, each node or component of these kinds of models is, in fact, operating **as a reason**. Extracting a meaningful explanation from them therefore factors down to following the path of connections between these reasons.

Note, though, that if a decision tree is excessively deep or a given decision list is overly long, it will be challenging to interpret the logic behind their outputs. Human-scale reasoning, generally speaking, operates on the basis of making connections between only a few variables at a time, so a tree or a list with thousands of features and relationships will be significantly harder to follow and thus less interpretable. In these more complex cases, an interpretable model may lose much of its global as well as its local explainability.

Similar advantages and disadvantages have long been recognised in the explainability of regression-based models. Clear-cut interpretability has made this class of algorithmic techniques a favoured choice in high-stakes and highly regulated domains because many of them possess linearity, monotonicity, and sparsity/non-complexity:

Characteristics of regression-based models that allow for optimal explainability and transparency

- **Linearity:** Any change in the value of the predictor variable is directly reflected in a change in the value of the response variable at a constant rate b. The interpretable prediction yielded by the model can therefore be directly inferred from the relative significance of the parameter/weights of the predictor variable and have high inferential clarity and strength.
- **Monotonicity:** When the value of the predictor changes in a given direction, the value of the response variable changes consistently either in the same or opposite direction. The interpretable prediction yielded by the model can therefore be directly inferred. This monotonicity dimension is a highly desirable interpretability condition of predictive models in many heavily regulated sectors, because it incorporates reasonable expectations about the consistent application of sector specific selection constraints into automated decision-making systems.
- **Sparsity/non-complexity**: The number of features (dimensionality) and feature interactions is low enough and the model of the underlying distribution is simple enough to enable a clear understanding of the function of each part of the model in relation to its outcome.

In general, it is helpful to get to know the range of techniques that are available for the explanation of interpretable AI models such as those listed above. These techniques not only make the rationale behind models like logic-based decision trees, lists, and sets, and of regression-based systems readily interpretable; they also form the basis of many of the supplementary explanation tools that are widely used to make 'black box' models more interpretable.

Technical strategies for explaining 'black box' AI models through supplementary explanation tools

If, after considering domain, impact, and technical factors, you have chosen to use a 'black box' AI system, your next step is to incorporate appropriate supplementary explanation tools into building your model.

There is no comprehensive or one-size-fits-all technical solution for making opaque algorithms interpretable. The supplementary explanation strategies available to support interpretability may shed light on significant aspects of a model's global processes and components of its local results.

However, often these strategies operate as imperfect approximations or as simpler surrogate models, which do not fully capture the complexities of the original opaque system. This means that overreliance on supplementary tools may be misleading.

With this in mind, 'fidelity' may be a suitable primary goal for your technical 'black box' explanation strategy. In order for your supplementary tool to achieve a high level of fidelity, it should provide a reliable and accurate approximation of the system's behaviour.

For practical purposes, you should think both locally and globally when choosing the supplementary explanation tools that will achieve fidelity.

Thinking locally is a priority, because the primary concern of AI explainability is to make the results of specific data processing activity clear and understandable to affected individuals. This is local explanation.

Even so, it is just as important to provide supplementary global explanations of your AI system. Understanding the relationship between your system's component parts (its features, parameters, and interactions) and its behaviour as a whole will often be a critical factor in setting up an accurate local explanation. It will also be essential to securing your AI system's fairness, safety and optimal performance.

This sort of global understanding may also provide crucial insights into your model's more general potential impacts on individuals and wider society, as well as allow your team to improve the model, so that concerns raised by such global insights can be properly addressed.

Below we provide you with a table containing some of the more widely used supplementary explanation strategies and tools. Keep in mind, though, that this is a rapidly developing field, so remaining up to date with the latest tools will mean that you and technical members of your team will need to move beyond the basic information we are offering here.

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 R^2 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	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

	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.	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. 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	Refining and extending PDP's, ICE plots graph	When used in combination with PDP's, ICE plots can

Expectations Plot (ICE)	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.	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. 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. ALE plots are also more computationally	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. 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.

	tractable than PDP's because they are able to use techniques to compute effects in smaller intervals and chunks of observations.	
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 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	While permuting variables to measure their relative importance, to some extent, accounts for interaction effects, there is still a 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. 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

	is being explained.	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 interactions between	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. 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. 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

	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.	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	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. 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

	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.	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.	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. 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

	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.	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. 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?	LIME offers a local and post-hoc form of supplementary	

Internal/post-hoc?	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. 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	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. 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. 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

	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 computed. 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.	when conditional contributions are approximated. The result is the introduction of uncertainty into the explanation that is produced, because the 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. 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. 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 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	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. Moreover, there are as yet limitations on the types of datasets and functions to which these kinds of explanations are applicable. 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

	with concise options for making practical changes that would improve their chances of obtaining the desired outcome. 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.	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,	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

	like a built-in recurrent neural net with an 'attention-directing' mechanism translates the extracted features into a natural language representation that produces a sentence-long explanation of the result to the user. 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.	evaluators and be understandable to affected individuals. Self-explaining systems, in other words, should themselves remain optimally interpretable. The task of formulating a primary strategy of supplementary explanation is still part of the process of building out a system with self-explaining capacity. 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.	

> Further reading on supplementary techniques

Combining and integrating supplementary explanation strategies

The main purpose of using supplementary explanation tools is to make the underlying rationale of the results of an AI system's data processing both optimally interpretable and more easily intelligible to those who use the system and to decision recipients.

For this reason, it is a good idea to think about using different explanation strategies in concert. You can combine explanation tools to enable affected individuals to make sense of the reasoning behind an AI-assisted decision with as much clarity and precision as possible.

With this in mind, it might be helpful to think about how to combine these different strategies into a portfolio of tools for explanation extraction that best serves the needs of your particular AI system, and that is most appropriate for providing meaningful information about the rationale of its results.

Keeping in mind the various strategies we have introduced in the table above, there are three interrelated layers of technical rationale that might be considered as especially significant components to include in such a portfolio:

- visualisation of how the model works;
- explanation of variable importance and interaction effects, both global and local; and
- counterfactual tools to explore alternative possibilities and actionable recourse.

Here are some questions that may assist you in thinking about how to integrate these layers of explanation extraction:

Visualisation of how the model works

How might graphical tools like ALE plots or a combination of PDP's and ICE plots make the logic behind both the global and the local behaviour of our model clearer to users, implementers, auditors and decision recipients? How might these tools be used to improve the model and to ensure that it operates in accordance with reasonable expectations?

How can domain knowledge and understanding the use case inform the insights derived from visualisation techniques? How might this knowledge

inform the integration of visualisation techniques with other explanation tools?

What are the most effective ways that such visualisations can be presented and explained to users and decision recipients so as to help them build a mental model of how the system works, both as a whole and in specific instances? How can they be used to enhance evidence-based reasoning?

Are other visualisation techniques available (like heat maps, interactive querying tools for ANN's, or more traditional 2D tools like principle components analysis) that would also be helpful to enhance the interpretability of our system?

Understanding of the role of variables and variable interactions

How can global measures of feature importance and feature interactions be utilised to help users and decision recipients better understand the underlying logic of the model as a whole?

How might they provide reassurance that the model is yielding results that are in line with reasonable expectations?

How might they support and enhance the information being provided in the visualisation tools?

How might measures of variable importance and interaction effects be used to confirm that our AI system is operating fairly and is not harming or discriminating against affected stakeholders?

Which local, post-hoc explanation tools—like LIME, SHAP, LOCO (Leave-One-Covariate-Out), etc—are reliable enough in the context of our particular AI system to be useful as part of its portfolio of explanation extraction tools?

Have we established through model exploration and testing that using these local explanation tools will help us to provide meaningful information that is informative rather than misleading or inaccurate?

Understanding of how the behaviours or circumstances that influence an AI-assisted decision would need to be changed to change that decision

Are counterfactual explanations appropriate for the use case of our AI application? If so, have alterable features been included in the input space that can provide decision recipients with reasonable options to change their behaviour in order to obtain different results?

Have we used a solid understanding of global feature importance, correlations, and interaction effects to set up reasonable and relevant options for the possible alternative outcomes that will be explored in our counterfactual explanation tool?

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

At a glance

- Once you have extracted the rationale of the underlying logic of your AI model, you will need to take the statistical output and incorporate it into your wider decision-making process.
- Implementers of the outputs from your AI system will need to recognise the factors that they see as legitimate determinants of the outcome they are considering.
- For the most part, the AI systems we consider in this guidance will produce statistical outputs that are based on correlation rather than causation. You therefore need to sense-check whether the correlations that the AI model produces make sense in the case you are considering.
- Decision recipients should be able to easily understand how the statistical result has been applied to their particular case.

Checklist

☐ We have taken the technical explanation delivered by our AI system and translated this into reasons that can be easily understood by the decision recipient.

☐ We have used tools such as textual clarification, visualisation media, graphical representations, summary tables, or a combination, to present information about the logic of the AI system's output.

☐ We have justified how we have incorporated the statistical inferences from the AI system into our final decision and rationale explanation.

Where there is a 'human in the loop' we have trained our implementers to:

☐ Understand the associations and correlations that link the input data to the model's prediction or classification.

☐ Interpret which correlations are consequential for providing a meaningful explanation by drawing on their domain knowledge or the decision recipient's specific circumstances.

☐ Combine the chosen correlations and outcome determinants with what they know of the individual affected to come to their conclusion.

☐ Apply the AI model's results to the individual case at hand, rather than uniformly across decision recipients.

☐ Where our decision-making is fully automated, we have made sure that our AI system is set up to provide understandable explanations to individuals.

In more detail

Introduction

The non-technical dimension to rationale explanation involves working out how you are going to convey your model's results in a way that is clear and understandable to users, implementers and decision recipients.

This involves presenting information about the logic of the output as clearly and meaningfully as possible. You could do this through textual clarification, visualisation media, graphical representations, summary tables, or any combination of them. The main thing is to make sure that there is a simple way for the implementer to describe the result to an affected individual.

However, it is important to remember that the technical rationale behind an AI model's output is only one component of the decision-making and explanation process. It reveals the statistical inferences (correlations) that implementers must then incorporate into their wider deliberation before they reach their ultimate conclusions and explanations.

Integrating statistical associations into their wider deliberations means implementers should be able to recognise the factors that they see as legitimate determinants of the outcome they are considering. They must be able to pick out, amongst the model's correlations, those associations that they think reasonably explain the outcome given the specifics of the case. They then need to be able to incorporate these legitimate determining factors into their thinking about the AI-supported decision, and how to explain it.

It is likely they will need training in order to do this.

Understand the statistical rationale

Once you have extracted your explanation, either from an inherently interpretable model or from supplementary tools, you should have a good idea of both the relative feature important and significant feature interactions. This is your local explanation, which you should combine with a more global picture of the behaviour of the model across cases. Doing this should help clarify where there is a meaningful relationship between the predictor and response variables.

Understanding the relevant associations between input variables and an AI model's result (ie its statistical rationale) is the first step in moving from the model's mathematical inferences to a meaningful explanation. However, on their own, these statistical inferences are not direct indicators of what determined the outcome, or of significant population-level insights in the real world.

As a general rule, the kinds of AI and machine learning models that we are exploring in this guidance generate statistical outputs that are based on **correlational** rather than **causal** inference. In these models, a set of relevant input features, X, is linked to a target or response variable, Y, where there is an established association or correlation between them. While it is justified, then, to say that the components of X are correlated (in some unspecified way) with Y, it is not justified (on the basis of the statistical inference alone) to say that the components of X cause Y, or that X is a

direct determinant of Y. This is a version of the well-known phrase 'correlation does not imply causation'.

Further steps need to be taken to assess the role that these statistical associations should play in a reasonable explanation, given the particulars of the case being considered.

Sense-check correlations and identify legitimate determining factors in a case-by-case manner

Next, you need to determine which of the statistical associations that the model's results have identified as important are legitimate and reasonably explanatory in the case you are considering. The challenge here is that there is no simple technical tool you can use to do this.

The model's prediction and classification results are observational rather than experimental, and they have been designed to minimise error rather than to be informative about causal structures. This means it is difficult to draw out an explanation.

You will therefore need to interpret and analyse which correlations and associations are consequential for providing a meaningful explanation. You can do this by drawing on your knowledge of the domain you are working in, and the decision recipient's specific circumstances.

This should help you do two things:

- Sense-check which correlations are relevant to an explanation. This involves not only ensuring that these correlations are not spurious or caused by hidden variables, but also determining how applicable the statistical generalisations are to the affected individual's specific circumstances.

  For example, a job candidate, who has spent several years in a full-time family care role, has been eliminated by an AI model because it identifies a strong statistical correlation between long periods of unemployment and poor work performance. This suggests that the correlation identified may not reasonably apply in this case. If such an outcome were challenged, the model's implementer would have to sense-check whether such a correlation should play a significant role given the decision recipient's particular circumstances. They would

also have to consider how other factors should be weighed in justifying that outcome.

- Identifying relevant determining factors involves picking out the features and interactions that could reasonably make a real-world difference when considering how they contribute to the outcome, as it specifically applies to the decision recipient under consideration.

  For example, a model predicts that a patient has a high chance of developing lung cancer in their lifetime. The features and interactions that have significantly contributed to this prediction include family history. The doctor knows that the patient is a non-smoker and has a family history of lung cancer, and concludes that, given risks arising from shared environmental and genetic factors, family history should be considered as a strong determinant in this patient's case.

Integrate your chosen correlations and outcome determinants into your reasoning

The final step involves integrating the correlations you have identified as most relevant into your reasoning. You should consider how this particular set of factors that influenced the model's result, combined with the specific context of the decision recipient, can support your overall conclusion on the outcome.

Similarly, implementers should be able to make their reasoning explicit and intelligible to affected individuals. Decision recipients should be able to easily understand how the statistical result has been applied to their particular case, and why the implementer assessed the outcome as they did. This could be through a plain-language explanation, or any other format they require to be able to make sense of the decision.

Step 5: Prepare implementers to deploy your AI system

At a glance

- In cases where decisions are not fully automated, implementers need to be meaningfully involved.
- This means that they need to be appropriately trained to use the model's results responsibly and fairly.
- Their training should cover:
    - the basics of how machine learning works;
    - the limitations of AI and automated decision-support technologies;
    - the benefits and risks of deploying these systems, particularly how they help humans come to judgements rather than replacing that judgement; and
    - how to manage cognitive biases.
- Where decisions are fully automated and provide a result directly to the decision recipient, you should set up the AI system to provide understandable explanations.

Checklist

☐ Where there is a 'human in the loop' we have trained our implementers to:

☐ Understand the associations and correlations that link the input data to the model's prediction or classification.

☐ Interpret which correlations are consequential for providing a meaningful explanation by drawing on their domain knowledge or the decision recipient's specific circumstances.

☐ Combine the chosen correlations and outcome determinants with what they know of the individual affected to come to their conclusion.

☐ Apply the AI model's results to the individual case at hand, rather than uniformly across decision recipients.

☐ Where our decision-making is fully automated, we have made sure that our AI system is set up to provide understandable explanations to individuals.

In more detail

Introduction

When human decision-makers are meaningfully involved in deploying an AI-assisted outcome (ie where the decision is not fully automated), you should make sure that they have been appropriate trained and prepared to use your model's results responsibly and fairly.

Implementer training

Implementer training should therefore include conveying basic knowledge about the statistical and probabilistic character of machine learning, and about the limitations of AI and automated decision-support technologies. This training should avoid any anthropomorphic (or human-like) portrayals of AI systems. It should also encourage the implementers to view the benefits and risks of deploying these systems in terms of their role in helping humans come to judgements, rather than replacing that judgement.

Further, training should address any cognitive or judgemental biases that may occur when implementers use AI systems in different settings. This should be based on the use-case, highlighting, for example, where over-reliance or over-compliance with the results of computer-based system can occur (known as automation bias). Cognitive biases may include overconfidence in a prediction based on the historical consistency of data, illusions that any clustering of data points necessarily indicates significant

insights, and discounting social patterns that exist beyond the statistical result.

Individuals are likely to expect that decisions produced about them do not treat them in terms of demographic probabilities and statistics. Inferences that are drawn from a model's results should therefore be applied to the particular circumstances of the decision recipient.

Fully automated systems

While it is usually safer to have a trained human to translate the result of an AI system for the affected individual, in many cases these processes will be automated, in which case you will have to ensure the AI system is set up to provide understandable explanations to individuals.

Step 6: Consider contextual factors when you deliver your explanation

At a glance

- Several contextual factors will have an effect on the purpose for which an individual wishes to use an explanation, and on how you should deliver your explanation. The factors are the:
  - domain you work in;
  - impact on the individual;
  - data used;
  - urgency of the decision; and ;
  - audience it is being presented to.

Checklist

☐ We have considered the contextual factors that affect what a decision recipient will find useful in an explanation.

☐ We have formulated all of our explanation types in a way that is most useful for the decision recipient, taking into account any reasonable adjustments.

In more detail

Introduction

The previous steps have shown you how to gather the information you need for each explanation type, and given further details on extracting rationale explanations in particular. However, there are also several factors relating to the context within which an AI-assisted decision is made that have an effect on the type of explanation which people will find useful and the purposes they wish to use it for.

From the primary research we carried out, particularly with members of the public, we identified five key contextual factors affecting why people want explanations of AI-assisted decisions. These contextual factors are set out below, along with suggestions of which explanations to prioritise in delivering an explanation of an AI-assisted decision given the factor.

Domain factor

What is this factor?

By 'domain', we mean the setting or the sector in which you deploy your AI model to help you make decisions about people. This can affect the explanations people want. For instance, what people want to know about AI-assisted decisions made in the criminal justice domain can differ significantly from other domains such as healthcare.

Likewise, domain or sector specific explanation standards can affect what people expect out of an explanation. For example, a person receiving an AI-assisted mortgage decision will expect to learn about the reasoning behind the determination in a manner that accords with established lending standards and practices.

Which explanations should we prioritise?

Considering the domain factor is perhaps the most crucial determiner of what explanations should be included and prioritised when communicating with affected individuals. If your AI system is operating in a safety-critical setting, decision recipients will obviously want appropriate safety and performance explanations. However, if your system is operating in a domain where bias and discrimination concerns are prevalent, they will likely want you to provide a fairness explanation.

In lower-stakes domains such as e-commerce, it is unlikely that people will, on average, want or expect extensive explanations of the fairness or performance of the outputs of recommender systems. Even so, in these

lower impact domains, provisions should be made for explaining the basic rationale and responsibility components (as well as all other relevant explanation types) of any decision system that affects people.

For example 'low' impact applications such as product recommendations and personalisation - eg of advertising or content - may give rise to sensitivities around targeting particular demographics, or ignoring others (eg advertising leadership roles targeted at men) raise obvious issues of fairness and impact on society, increasing the importance of explanations addressing these issues.

Impact factor

What is this factor?

The 'impact' factor is about the effect an AI-assisted decision can have on an individual and wider society. Varying levels of severity and different types of impact can change what explanations people will find useful, and the purpose the explanation serves.

Are the decisions safety-critical, relating to life or death situations (most often in the healthcare domain)? Do the decisions affect someone's liberty or legal status? Is the impact of the decision less severe but still significant – eg denial of a utility or targeting of a political message? Or is the impact more trivial – eg being directed to a specific ticket counter by an AI system that sorts queues in an airport?

Which explanations should we prioritise?

In general, where an AI-assisted decision has a high impact on an individual, explanations such as fairness, safety and performance, and impact are often important, because individuals want to be reassured as to the safety of the decision, to trust that they are being treated fairly, and to understand the consequences.

However, the rationale and responsibility explanations can be equally as important depending on the other contextual factors at play – for instance if the features of the data used by the AI model are changeable, or the inferences drawn are open to interpretation and can be challenged.

Considering impact as a contextual factor is not straightforward. There is no hard and fast rule. It should be done on a case by case basis, and considered in combination with all the other contextual factors.

Data factor

What is this factor?

'Data' as a contextual factor relates to both the data used to train and test your AI model, as well as the input data at the point of the decision. The type of data used by your AI model can influence an individual's willingness to accept or contest an AI-assisted decision, and the actions they take as a result of it.

This factor suggests that you should think about the nature of the data your model is trained on and uses as inputs for its outputs when it is deployed. You should consider whether the data is biological or physical (eg biomedical data used for research and diagnostics), or if it is social data that relates to demographic characteristics or measurements of human behaviour.

You should also consider whether an individual can change the outcome of a decision. If the factors that go into your decision are ones that can be influenced by changes to someone's behaviour or lifestyle, it is more likely that individuals that don't agree with the outcome may want to make such changes.

For example, if a bank loan decision was made based on a customer's financial activity, the customer may want to alter their spending behaviour to change that decision in the future. This will affect the type of explanation an individual wants. However, if the data is less flexible, such as biophysical data, it will be less likely that an individual will disagree with the output of the AI system. For example in healthcare, an output that is produced by an AI system on a suggested diagnosis based on genetic data about a patient is more 'fixed' – this is not something the patient can easily change.

Which explanations should we prioritise?

It will often be useful to prioritise the rationale explanation, for both social data and biophysical data. Where social data is used, individuals receiving an unfavourable decision can understand the reasoning and learn from this to appropriately adapt their behaviour for future decisions. For biophysical data, this can help people understand why a decision was made about them.

However, where biophysical data is used such as in medical diagnoses, individuals often prefer to simply know what the decision outcome means for them, and to be reassured about the safety and reliability of the decision. In these cases it makes sense to prioritise the impact and safety and performance explanations to meet these needs.

On the other hand, where the nature of the data is social, or subjective, individuals are more likely to have concerns about what data was taken into account for the decision, and the suitability or fairness of this in influencing an AI-assisted decision about them. In these circumstances, the data and fairness explanations will help address these concerns by telling people what the input data was, where it was from, and what measures you put in place to ensure that using this data to make AI-assisted decisions does not result in bias or discrimination.

Urgency factor

What is this factor?

The 'urgency' factor concerns the importance of receiving, or acting upon the outcome of an AI-assisted decision within a short timeframe. What people want to know about a decision can change depending on how little or much time they have to reflect on it.

The urgency factor recommends that you give thought to how urgent the AI-assisted decision is. Think about whether or not a particular course of action is often necessary after the kind of decisions you make, and how quickly that action needs to be taken.

Which explanations should we prioritise?

Where urgency is a key factor in the context within which your AI-assisted decision is made, it is more likely that individuals will want to know what the consequences are for them, and to be reassured that the AI model helping to make the decision is safe and reliable. As such, the impact and safety and performance explanations are suitable in these cases. This is because these explanations will help individuals to understand how the decision affects them, what happens next, and what measures and testing were implemented to maximise and monitor the safety and performance of the AI model.

Audience factor

What is this factor?

'Audience' as a contextual factor is about the individuals to whom you are explaining an AI-assisted decision. The groups of people you make decisions about, and the individuals within those groups have an effect on what type of explanations are meaningful or useful for them.

What level of expertise (eg about AI) do they have in relation to what the decision is about? Are a broad range of people subject to decisions you make (eg the UK general public), thus indicating that there might also be a broad range of knowledge or expertise? Or are the people you make decisions about limited to a smaller subset (eg your employees), suggesting they may be more informed on the things you are making decisions about? Also consider the decision recipients require any reasonable adjustments in how they receive the explanation (Equality Act 2010).

Which explanations should we prioritise?

If the people about whom you are making AI-assisted decisions are likely to have some domain expertise, you might consider using the rationale explanation. This is because you can be more confident that they can understand the reasoning and logic of an AI model, or a particular decision, due to being more familiar with the topic of the decisions. Additionally, if people subject to your AI-assisted decisions have some technical expertise, or are likely to be interested in the technical detail underpinning the decision, the safety and performance explanation will help.

Alternatively, where you think it's often likely the people you make AI-assisted decisions about do not have any specific expertise or knowledge about either the topic of the decision, or its technical aspects, other explanation types such as responsibility, or particular aspects of the safety and performance explanation may be more helpful. This is so that people can be reassured about the safety of the system, and know who to contact to query or question an AI decision.

Of course, even for those with little knowledge of an area about which an AI-assisted decision is made, the rationale explanation can still be useful to help illuminate the reasons why a decision was made in plain and simple terms. But there may also be occasions where the data used and inferences drawn by an AI model are particularly complex (see the 'data' factor above), and individuals would rather delegate the rationale explanation to a relevant domain expert to review and come to their own informed conclusions about the validity or suitability of the reasons for the decision (eg a doctor in a healthcare setting).

Step 7: Consider how to present your explanation

At a glance

- How you present your explanation depends on the way in which you make AI-assisted decisions, and on how people might expect you to deliver explanations you make without using AI.
- You can 'layer' your explanation by proactively providing individuals first with the explanations you have prioritised, and making additional explanations available in further layers. This helps to avoid information (or explanation) overload.
- You should think of delivering your explanation as a conversation, as opposed to a one-way process. People should be able to discuss a decision with a competent human being.
- Providing your explanation at the right time is also important.
- To increase trust and awareness of your use of AI, you can proactively engage with your customers by making information available about how you use AI systems to help you make decisions.

Checklist

☐ We have presented our explanation in a layered way, giving the most relevant explanation type(s) upfront, and providing the other types in additional layers.

☐ We have made it clear how decision recipients can contact us if they would like to discuss the AI-assisted decision with a human being.

☐ We have provided the decision recipient with the process-based and relevant outcome-based explanation for each explanation type, in advance of making a decision.

☐ We have proactively made information about our use of AI available in order to build trust with our customers and stakeholders.

In more detail

Introduction

You should determine the most appropriate method of delivery based on the way in which you make AI-assisted decisions about people, and how they might expect you to deliver explanations of decisions you make without using AI. This might be verbally, face to face, in hard-copy or electronic format. Think about any reasonable adjustments you might need to make for people under the Equality Act 2010. The timing for delivery of explanations will also affect the way you deliver the explanation.

If you deliver explanations in hard-copy or electronic form, you may also wish to consider whether there are design choices that can help make what you're telling people more clear and easy to understand. For example, in addition to text, simple graphs and diagrams may help explain certain explanations such as rationale and safety and performance. Depending on the size and resources of your organisation, you may be able to draw on the expertise of user experience and user interface designers.

Layering explanations

Based on the guidance we've provided above, and engagement with industry, we think it makes sense to build a 'layered' explanation.

By layered we mean proactively providing individuals with the prioritised explanations (the first layer), and making the additional explanations available on a second, and possibly third, layer. If you deliver your explanation on a website, you can use expanding sections, tabs, or simply link to webpages with the additional explanations.

The purpose of this layered approach is to avoid information (or explanation) fatigue. It means you won't overload people. Rather, they are provided with what is likely to be the most relevant and important information, while still having clear and easy access to other explanatory information, should they wish to know more about the AI decision.

Explanation as a dialogue

However you choose to deliver your explanations to individuals, it is important to think of this exercise as a conversation as opposed to a one-way process. By providing the priority explanations, you are the initiating a conversation, not ending it. Individuals should not only have easy access to additional explanatory information (hence layered explanations), but they should also be able to discuss the AI-assisted decision with a human being. This ties in with the responsibility explanation and having a human reviewer. However, as well as being able to contest decisions, it's important to provide a way for people to talk about and clarify explanations with a competent human being.

Explanation timing

It is important to provide explanations of AI-assisted decisions to individuals at the right time.

Delivering an explanation is not just about telling people the result of an AI decision. It is equally about telling people how decisions are made in advance.

What explanation can I provide in advance?

In Step 2 we provided two categories for each type of explanation: process-based and outcome-based.

You can provide the process-based explanations in advance of a specific decision. In addition, there will be some outcome-based explanations that you can provide in advance, particularly those related to:

- Responsibility - who is responsible for taking the decision that is supported by the result of the AI system, for reviewing and for implementing it;
- Impact – how you have assessed the potential impact of the model on the individual and the wider community; and
- Data - what data was input into the AI system to train, test, and validate it.

There will also be some situations when you can provide the same explanation in advance of a decision as you would afterwards. This is

because in some sectors it is possible to run a simulation of the model's output. For example, if you applied for a loan some organisations could explain the computation and tell you which factors matter in determining whether or not your application would be accepted. In cases like this, the distinction between explanations before and after a decision is less important. However, in many situations this won't be the case.

What should I do?

After you have prioritised the explanations (see Step 1), you should provide the relevant process-based explanations before the decision, and the outcome-based explanations if you are able to.

What explanation can I provide after a decision?

You can provide the full explanation after the decision, however there are some specific outcome-based explanations that you will not have been able to explain in advance - ie rationale, fairness, and safety and performance of the system, which are specific to a particular decision and are likely to be queried after the decision has been made. These explain the underlying logic of the system that led to the specific decision or output, whether the decision recipient was treated fairly compared with others who were similar, and whether the system functioned properly in that particular instance.

> The basics of explaining AI

Example

In this example, clinicians are using an AI system to help them detect cancer.

Example: Explanations in health care - cancer diagnosis		
Before decision	**Process-based explanation**	Responsibility – who is responsible for ensuring the AI system used in detecting cancer works in the intended way. Rationale – what steps you have taken to ensure that the components or measurements used in the model make sense for detecting cancer and can be made understandable to

		affected patients. Fairness – what measures you have taken to ensure the model is fair, prevents discrimination and mitigates bias (this may be less relevant here where biophysical data is being used). Safety and performance – what measures you have taken to ensure the model chosen to detect cancer is secure, accurate, reliable and robust, and how it has been tested, verified and validated. Impact – what measures you have taken to ensure that the AI model does not negatively impact the patient in how it has been designed or used. Data – how you have ensured that the source(s), quantity, and quality of the data used to train the system is appropriate for the type(s) of cancer detection for which you are utilising your model.
	Outcome-based explanation	Responsibility – who is responsible for taking the diagnosis resulting from the AI system's output, implementing it, and providing an explanation for how the diagnosis came about, and who the patient can go to in order to query the diagnosis. Impact – how the design and use of the AI system in the particular case of the patient will impact the patient. For example, if the system detects cancer but the result is a false positive, this could have a significant impact on the mental health of the patient. Data – the patient's data that will be used in this particular instance.

After decision	**Outcome-based explanation**	Rationale – whether the AI system's output, ie what it has detected as being cancerous or not, makes sense in the case of the patient, given the doctor's domain knowledge. Fairness – whether the model has produced results consistent with those it has produced for other patients with similar characteristics. Safety and performance – how secure, accurate, reliable and robust the AI model has been in the patient's particular case, and which safety and performance measures were used to test this.

Why is this important?

Not only is this a good way to provide an explanation to an individual when they might need it, it is also a way to comply with the law.

Articles 13-14 of the GDPR require that you proactively provide individuals with '…meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject…' in the case of solely automated decisions with a legal or similarly significant effect.

Article 15 of the GDPR also gives individuals a right to obtain this information at any time on request.

This is also good practice for systems where there is a 'human in the loop'.

The process-based and outcome-based explanations relating to the rationale of the AI system, and the outcome-based explanation relating to the AI system's impact on the individual, fulfil this requirement of the GDPR.

It is up to you and your organisation to determine the most appropriate way to deliver the explanations you choose to provide.

However, you might consider what the most direct and helpful way would be to deliver explanations that you can provide in advance of a decision. You should consider where individuals are most likely to go to find an explanation or information on how you make decisions with support of AI systems.

You should think about using the same platform for providing an advance explanation that you will use to provide the ultimate decision. This means that the information that an individual needs is in one place. You should also ensure that the explanation is prominent, to make it easier for individuals to find it.

Proactive engagement

How can we build trust?

Proactively making information available about how you use AI systems to help you make decisions is a good way to increase awareness among your customers. This will help them know more about when and why you use an AI system and how it works.

By being open and inclusive in how you share this information, you can increase the trust your customers have in how you operate, and build confidence in your organisation using AI to help them get a better service.

In the primary research we conducted, we found that the public is looking for more engagement from organisations and awareness raising about how they use AI for decision-making. By being proactive, you can use this engagement to help you fulfil the principle of being transparent.

What should we proactively share?

Among the things you could consider sharing are the following:

- What is AI?

This helps to demystify the technologies involved. It might be useful to outline these technologies, and provide a couple of examples of where AI is used in your sector.

> A good example is this animation about machine learning produced by researchers at the University of Oxford.

> ‘What is Machine Learning?’ animation

- How can it be used for decision-making?

  This outlines the different ways AI is useful for supporting decision-making – this tells people what the tools do. You could provide some examples of how you use it to help you make decisions.

- What are the benefits?

  This should lay out how AI can be beneficial, specifically for the individuals that are affected by the decisions you make. For example, if you are a service provider, you can outline how it can personalise your services so that your customers can get a better experience. The benefits you outline could also explore ways that the AI tools available can be better than more traditional decision-support tools. Examples could help you to make this clear.

- What are the risks?

  You should be honest about how AI can go wrong in your sector, for example how it can lead to discrimination or misinformation, and how you will mitigate this. This helps to set people’s expectations about what AI can do in their situation, and helps them understand what your organisation will do to look after them.

  You should also provide information about people’s rights under the GDPR, for example the right to object or challenge the use of AI, and the right to obtain human review or intervention.

- Why does our organisation use AI for decisions?

  This should clearly and comprehensively outline why you have chosen to use AI systems in your particular organisation. It should expand on the more general examples you have provided above for how it improves the service you offer compared with other approaches (if applicable), and what the benefits are for your customers.

- Where/when do we do this?

Here you can describe which parts of your organisation and in which parts of the decision-making process you are using AI. You should make this as informative as possible. You could also outline what measures you have put in place to ensure that the AI system you are using in each of these areas is designed in a way to maximise the benefits and minimise the risks. In particular, you should be clear about whether there is a 'human in the loop' or whether the AI is solely automated. In addition, it might be helpful to show how you are managing the system's use to make sure it is maximising the interests of your customers.

- Who can I speak to about it?

  You could provide an email address or helpline for interested members of the public to contact in order to get more information on how you are using AI. Those answering these queries should have good knowledge of AI and how you are using it, and be able to explain it in a clear, open and accessible way. The amount of detail you provide should be proportionate to the information people ask for.

How should we share this?

There are many different ways you could proactively share information with your customers and stakeholders:

- Your usual communications to customers and stakeholders, such as regular newsletters or customer information.
- Providing a link to a dedicated part of your organisation's website outlining the sections above.
- Flyers and leaflets distributed in your offices and to those of other relevant or partner organisations.
- An information campaign or other initiative in partnership with other organisations.
- Information you distribute through trade bodies.

Your communications team will have an important role to play in making sure the information is targeted and relevant to your customers.

The ICO has written guidance on the right to be informed, which will help you with this communication task.

> Guidance on the right to be informed (GDPR)

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.

Example: Explanations in healthcare - cancer diagnosis

Step 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…

Fairness – Less important due to use of biophysical data, as opposed to social or demographic data, but still relevant in areas such as data representativeness…

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…

Step 2: Collect the information you need for each explanation type

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.

Data collection and pre-processing: information on how data has been labelled and how that shows the reasons for classifying, for example, certain images as tumours.

Responsibility

Process-based explanation: information on those responsible within the healthcare organisations 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 taking the output reached by the AI system, implementing the output into diagnosis, reviewing it, and providing explanations for how the diagnosis came about (ie who the patient can go to in order to query the diagnosis).

Data collection and pre-processing: information on who or which part of the healthcare organisation is responsible for collecting and pre-processing the patient's data. Being transparent about the process can help the healthcare organisation to build trust and confidence in their use of AI.

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.

Data collection and pre-processing: information on the accuracy rate of the model and the accuracy-related measures the healthcare organisation 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.

Data collection and pre-processing: the data the healthcare organisation uses has a bearing on its impact and risk assessment.

Step 3: Build your rationale explanation to provide meaningful information about the underlying logic of your AI system

The healthcare organisation 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.

Step 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.

Step 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.

Step 6: Consider the contextual factors when you deliver your explanation

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…

Step 7: Consider how to present your explanation

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 explanation face to face with patient & leaflet.

Layer 2

- Data explanation
- Fairness explanation

Delivery – eg the clinician gives the patient a leaflet.

Appendix 2: Further reading

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. https://projecteuclid.org/download/pdfview_1/euclid.aoas/1324399600

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. https://link.springer.com
/content/pdf/10.1023/A:1010933404324.pdf

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

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems* (pp. 4066-4076). http://papers.nips.cc/paper/6995-counterfactual-fairness.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