

# Explaining decisions made with AI

Introduction	3
Part 1 The basics of explaining AI	4
Definitions	6
Legal framework	10
Benefits and risks	17
What goes into an explanation?	21
What are the contextual factors?	35
The principles to follow	40
Part 2: Explaining AI in practice	47
Summary of the tasks to undertake	49
Task 1: Select priority explanations by considering the domain, use case and impact on the individual	57
Task 2: Collect and pre-process your data in an explanation-aware manner	57
Task 3: Build your system to ensure you are able to extract relevant information for a range of explanation types	64
Task 4: Translate the rationale of your system's results into useable and easily understandable reasons	80
Task 5: Prepare implementers to deploy your AI system	85
Task 6: Consider how to build and present your explanation	94
Part 3: What explaining AI means for your organisation	96
Organisational roles and functions for explaining AI	100
Policies and procedures	105
Documentation	113
Annexe 1: Example of building and presenting an explanation of a cancer diagnosis	118
Annexe 2: Algorithmic techniques	123
Annexe 3: Supplementary models	133
Annexe 4: Further reading	139
Annexe 5: Argument-based assurance cases	

# Introduction

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

## At a glance

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

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

<p><b>Part 1: The basics of explaining AI</b></p> <p>Aimed at DPOs and compliance teams, part one defines the key concepts and outlines a number of different types of explanations. It will be relevant for all members of staff involved in the development of AI systems.</p>	<p><b>Part 2: Explaining AI in practice</b></p> <p>Aimed at technical teams, part two helps you with the practicalities of explaining these decisions and providing explanations to individuals. This will primarily be helpful for the technical teams in your organisation, however your DPO and compliance team will also find it useful.</p>	<p><b>Part 3: What explaining AI means for your organisation</b></p> <p>Aimed at senior management, part three goes into the various roles, policies, procedures and documentation that you can put in place to ensure your organisation is set up to provide meaningful explanations to affected individuals. This is primarily targeted at your organisation’s senior management team, however your DPO and compliance team will also find it useful.</p>
--	--	---

# Part 1 The basics of explaining AI

## About this guidance

### What is the purpose of this guidance?

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

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

This part of the guidance outlines the:

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

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

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

### How should we use this guidance?

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

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

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

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

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

## What is the status of this guidance?

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

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

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

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

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

“

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

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

“

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

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

# Definitions

## At a glance

Artificial Intelligence (AI) can be defined in many ways. However, within this guidance, we define it as an umbrella term for a range of algorithm-based technologies that solve complex tasks by carrying out functions that previously required human thinking. Decisions made using AI are either fully automated, or with a 'human in the loop'. As with any other form of decision-making, those impacted by an AI supported decision should be able to hold someone accountable for it.

## In more detail

- [What is AI?](#)
- [What is an output or an AI-assisted decision?](#)
- [How is an AI-assisted decision different to one made only by a human?](#)

## What is AI?

AI is an umbrella term for a range of technologies and approaches that often attempt to mimic human thought to solve complex tasks. Things that humans have traditionally done by thinking and reasoning are increasingly being done by, or with the help of, AI.



In **healthcare** AI can be used to spot early signs of illness and diagnose disease.



In **policing** AI can be used to target interventions and identify potential offenders.



In **marketing** AI can be used to target products and services to consumers.

While AI has existed for some time, recent advances in computing power, coupled with the increasing availability of vast swathes of data, mean that AI designers are able to build systems capable of undertaking these complex tasks.

As information processing power has dramatically increased, it has become possible to expand the number of calculations AI models complete to effectively map a set of inputs into a set of outputs. This means that

the correlations that AI models identify and use to produce classifications and predictions have also become more complex and less intrinsically understandable to human thinking. It is therefore important to consider how and why these systems create the outputs they do.

There are several ways to build AI systems. Each involves the creation of an algorithm that uses data to model some aspect of the world, and then applies this model to new data in order to make predictions about it.

Historically, the creation of these models required incorporating considerable amounts of hand-coded expert input. These 'expert systems' applied large numbers of rules, which were taken from domain specialists, to draw inferences from that knowledge base. Though they tended to become more accurate as more rules were added, these systems were expensive to scale, labour intensive, and required significant upkeep. They also often responded poorly to complex situations where the formal rules upon which they generated their inferences were not flexible enough.

More recently data-driven, machine learning (ML) models have emerged as the dominant AI technology. These kinds of models may be constructed using a few different learning approaches that build from the past information contained in collected data to identify patterns and hone classificatory and predictive performance. The three main ML approaches are supervised, unsupervised, and reinforcement learning:

Supervised learning models are trained on a dataset which contains labelled data. 'Learning' occurs in these models when numerous examples are used to train an algorithm to map input variables (often called features) onto desired outputs (also called target variables or labels). On the basis of these examples, the ML model is able to identify patterns that link inputs to outputs. ML models are then able to reproduce these patterns by employing the rules honed during training to transform new inputs received into classifications or predictions.

Unsupervised learning models are trained on a dataset without explicit instructions or labelled data. These models identify patterns and structures by measuring the densities or similarities of data points in the dataset. Such algorithmic models can be used to:

- cluster data (grouping similar data together);
- detect anomalies (flagging inputs that are outliers compared to the rest of the dataset); and
- associate a data point with other attributes that are typically seen together.

Reinforcement learning models learn on the basis of their interactions with a virtual or real environment rather than existing data. Reinforcement learning 'agents' search for an optimal way to complete a task by taking a series of steps that maximise the probability of achieving that task. Depending on the steps they take, they are rewarded or punished. These 'agents' are encouraged to choose their steps to maximise their reward. They 'learn' from past experiences, improve with multiple iterations of trial and error, and may have long-term strategies to maximise their reward overall rather than looking only at their next step.

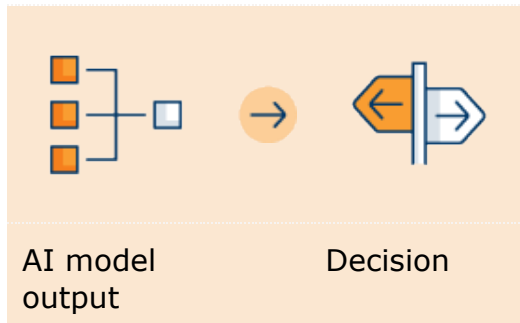
While this guidance is applicable to all three of these ML methods, it mainly focuses on supervised learning, the most widely used of the approaches.

## What is an AI output or an AI-assisted decision?

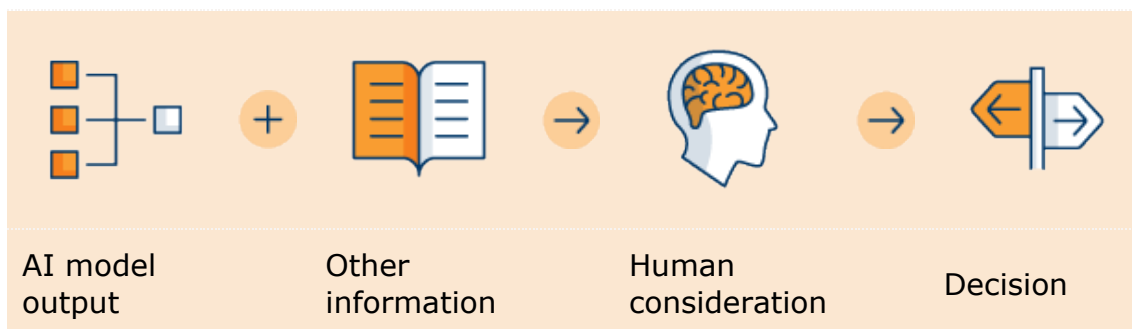
The output of an AI model varies depending on what type of model is used and what its purpose is. Generally, there are three main types of outputs:

- a prediction (eg you will not default on a loan);
- a recommendation (eg you would like this news article); or
- a classification (eg this email is spam).

In some cases, an AI system can be fully automated when deployed, if its output and any action taken as a result (the decision) are implemented without any human involvement or oversight.



In other cases, the outputs can be used as part of a wider process in which a human considers the output of the AI model, as well as other information available to them, and then acts (makes a decision) based on this. This is often referred to as having a 'human in the loop'.



We use the term 'AI decision' broadly, incorporating all the above. So, an AI decision can be based on a prediction, a recommendation or a classification. It can also refer to a solely automated process, or one in which a human is involved.

### Further reading

For more information on what constitutes meaningful human involvement in an AI-assisted decision process, read our [guidance on automated decision-making and profiling](#) in the Guide to the GDPR, and [advice on this topic in our draft AI auditing framework](#).

### How is an AI- assisted decision different to one made only by a human?

One of the key differences is who an individual can hold accountable for the decision made about them. When it is a decision made directly by a human, it is clear who the individual can go to in order to get an explanation about why they made that decision. Where an AI system is involved, the responsibility for the decision can be less clear.



There should be no loss of accountability when a decision is made with the help of, or by, an AI system, rather than solely by a human. Where an individual would expect an explanation from a human, they should instead expect an explanation from those accountable for an AI system.

# Legal framework

## At a glance

The General Data Protection Regulation (GDPR) and the Data Protection Act 2018 (DPA 2018) regulate the collection and use of personal data. Where AI uses personal data it falls within the scope of this legislation. This can be through the use of personal data to train, test or deploy an AI system. Administrative law and the Equality Act 2010 are also relevant to providing explanations when using AI.

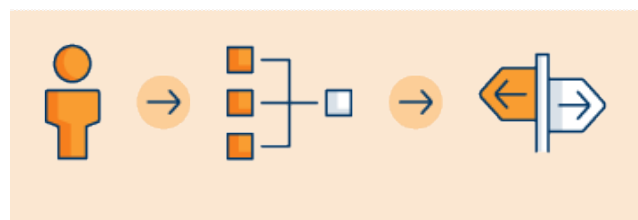
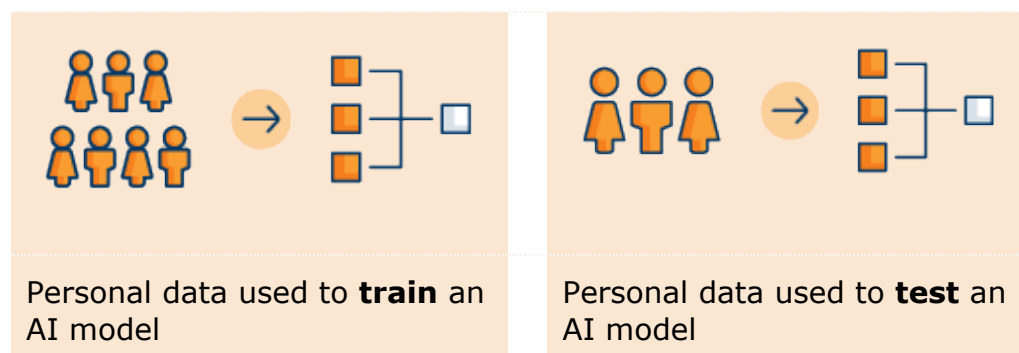
## In more detail

- [What does data protection law have to do with AI?](#)
- [Does data protection law actually mention AI?](#)
- [Does data protection law require that we explain AI-assisted decisions to individuals?](#)
- [Are there other relevant laws?](#)

## What does data protection law have to do with AI?

In the UK, data protection law is made up of the GDPR and the DPA 2018. Together, they regulate the collection and use of personal data – information about identified or identifiable individuals. Please note that from January 2021 references to the GDPR should be read as references to the equivalent articles in the UK GDPR.

Where AI doesn't involve the use of personal data, it falls outside the remit of data protection law. For example, the use of AI for weather forecasting or astronomy. But very often, AI does use or create personal data. In some cases, vast amounts of personal data are used to train and test AI models. On deployment, more personal data is collected and fed through the model to make decisions about individuals. Those decisions about individuals – even if they are only prediction or inferences – are themselves personal data.



On **deployment**, personal data used or created to make decisions about individuals

In any of these cases, AI is within the scope of data protection law.

## Does data protection law actually mention AI?

Data protection law is technology neutral. It does not directly reference AI or any associated technologies such as machine learning.

However, the GDPR and the DPA 2018 do have a significant focus on large scale automated processing of personal data, and several provisions specifically refer to the use of profiling and automated decision-making. This means it applies to the use of AI to provide a prediction or recommendation about someone.

## The right to be informed

Articles 13 and 14 of the GDPR give individuals the right to be informed of:

- the existence of solely automated decision-making producing legal or similarly significant effects;
- meaningful information about the logic involved; and
- the significance and envisaged consequences for the individual.

## The right of access

Article 15 of the GDPR gives individuals the right of access to:

- information on the existence of solely automated decision-making producing legal or similarly significant effects;
- meaningful information about the logic involved; and
- the significance and envisaged consequences for the individual.

Recital 71 provides interpretative guidance on rights related to automated decision-making. It mainly relates to Article 22 rights, but also makes clear that individuals have the right to obtain an explanation of a solely automated decision after it has been made.

## The right to object

Article 21 of the GDPR gives individuals the right to object to processing of their personal data, specifically including profiling, in certain circumstances.

There is an absolute right to object to profiling for direct marketing purposes.

## Rights related to automated decision-making including profiling

Article 22 of the GDPR gives individuals the right not to be subject to a solely automated decision producing legal or similarly significant effects. There are some exceptions to this and in those cases it obliges organisations to:

- adopt suitable measures to safeguard individuals, including the right to obtain human intervention;
- express their view; and
- contest the decision.

Recital 71 also provides interpretive guidance for Article 22.

## Data protection impact assessments

Article 35 of the GDPR requires organisations to carry out Data Protection Impact Assessments (DPIAs) if their processing of personal data, particularly when using new technologies, is likely to result in a high risk to individuals.

A DPIA is always required for any systematic and extensive profiling or other automated evaluation of personal data which are used for decisions that produce legal or similarly significant effects on people.

DPIAs are therefore likely to be an obligation if you are looking to use AI systems to process personal data, and you should carry them out prior to the processing in order to identify and assess the levels of risk involved. DPIAs should be 'living documents' that you review regularly, and when there is any change to the nature, scope, context or purposes of the processing.

The ICO has published additional guidance on DPIAs, including a list of processing operations which require a DPIA. The list mentions AI, machine learning, large-scale profiling and automated decision-making resulting in denial of a service, product or benefit.

If a DPIA indicates there are residual high risks to the rights and freedoms of individuals that cannot be reduced, you must consult with the ICO prior to the processing.

## Further Reading

### Data Protection Impact Assessments (DPIAs)

For organisations

## Does data protection law require that we explain AI-assisted decisions to individuals?

As above, the GDPR has specific requirements around the provision of information about, and an explanation of, an AI-assisted decision where:

- it is made by a process without any human involvement; and
- it produces legal or similarly significant effects on an individual (something affecting an individual's legal status/ rights, or that has equivalent impact on an individual's circumstances, behaviour or opportunities, eg a decision about welfare, or a loan).

In these cases, the GDPR requires that you:

- are proactive in "...[giving individuals] meaningful information about the logic involved, as well as the significance and envisaged consequences..." (Articles 13 and 14);
- "... [give individuals] at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision." (Article 22); and

- "... [give individuals] the right to obtain... meaningful information about the logic involved, as well as the significance and envisaged consequences..." (Article 15) "...[including] an explanation of the decision reached after such assessment..." (Recital 71).

The GDPR's recitals are not legally binding, but they do clarify the meaning and intention of its articles. So, the reference to an explanation of an automated decision after it has been made in Recital 71 makes clear that such a right is implicit in Articles 15 and 22. You need to be able to give an individual an explanation of a fully automated decision to enable their rights to obtain meaningful information, express their point of view and contest the decision.

But even where an AI-assisted decision is not part of a solely automated process (because there is meaningful human involvement), if personal data is used, it is still subject to all the GDPR's principles. The GDPR principles of fairness, transparency and accountability are of particular relevance.

## **Fairness**

Part of assessing whether your use of personal data is fair is considering how it affects the interests of individuals. If an AI-assisted decision is made about someone without some form of explanation of (or information about) the decision, this may limit their autonomy and scope for self-determination. This is unlikely to be fair.

## **Transparency**

Transparency is about being clear, open and honest with people about how and why you use their personal data. In addition to the information requirements on automated processing laid out in Articles 13 and 14 of the GDPR, Recital 60 states that you should provide any further information necessary to ensure fair and transparent processing taking into account the specific circumstances and context in which you process the personal data. It is unlikely to be considered transparent if you are not open with people about how and why an AI-assisted decision about them was made, or where their personal data was used to train and test an AI system. Providing an explanation, in some form, will help you be transparent. Information about the purpose for which you are processing someone's data under Articles 13-15 of the GDPR could also include an explanation in some cases.

## **Accountability**

To be accountable, you must be able to demonstrate your compliance with the other principles set out in Article 5 of the GDPR, including those of data minimisation and accuracy. How can you show that you treated an individual fairly and in a transparent manner when making an AI-assisted decision about them? One way is to provide them with an explanation of the decision and document its provision.

So, whichever type of AI-assisted decision you make (involving the use of personal data), data protection law still expects you to explain it to the individuals affected.

## **Parts 3 and 4 of the DPA 2018**

In addition, there are separate provisions in Part 3 of the DPA 2018 for solely automated decisions that have an adverse legal effect or significantly affect the data subject and which are carried out for law enforcement purposes by competent authorities. Individuals can obtain human intervention, express their point of view, and obtain an explanation of the decision and challenge it. Currently, instances of solely

automated decision-making in law enforcement are likely to be rare.

There are also separate provisions in Part 4 of the DPA 2018 for solely automated decision-making carried out by the intelligence services that significantly affect a data subject. Individuals have a right to obtain human intervention in these cases. There is also a general right for individuals to have information about decision-making where the controller is processing their data and the results produced by the processing are applied to them. In these cases, they can request “knowledge of the reasoning underlying the processing.” However, these rights may be limited by the exemption for safeguarding national security in Part 4.

## Are there other relevant laws?

The GDPR is the main legislation in the United Kingdom that explicitly states a requirement to provide an explanation to an individual. Other laws may be relevant that mean it is good practice to explain AI-assisted decisions, and we have listed examples below.

### **Equality Act 2010**

The Equality Act 2010 applies to a range of organisations, including government departments, service providers, employers, education providers, transport providers, associations and membership bodies, as well as providers of public functions.

Behaviour prohibited under the Equality Act 2010 is any that discriminates, harasses or victimises another person on the basis of any of these “protected characteristics”:

- Age
- Disability
- Gender reassignment
- Marriage and civil partnership
- Pregnancy and maternity
- Race
- Religion and belief
- Sex
- Sexual orientation

If you are using an AI system in your decision-making process, you need to ensure, and be able to show, that this does not result in discrimination that:

- causes the decision recipient to be treated worse than someone else because of one of these protected characteristics; or
- results in a worse impact on someone with a protected characteristic than someone without one.

Reasonable adjustments mean that employers or those providing a service have a duty to avoid as far as possible by reasonable means the disadvantage that a disabled person experiences because of their impairments.

Therefore you should explain to the decision recipient that the decision is not discriminatory about any of

the protected characteristics listed above. This explanation must be in a format that the decision recipient can meaningfully engage with.

## Further Reading

 [Equality and Human Rights Commission](#) 

External link

 [Reasonable adjustments](#) 

External link

## Judicial review under administrative law

Anyone can apply to challenge the lawfulness of government decisions. This means that individuals are able to challenge the decision made by a public sector agency, or by private bodies contracted by government to carry out public functions, where they have deployed AI systems to support decision-making. It should be possible to judicially review these systems where public agencies have used them to make decisions about individuals, on the basis that the decision was illegal, irrational, or the way in which it was made was 'improper'.

## Further Reading

 [Administrative Law and the Machines of Government](#) 

External link

 [Algorithm-assisted decision-making in the public sector](#) 

External link

## Additional legislation

An explanation may also indicate compliance or 'best practice' with other legislation. We do not plan to go into detail about these laws here, as they may be specific to your sector. We recommend you contact your own regulatory body if you are concerned this may apply to you.

Further legislation this may apply to includes (please note that this list is not intended to be exhaustive, and that further legislation may apply in some cases):

- e-Privacy legislation
- Law Enforcement Directive
- Consumer Rights legislation
- Financial Services legislation
- Competition law
- Human Rights Act
- Legislation about health and social care
- Regulation around advertising and marketing

- Legislation about school admissions procedures



# Benefits and risks

## At a glance

Explaining AI-assisted decisions has benefits for your organisation. It can help you comply with the law, build trust with your customers and improve your internal governance. Society also benefits by being more informed, experiencing better outcomes and being able to engage meaningfully in the decision-making process. If your organisation does not explain AI-assisted decisions, it could face regulatory action, reputational damage and disengagement by the public.

## In more detail

- [What are the benefits to your organisation?](#)
- [What are the benefits to individuals and society?](#)
- [What are the risks of explaining AI decisions?](#)
- [What are the risks of not explaining AI decisions?](#)

What are the benefits to your organisation?

### **Legal compliance**

As set out in the legal framework section of this guidance, a number of laws (both sectoral and cross-sector) have something relevant to say on this topic. Some explicitly require explanations of AI-assisted decisions in certain circumstances, others have broader requirements around the fair treatment of citizens. But whatever sector or business you are in, explaining your AI-assisted decisions to those affected will help to give you (and your board) better assurance of legal compliance, mitigating the risks associated with non-compliance.

### **Trust**

Explaining AI-assisted decisions to affected individuals makes good business sense. This will help to empower them to better understand the process and allow them to challenge and seek recourse where necessary. Handing a degree of control back to individuals in this way may help to foster trust in your use of AI decisions and give you an edge over other organisations and competitors that are not as progressive and respectful in their interactions with customers.

### **Internal governance**

Explaining AI-assisted decisions to affected individuals requires those within your organisation to understand the models, choices and processes associated with the AI decisions you make. So, by making 'explainability' a key requirement, you will also have better oversight of what these systems do and why. This will help to ensure your AI systems continue to meet your objectives and support you in refining them to increase precision.

## What are the benefits to individuals and society?

### Informed public

As more organisations incorporate explanations to individuals as a core element of their AI-assisted decision-making systems, the general public will gain an increasing awareness of when and where such decisions are made. In turn, this may help the public to have a meaningful involvement in the ongoing conversation about the deployment of AI and its associated risks and benefits. This could help address concerns about AI and support a more constructive and mutually beneficial debate for business and society.

### Better outcomes

Organisations are required to identify and mitigate discriminatory outcomes, which may already be present in human decision-making, or may be exacerbated or introduced by the use of AI. Providing explanations to affected individuals can help you to do this, and highlight issues that may be more difficult to spot. Explanations should therefore support more consistency and fairness in the outcomes for different groups across society.

### Human flourishing

Giving individuals explanations of AI-assisted decisions helps to ensure that your use of AI is human-centric. The interests of your customers are paramount. As long as you have well-designed processes to contest decisions and continuously improve AI systems based on customer feedback, people will have the confidence to express their point of view.

## What are the risks of explaining AI decisions?

Industry engagement activities we carried out highlighted a number of elements that may have a limiting effect on the information that can be provided to individuals when explaining AI-assisted decisions. The explanations set out in this guidance have largely been designed to take these issues into account and mitigate the associated risks, as explained below.

### Distrust

It could be argued that providing **too much** information about AI-assisted decisions may lead to increased distrust due to the complex, and sometimes opaque, nature of the process.

While AI-assisted decisions are often undeniably complex, the explanation types and explanation extraction methods offered in this guidance are designed to help you, where possible, to simplify and transform this complexity into understandable reasoning. In cases where fairness and physical wellbeing are a central issue, focusing on relevant explanation types will help you to build trust and reassure individuals about the safety and equity of an AI model without having to dive deeply into the complexity of the system's rationale. This is particularly the case with the safety and performance explanation and fairness explanation. These show how you have addressed these issues, even if the rationale of a decision is particularly complex and difficult to convey.

### Commercial sensitivities

You may have concerns that your explanations of AI-assisted decisions disclose commercially sensitive

material about how your AI model and system works.

We don't think the explanations we set out here will normally risk such disclosures. Neither the rationale nor the safety and performance explanations require you to provide information so in-depth that they reveal your source code or any algorithmic trade secrets. However, you will have to form a view based on your specific situation.

Where you do think it's necessary to limit detail (eg feature weightings or importance), you should justify and document your reasons for this.

### **Third-party personal data**

Due to the way you train your AI model, or input data for particular decisions, you may be concerned about the inappropriate disclosure of personal data of someone other than the individual the decision is about.

For some of the explanations we identify in this guidance this is not a problem. However, there are potential risks with the rationale, fairness and data explanation types – information on how others similar to the individual were treated and detail on the input data for a particular decision (which is about more than one person).

You should assess this risk as part of a data protection impact assessment (DPIA), and make justified and documented choices about the level of detail it is safe to provide for these explanations.

### **Gaming**

Depending on what you make AI-assisted decisions about, you may need to protect against the risk that people may game or exploit your AI model if they know too much about the reasons underlying its decisions.

Where you are using AI-assisted decisions to identify wrongdoing or misconduct (eg fraud detection), the need to limit the information you provide to individuals will be stronger, particularly about the rationale explanation. But you should still provide as much information on reasoning and logic as you can.

However, in other settings, there will be relatively few risks associated with giving people more detail on the reasons for decisions. In fact, it will often help individuals to legitimately adjust their behaviour or the choices they make in order to achieve a desirable decision outcome for both parties.

You should consider this as part of your initial risk or impact assessment for your AI model. It may form part of your DPIA. Start with the assumption that you will be as open and transparent as possible about the rationale of your AI-assisted decisions, and work back from there to limit what you tell people if you decide this is necessary. You should justify and document your reasons for this.

What are the risks of not explaining AI decisions?

### **Regulatory action**

While we cannot speak for other regulators, a failure to meet legal requirements around explaining AI-assisted decisions and treating people fairly may lead a regulator to take regulatory intervention or action. The ICO uses education and engagement to promote compliance by the organisations we regulate. But if the rules are broken, organisations risk formal action, including mandatory audits, orders to cease

processing personal data, and fines.

## **Reputational damage**

Public and media interest in AI is increasing, and often the spotlight falls on organisations that get things wrong. If you don't provide people with explanations of AI-assisted decisions you make about them, you risk being left behind by organisations that do, and getting singled out as unethical and uncaring towards your customers and citizens.

## **Disengaged public**

Not explaining AI-assisted decisions to individuals may leave them wary and distrustful of how and why AI systems work. If you choose not to do this, you risk a disengaged public that is slower to embrace, or even reject AI more generally.

# What goes into an explanation?

## At a glance

You need to consider how to provide information on two subcategories of explanation:

- 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.

There are different ways of explaining AI decisions. We have identified six main types of explanation:

- **Rationale explanation:** the reasons that led to a decision, delivered in an accessible and non-technical way.
- **Responsibility explanation:** who is involved in the development, management and implementation of an AI system, and who to contact for a human review of a decision.
- **Data explanation:** what data has been used in a particular decision and how.
- **Fairness explanation:** steps taken across the design and implementation of an AI system to ensure that the decisions it supports are generally unbiased and fair, and whether or not an individual has been treated equitably.
- **Safety and performance explanation:** steps taken across the design and implementation of an AI system to maximise the accuracy, reliability, security and robustness of its decisions and behaviours.
- **Impact explanation:** steps taken across the design and implementation of an AI system to consider and monitor the impacts that the use of an AI system and its decisions has or may have on an individual, and on wider society.

## In more detail

- [What do we mean by 'explanation'?](#)
- [Process-based vs outcome-based explanations](#)
- [Rationale explanation](#)
- [Responsibility explanation](#)
- [Data explanation](#)
- [Fairness explanation](#)
- [Safety and performance explanation](#)
- [Impact explanation](#)

## What do we mean by 'explanation'?

The Cambridge dictionary defines 'explanation' as:



“The details or reasons that someone gives to make something clear or easy to understand.”

While this is a general definition, it remains valid when considering how to explain AI- assisted decisions to the individuals affected (who are often also data subjects). It suggests that you should not always approach explanations in the same way. What people want to understand, and the ‘details’ or ‘reasons’ that make it ‘clear’ or ‘easy’ for them to do so may differ.

Our own research, and that of others, reveals that context is a key aspect of explaining decisions involving AI. Several factors about the decision, the person, the application, the type of data, and the setting, all affect what information an individual expects or finds useful.

Therefore, when we talk about explanations in this guidance, we do not refer to just one approach to explaining decisions made with the help of AI, or to providing a single type of information to affected individuals. Instead, the context affects which type of explanation you use to make an AI-assisted decision clear or easy for individuals to understand.

You should remember which audience you are aiming your explanation at, such as:

- staff whose decisions are supported by the AI system and who need to relay meaningful information to an individual affected by the AI-assisted decisions;
- those affected by the decision, with particular thought given to:
  - vulnerable groups; and
  - children; or
- auditors or external reviewers who are charged with monitoring or overseeing the production and deployment of the system.

Each group may require different levels of detail within the explanation they receive. The level of knowledge that the explanation recipient has about the subject of the explanation will affect the detail and language you need to use.

You should also take into account the transparency requirements of the GDPR, which (at least in cases of solely automated AI decisions) includes:

- providing meaningful information about the logic, significance and envisaged consequences of the AI decision;
- the right to object; and
- the right to obtain human intervention.

Where there is a “human in the loop” you still have to comply with the transparency requirements. In these situations, you should consider information about the decisions or recommendations made by the system and how this informs the human decision.

As a result of our research and engagement we identified six different types of explanation. You can combine these into your explanation in various ways depending on the specific decision and the audience

you are clarifying that decision about. You may not need to supply a decision recipient with information about all of these explanation types. However, you should consider what information will be required by all affected individuals, as well as the context in which that decision will be made, and plan your explanation accordingly.

You should also keep in mind that your system will need to be appropriately explainable to others who are, or may be, involved in the decision process. This could include, for example, implementers using the system or auditors checking up on it. These explanation types are designed to help you do this in a concise and clear way.

## Process-based vs outcome-based explanations

Before we explore the six explanation types, it is useful to make a distinction, which applies to all of them, between process-based and outcome-based explanations.

The primary aim of explaining fully automated or AI-assisted decisions is justifying a particular result to the individual whose interests are affected by it. This means:

- demonstrating how you and all others involved in the development of your system acted responsibly when choosing the processes behind its design and deployment; and
- making the reasoning behind the outcome of that decision clear.

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 outcome 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.

## Rationale explanation

### What does this explanation help people to understand?

It is about the 'why?' of an AI decision. It helps people understand the reasons that led to a decision

outcome, in an accessible way.

## **What purposes does this explanation serve?**

- **Challenging a decision**

It is vital that individuals understand the reasons underlying the outcome of an automated decision, or a human decision that has been assisted by the results of an AI system. If the decision was not what they wanted or expected, this allows them to assess whether they believe the reasoning of the decision is flawed. If they wish to challenge the decision, knowing the reasoning supports them to formulate a coherent argument for why they think this is the case.

- **Changing behaviour**

Alternatively, if an individual feels the reasoning for the decision was sound, they can use this knowledge to consider how they might go about changing their behaviour, or aspects of their lifestyle, to get a more favourable outcome in the future. If the individual is already satisfied with the outcome of the AI decision, the rationale explanation may still be useful so that they may validate their belief of why this was the case, or adjust it if the reasons for the favourable outcome were different to those they expected.

## **What you may 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 how the system's results apply to the concrete context and life situation of the affected individual.

Rationale explanations 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?

## **What information goes into rationale explanations**

As with the other types of explanation, rationale explanations can be process-based or outcome-based.

Process-based explanations clarify:

- How the procedures you have set up help you provide meaningful explanations of the underlying logic of your AI model's results.



- How these procedures are suitable given the model's particular domain context and its possible impacts on the affected decision recipients and wider society.
- How you have set up your system's design and deployment workflow so that it is appropriately interpretable and explainable, including its data collection and pre-processing, model selection, explanation extraction, and explanation delivery procedures.

Outcome-based explanations provide:

- The formal and logical rationale of the AI system – how the system is verified against its formal specifications, so you can verify that the AI system will operate reliably and behave in accordance with its intended functionality.
- The technical rationale of the system's output – how the model's components (its variables and rules) transform inputs into outputs, so you know what role these components play in producing that output. By understanding the roles and functions of the individual components, it is possible to identify the features and parameters that significantly influence a particular output.
- Translation of the system's workings – its input and output variables, parameters and so on – into accessible everyday language, so you can clarify, in plain and understandable terms, what role these factors play in reasoning about the real-world problem that the model is trying to address or solve.
- Clarification of how a statistical result is applied to the individual concerned. This should show how the reasoning behind the decision takes into account the specific circumstances, background and personal qualities of affected individuals.

The GDPR also refers 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 of Part 2 of this guidance for more detail about how to do this.

## **How can the guidance can help me with this?**

See Part 2 ([Explaining AI in practice](#)) for more information on extracting the technical rationale and translating this into understandable and context-sensitive reasons.

## **Responsibility explanation**

### **What does this explanation help people to understand?**

It helps people understand 'who' is involved in the development and management of the AI model, and 'who' to contact for a human review of a decision.

### **What purposes does this explanation serve?**

- **Challenging a decision**

Individuals in receipt of other explanations, such as rationale or fairness, may wish to challenge the AI decision based on the information provided to them. The responsibility explanation helps by directing the individual to the person or team responsible for carrying a human review of a decision. It also makes accountability traceable.

## • Informative

This explanation can also serve an informative purpose by shedding some light on the different parts of your organisation involved in the design and deployment of your AI decision-support system.

### What you may need to show

- 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.
- Definitions of 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 responsibility explanations

Process-based explanations clarify:

- The roles and functions across your organisation that are involved in the various stages of developing and implementing your AI system, including any human involvement in the decision-making. If your system, or parts of it, are procured, you should include information about the providers or developers involved.
- Broadly, what the roles do, why they are important, and where overall responsibility lies for management of the AI model – who is ultimately accountable.
- 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 explanations:

Because a responsibility explanation largely has to do with the governance of the design and implementation of AI systems, it is, in a strict sense, entirely process-based. Even so, there is important information about post-decision procedures that you should be able to provide:

- 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).
- Give individuals a way 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 have implemented the decision, and used the statistical results of a decision-support system to come to a determination about an individual.

### How can the guidance help me with this?

See Part 3 of this guidance ([What explaining AI means for your organisation](#)) for more information on identifying the roles involved in explaining an AI-assisted decision. See Part 2 of this guidance ([Explaining AI in practice](#)) for more details on the information you need to provide for this explanation.

### Data explanation

### What does this explanation help people to understand?

Data explanations are about the 'what' of AI-assisted decisions. They help people understand what data about them, and what other sources of data, were used in a particular AI decision. Generally, they also help individuals understand more about the data used to train and test the AI model. You could provide some of this information within the fair processing notice you are required to provide under Articles 13 and 14 of the GDPR.

## **What purposes does this explanation serve?**

- **Challenging a decision**

Understanding what data was input into an AI decision-support system will allow an individual to challenge the outcome if they think it was flawed (eg if some of the input data was incorrect or irrelevant, or additional data wasn't taken into account that the individual thinks is relevant).

- **Providing reassurance**

Knowing more about the actions you took when collecting and preparing the training and test data for your AI model will help to reassure individuals that you made appropriate and responsible choices in the best interests of developing an understandable, fair and accurate AI decision-support system.

## **What you may need to show**

- How the data used to train, test, and validate your AI model was managed and utilised from collection through processing and monitoring.
- What data you used in a particular decision and how.

## **What information goes into data explanations**

Process-based explanations include:

- What training/ testing/ validating data was collected, the sources of that data, and the methods that were used to collect it.
- Who took part in choosing the data to be collected or procured and who was involved in its recording or acquisition. How procured or third-party provided data was vetted.
- How data quality was assessed and the steps that were taken to address any quality issues discovered, such as completing or removing data.
- What the training/ testing/ validating split was and how it was determined.
- How data pre-processing, labelling, and augmentation supported the interpretability and explainability of the model.
- What measures were taken to ensure the data used to train, test, and validate the system was representative, relevant, accurately measured, and generalisable.
- How you ensured that any potential bias and discrimination in the dataset have been mitigated.

Outcome-based explanations:

- Clarify the input data used for a specific decision, and the sources of that data. This is outcome-based because it refers to your AI system's result for a particular decision recipient.
- In some cases the output data may also require an explanation, particularly where the decision recipient

has been placed in a category which may not be clear to them. For example, in the case of anomaly detection for financial fraud identification, the output might be a distance measure which places them at a certain distance away from other people based on their transaction history. Such a classification may require an explanation.

## How can the guidance help me with this?

See the data collection section in Part 2 ([Explaining AI in practice](#)) for more on deriving this explanation type.

In Part 3 ([What explaining AI means for your organisation](#)) we provide some further pointers on how to document and demonstrate responsible data management practices across the design and implementation of your AI model.

## Fairness explanation

### What does this explanation help people to understand?

The fairness explanation is about helping people understand the steps you took (and continue to take) to ensure your AI decisions are generally unbiased and equitable. It also gives people an understanding of whether or not they have been treated equitably themselves.

### What purposes does this explanation serve?

- **Trust**

The fairness explanation is key to increasing individuals' confidence in your AI system. You can foster trust by explaining to an individual how you avoid bias and discrimination in the AI-assisted decisions you make and by proving that they were not treated differently than others like them.

- **Challenging a decision**

It also allows individuals to challenge a decision made using an AI system. An individual might feel the explanation you provide actually suggests they were treated unfairly.

### What you may need to show

An explanation of fairness can relate to several stages of the design, development and deployment of AI systems:

**Dataset fairness:** The system is trained and tested on properly representative, relevant, accurately measured, and generalisable datasets (note that this dataset fairness component will overlap with data explanation). This may include showing that you have made sure your data is:

- as representative as possible of all those affected;
- sufficient in terms of its quantity and quality, so it represents the underlying population and the phenomenon you are modelling;
- assessed and recorded through suitable, reliable and impartial sources of measurement and has been sourced through sound collection methods;

- up-to-date and accurately reflects the characteristics of individuals, populations and the phenomena you are trying to model; and
- 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. This may include showing that you have done the following:

- Attempted to identify any underlying structural biases that may play a role in translating your objectives into target variables and measurable proxies. When defining the problem at the start of the AI project, these biases could influence what system designers expect target variables to measure and what they statistically represent.
- Mitigated bias in the data pre-processing phase by taking into account the sector or organisational context in which you are operating. When this process is automated or outsourced, show that you have reviewed what has been done, and maintained oversight. You should also attach information on the context to your metadata, so that those coming to the pre-processed data later on have access to the relevant properties when they undertake bias mitigation.
- Mitigated bias when the feature space was determined (ie when relevant features were selected as input variables for your model). Choices made 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.
- Mitigated bias 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.
- Mitigated bias by watching 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 it affects. This may include showing that:

- you have been explicit about the formal definition(s) of fairness you have chosen and why. 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; and
- the method you have applied in operationalising your formalised fairness criteria, for example, 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. This may include showing that you have appropriately prepared and trained the implementers of your system 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);
- use its results with an active awareness of the specific context in which they are being applied. They should understand the particular circumstances of the individual to which that output is being applied;

and

- understand the limitations of the system. This includes understanding the statistical uncertainty associated with the result as well as the relevant error rates and performance metrics.

## **What information goes into fairness explanations**

This explanation is about providing people with appropriately simplified and concise information on the considerations, measures and testing you carry out to make sure that your AI system is equitable and that bias has been optimally mitigated. Fairness considerations come into play through the whole lifecycle of an AI model, from inception to deployment, monitoring and review.

Process-based explanations include details about:

- your chosen measures to mitigate risks of bias and discrimination at the data collection, preparation, model design and testing stages;
- how these measures were chosen and how you have managed informational barriers to bias-aware design such as limited access to data about protected or sensitive traits of concern; and
- the results of your initial (and ongoing) fairness testing, self-assessment, and external validation – showing that your chosen fairness measures are deliberately and effectively being integrated into model design. 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 explanations include:

- details about how your formal fairness criteria were implemented in the case of a particular decision or output;
- presentation of the relevant fairness metrics and performance measurements in the delivery interface of your model. This should be geared to a non-technical audience and done in an easily understandable way; and
- explanations of 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.

## **How can the guidance help me with this?**

See Part 2 ([Explaining AI in practice](#)) for more information on building fairness into the design and deployment of your AI model. See also Part 3 ([What explaining AI means for your organisation](#)) for information on how to document what you have done to achieve fairness.

## **Safety and performance explanation**

### **What does this explanation help people to understand?**

The safety and performance explanation helps people understand the measures you have put in place, and the steps you have taken (and continue to take) to maximise the accuracy, reliability, security and robustness of the decisions your AI model helps you to make. It can also be used to justify the type of AI system you have chosen to use, such as comparisons to other systems or human decision makers.

## What purposes does this explanation serve?

- **Reassurance**

Individuals often want to be reassured that an AI system is safe and reliable. The safety and performance explanation helps to serve this purpose by demonstrating what you have done to test and monitor the accuracy, reliability, security and robustness of your AI model.

- **Informative**

If an individual receiving an explanation of an AI-assisted decision is technically knowledgeable or proficient, this explanation will allow them to assess the suitability of the model and software for the types of decision being made. This explanation helps you to be as transparent as you can with people about the integrity of your AI decision-support system.

- **Challenging a decision**

Individuals can make an informed choice about whether they want to contest an AI decision on the basis that it may be incorrect for them, or carried out in an unsafe, hazardous, or unreliable way. This is closely linked with challenging a decision on the basis of fairness.

## What you may 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 in practice. Individuals may want to know how well the system works if things go wrong, how this has been anticipated and tested, and how the system has been immunised from adversarial attacks.

## What information goes into safety and performance explanations

Process-based explanations include:

For accuracy:

- How you measure it (eg maximising precision to reduce the risk of false negatives).
- Why you chose those measures, and how you went about assuring it.
- 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 to whom the results apply.

- What kinds of external validation you have undertaken to test and confirm your model's 'ground truth'.
- What the overall accuracy rate of the system was at testing stage.
- What you do to monitor this (eg measuring for concept drift over time).

For reliability:

- How you measure it and how you went about assuring it.
- Results of the formal verification of the system's programming specifications, ie how encoded requirements have been mathematically verified.

For security:

- How you measure it and how you went about assuring it, eg how limitations have been set on who is able to access the system, when, and how.
- How you manage the security of confidential and private information that is processed in the model.

For robustness:

- How you measure it.
- Why you chose those measures.
- How you went about assuring it, 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).

Outcome-based explanations:

While you may not be able to guarantee accuracy at an individual level, you should be able to provide assurance that, at run-time, your AI system operated reliably, securely, and robustly for a specific decision.

- 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). Include 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).

## How can the guidance help me with this?

See Part 2 ([Explaining AI in practice](#)) for more information on ensuring the accuracy, reliability, security and robustness of your AI system. See also Part 3 ([What explaining AI means for your organisation](#)) for information on how to document what you have done to achieve these objectives.

Impact explanation

## What does this explanation help people to understand?



An impact explanation helps people understand how you have considered the effects that your AI decision-support system may have on an individual, ie what the outcome of the decision means for them. It is also about helping individuals to understand the broader societal effects that the use of your system may have. This may help reassure people that the use of AI will be of benefit. Impact explanations are therefore often well suited to delivery before an AI-assisted decision has been made. See Task 6 of Explaining AI in practice for guidance on when to deliver explanations.

## **What purposes does this explanation serve?**

- **Consequences**

The purpose of the impact explanation is primarily to give individuals some power and control over their involvement in an AI-assisted decision made about them. By understanding the possible consequences of the decision (negative, neutral and positive) an individual can better assess their willingness to take part in the process, and can anticipate how the outcomes of the decision may affect them.

- **Reassurance**

Knowing that you took the time to consider and manage the potential effects that your AI system has on society can help to reassure individuals that issues such as safety, equity, and reliability are core components of the AI model they are subject to. It also helps individuals to be more informed about the benefits and risks of AI decision-support systems, and therefore, more confident and active in the debate about its development and use.

## **What you may need to show**

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

## **What information goes into impact explanations**

Process-based explanations include:

- Showing the considerations you gave to your AI system's potential effects, how you undertook these considerations, and the measures and steps you took to mitigate possible negative impacts on society, and to amplify the positive effects.
- Information about how you plan to monitor and re-assess impacts while your system is deployed.

Outcome-based explanations:

Although the impact explanation is mainly about demonstrating that you have put appropriate forethought into the potential 'big picture' effects, you should also consider how to help decision recipients understand the impact of the AI-assisted decisions that specifically affect them. For instance, you might explain the consequences for the individual of the different possible decision outcomes and how, in some cases, changes in their behaviour would have brought about a different outcome with more positive impacts. This use of counterfactual assessment would help decision recipients make changes that could lead to a different outcome in the future, or allow them to challenge the decision.

## How can the guidance help me with this?

See Part 2 ([Explaining AI in practice](#)) for more information on considering impact in how you select an appropriately explainable AI model. See also Part 3 ([What explaining AI means for your organisation](#)) for information on how to document this.

# What are the contextual factors?

## At a glance

Five contextual factors have an effect on the purpose an individual wishes to use an explanation for, and on how you should deliver your explanation:

- domain you work in;
- impact on the individual;
- data used;
- urgency of the decision; and
- audience it is being presented to.

## In more detail

- [Introduction to the contextual factors](#)
- [Domain factor](#)
- [Impact factor](#)
- [Data factor](#)
- [Urgency factor](#)
- [Audience factor](#)

### Introduction

When constructing an explanation for an individual, there are several factors about the context an AI-assisted decision is made in. These 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 for which explanations to prioritise in delivering an explanation of an AI-assisted. You should consider these factors at all stages of the process outlined in Part 2 of this guidance.

When considering these contextual factors, keep in mind that providing explanations to decision recipients will also educate them about AI systems. It may therefore be worth thinking about the information you can provide in advance of processing in order to help develop knowledge and understanding of AI use among the general public.

### Domain factor

#### **What is this factor?**

By 'domain', we mean the setting or the sector you deploy your AI model in 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 from an explanation. For example, a person receiving an AI-assisted mortgage decision will expect to learn about the reasoning behind the determination in a way that matches established lending standards and practices.

### **Which explanations should we prioritise?**

Considering the domain factor is perhaps the most crucial determiner of what explanations you should include and prioritise 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 are likely to want you to provide a fairness explanation.

In lower-stakes domains such as e-commerce, it is unlikely that people will want or expect extensive explanations of the safety and performance of the outputs of recommender systems. Even so, in these lower impact domains, you should explain 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). These 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 about 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. For example, 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. You should do it on a case by case basis, and consider it in combination with all the other contextual factors. It should also involve inclusive dialogue across the fields of expertise that are involved in the design, development, and deployment of the AI system. Getting together different team members, who have technical, policy, compliance, and domain expertise can provide a more informed vision of the impact factor of an AI model.

## 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.

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 about 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 may want to make these changes if they don’t agree with the outcome.

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 may 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 you need to take that action.

### **Which explanations should we prioritise?**

Where urgency is a key factor, 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. Therefore, 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 you are explaining an AI-assisted decision to. 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 about what the decision is about? Are a broad range of people subject to decisions you make (eg the UK general public), which indicates 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 whether the decision recipients require any reasonable adjustments in how they receive the explanation (Equality Act 2010).

As a general rule, it is a good idea to accommodate the explanation needs of the most vulnerable individuals. You should ensure that these decision recipients are able to clearly understand the information that you are giving them. Using plain, non-technical language and visualisation tools, where possible, may often help.

Note as well that, while we are focusing on the decision recipient, you are also likely to have to put significant forethought into how you will provide other audiences with appropriate information about the outputs of your AI model. For instance, in cases where the models are supporting decision-making, you will have to provide the end-users or implementers of these models with a depth and level of explanation that is appropriate to assist them in carrying out evidence-based reasoning in way that is context-sensitive and aware of the model's limitations. Likewise, in instances where models and their results are being reviewed by auditors, you will have to provide information about these systems at a level and depth that is fit for the purpose of the relevant review.

## **Which explanations should we prioritise?**

If the people you are making AI-assisted decisions about 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, as they are 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 likely the people will 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 ask about an AI decision.

Of course, even for those with little knowledge of an area, the rationale explanation can still be useful to explain 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. The expert can then 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).

# The principles to follow

## At a glance

To ensure that the decisions you make using AI are explainable, you should follow four principles:

- be transparent;
- be accountable;
- consider the context you are operating in; and,
- reflect on the impact of your AI system on the individuals affected, as well as wider society.

## In more detail

- [Why are principles important?](#)
- [What are the principles?](#)
- [Be transparent](#)
- [Be accountable](#)
- [Consider context](#)
- [Reflect on impacts](#)
- [How do these principles relate to the explanation types?](#)

### Why are principles important?

AI-assisted decisions are not unique to one sector, or to one type of organisation. They are increasingly used in all areas of life. This guidance recognises this, so you can use it no matter what your organisation does. The principles-based approach of this guidance gives you a broad steer on what to think about when explaining AI-assisted decisions to individuals. Please note that these principles relate to providing explanations of AI-assisted decision to individuals, and complement the data protection principles in the GDPR.

The first two principles – be transparent and be accountable – share their names with GDPR principles, as they are an extension of the requirements under GDPR. This means that you are required to comply with your obligations under the GDPR, but we provide further guidance that should enable you to follow ‘best practice’ when explaining AI decisions.

### What are the principles?

Each principle has two key aspects detailing what the principles are about and what they mean in practice. Parts of the guidance that support you to act in accordance with the different aspects of each principle are signposted.

### Be transparent



## What is this principle about?

The principle of being transparent is an extension of the transparency aspect of principle (a) in the GDPR (lawfulness, fairness and transparency).

In data protection terms, transparency means being open and honest about who you are, and how and why you use personal data.

Being transparent about AI-assisted decisions builds on these requirements. It is about making your use of AI for decision-making obvious and appropriately explaining the decisions you make to individuals in a meaningful way.

## What are the key aspects of being transparent?

Raise awareness:

- Be open and candid about:
  - your use of AI-enabled decisions;
  - when you use them; and
  - why you choose to do this.
- Proactively make people aware of a specific AI-enabled decision concerning them, in advance of making the decision.

Meaningfully explain decisions:

Don't just giving **any** explanation to people about AI-enabled decisions - give them:

- a truthful and meaningful explanation;
- written or presented appropriately; and
- delivered at the right time.

(This is closely linked with the context principle.)

## How can this guidance help us be transparent?

To help with raising awareness about your use of AI decisions read:

- Policies and procedures section of 'What Explaining AI means for your organisation'; and
- Proactive engagement section in Task 6 of 'Explaining AI in Practice'.

To support you with meaningfully explaining AI decisions read:

- Policies and procedures section of 'What explaining AI means for your organisation';
- Building your system to aid in a range of explanation types in Task 3 of 'Explaining AI in Practice'.
- Selecting your priority explanations in Task 1 of 'Explaining AI in Practice'.
- Explanation timing in Task 6 of 'Explaining AI in practice'.

Be accountable

## What is this principle about?

The principle of being accountable is derived from the accountability principle in the GDPR.

In data protection terms, accountability means taking responsibility for complying with the other data protection principles, and being able to demonstrate that compliance. It also means implementing appropriate technical and organisational measures, and data protection by design and default.

Being accountable for explaining AI-assisted decisions concentrates these dual requirements on the processes and actions you carry out when designing (or procuring/ outsourcing) and deploying AI models.

It is about ensuring appropriate oversight of your AI decision systems, and being answerable to others in your organisation, to external bodies such as regulators, and to the individuals you make AI-assisted decisions about.

## What are the key aspects of being accountable?

Assign responsibility:

- Identify those within your organisation who manage and oversee the 'explainability' requirements of an AI decision system, and assign ultimate responsibility for this.
- Ensure you have a designated and capable human point of contact for individuals to query or contest a decision.

Justify and evidence:

- Actively consider and make justified choices about how to design and deploy AI models that are appropriately explainable to individuals.
- Take steps to prove that you made these considerations, and that they are present in the design and deployment of the models themselves.
- Show that you provided explanations to individuals.

## How can this guidance help us be accountable?

To help with assigning responsibility for explaining AI decisions read:

- the Organisational roles and Policies and procedures sections of 'What explaining AI means for your organisation'.

To support you with justifying the choices you make about your approach to explaining AI decisions read:

- all the tasks in 'Explaining AI in practice'.

To help you evidence this read:

- the Policies and procedures and Documentation sections of 'What explaining AI means for your organisation'.

Consider context

## What is this principle about?

There is no one-size-fits-all approach to explaining AI-assisted decisions. The principle of considering context underlines this.

It is about paying attention to several different, but interrelated, elements that can have an effect on explaining AI-assisted decisions, and managing the overall process.

This is not a one-off consideration. It's something you should think about at all stages of the process, from concept to deployment and presentation of the explanation to the decision recipient.

There are therefore several types of context that we address in this guidance. These are outlined in more detail in the 'contextual factors' section above.

## **What are the key aspects of considering context?**

Choose appropriate models and explanation:

When planning on using AI to help make decisions about people, you should consider:

- the setting in which you will do this;
- the potential impact of the decisions you make;
- what an individual should know about a decision, so you can choose an appropriately explainable AI model; and
- prioritising delivery of the relevant explanation types.

Tailor governance and explanation:

Your governance of the 'explainability' of AI models should be:

- robust and reflective of best practice; and
- tailored to your organisation and the particular circumstances and needs of each decision recipient.

## **How can this guidance help us consider context?**

To support your choice of appropriate models and explanations for the AI decisions you make read:

- 'Explaining AI in practice'.
- The Contextual factors section of this document.

To help you tailor your governance of the explainability of AI decision systems you use read:

- the Organisational roles and Policies and procedures sections of 'What explaining AI means for your organisation'.

Reflect on impacts

## **What is this principle about?**

In making decisions and performing tasks that have previously required the thinking and reasoning of responsible humans, AI systems are increasingly serving as trustees of human decision-making. However, individuals cannot hold these systems directly accountable for the consequences of their outcomes and

behaviours.

The value of reflecting on the impacts of your AI system helps you explain to individuals affected by its decisions that the use of AI will not harm or impair their wellbeing.

This means asking and answering questions about the ethical purposes and objectives of your AI project at the initial stages.

You should then revisit and reflect on the impacts identified in the initial stages of the AI project throughout the development and implementation stages. If any new impacts are identified, you should document them, alongside any mitigating factors you implement where relevant. This will help you explain to decision recipients what impacts you have identified and how you have reduced any potentially harmful effects as far as possible.

## **What are the key aspects of reflecting on impacts?**

Individual wellbeing:

Think about how to build and implement your AI system in a way that:

- fosters the physical, emotional and mental integrity of affected individuals;
- ensures their abilities to make free and informed decisions about their own lives;
- safeguards their autonomy and their power to express themselves;
- supports their abilities to flourish, to fully develop themselves, and to pursue their interests according to their own freely determined life plans;
- preserves their ability to maintain a private life independent from the transformative effects of technology; and
- secures their capacities to make well-considered, positive and independent contributions to their social groups and to the shared life of the community, more generally.

Wellbeing of wider society:

Think about how to build and implement your AI system in a way that:

- safeguards meaningful human connection and social cohesion;
- prioritises diversity, participation and inclusion;
- encourages all voices to be heard and all opinions to be weighed seriously and sincerely;
- treats all individuals equally and protects social equity;
- uses AI technologies as an essential support for the protection of fair and equal treatment under the law;
- utilises innovation to empower and to advance the interests and well-being of as many individuals as possible; and
- anticipates the wider impacts of the AI technologies you are developing by thinking about their ramifications for others around the globe, for the biosphere as a whole and for future generations.

## **How can this guidance help us reflect on impacts?**

For help with reflecting on impacts read:

---

- the different types of explanation above; and
- 'Explaining AI in practice'.

To support you with justifying the choices you make about your approach to explaining AI decisions read:

- the different types of explanation above; and
- ['Explaining AI in practice'](#).

To help you evidence this read:

- the [Policies and procedures](#) and [Documentation](#) sections of 'What explaining AI means for your organisation'.

## How do the principles relate to the explanation types?

The principles are important because they underpin how you should explain AI-assisted decisions to individuals. Here we set out how you can put them into practice by directly applying them through the explanations you use:

Principle	AI explanation and relevant considerations
<b>Be transparent</b>	<p><b>Rationale</b></p> <ul style="list-style-type: none"> <li>• What is the technical logic or reasoning behind the model's output?</li> <li>• Which input features, parameters and correlations played a significant role in the calculation of the model's result and how?</li> <li>• How can you explain the technical rationale underlying the model's output in easily understandable reasons that may be subjected to rational evaluation by affected individuals or their representatives?</li> <li>• How can you apply the statistical results to the specific circumstances of the individual receiving the decision?</li> </ul> <p><b>Data</b></p> <ul style="list-style-type: none"> <li>• What data did you use to train the model?</li> <li>• Where did the data come from?</li> <li>• How did you ensure the quality of the data you used?</li> </ul>
<b>Be accountable</b>	<p><b>Responsibility</b></p> <ul style="list-style-type: none"> <li>• Who is accountable at each stage of the AI system's design and deployment, from the initial phase of defining outcomes to the concluding phase of providing explanations?</li> <li>• What are the mechanisms they will be held accountable by?</li> <li>• How have you made design and implementation processes traceable and auditable across the entire project?</li> </ul>

## **Consider context**

See Task 1 of 'Explaining AI in practice' for more information on how context matters when choosing which explanation type to use, and which AI model.

See the section above on contextual factors to see how these can help you choose which explanation types to prioritise in presenting your explanation to the decision recipient.

## **Reflect on impacts**

### **Fairness**

- Do the AI system's outputs have discriminatory effects?
- Have you sufficiently integrated the objectives of preventing discrimination and of mitigating bias into the design and implementation of the system?
- Have you incorporated formal criteria of fairness that determine the distribution of outcomes into the system and made these explicit to customers in advance?
- Has the model prevented discriminatory harm?

### **Safety and performance**

- Is the AI system safe and technically sustainable when operating in practice?
- Is the system's operational integrity worthy of public trust?
- Have you designed, verified, and validated the model in a way that sufficiently ensures that it is secure, accurate, reliable, and robust?
- Have you taken sufficient measures to ensure that the system dependably operates in accordance with its designers' expectations when confronted with unexpected changes, anomalies, and perturbations?

### **Impact**

- Have you sufficiently considered impacts on the wellbeing of affected individuals and communities from start to finish of the AI model's design and deployment?

# Part 2: Explaining AI in practice

## About this guidance

### What is the purpose of this guidance?

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

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

### How should we use this guidance?

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

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

### What is the status of this guidance?

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

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

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

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

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



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

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



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

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



# Summary of the tasks to undertake

We have set out a number of tasks both to help you design and deploy appropriately explainable AI systems and to assist you in providing clarification of the results these systems produce to a range of affected individuals (from operators, implementers, and auditors to decision recipients).

These tasks offer a systematic approach to:

- developing AI models in an explanation-aware fashion; and
- selecting, extracting and delivering explanations that are differentiated according to the needs and skills of the different audiences they are directed at.

They should help you navigate the detailed technical recommendations in this part. However, we recognise that in practice some tasks may be concurrent, cyclical, and iterative rather than consecutive or linear, and you may wish to develop your own plan for doing this.

In [Annexe 1](#) we give an example of how these tasks may be carried out in a particular case in the health sector.

## **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 use, or plan to use, AI to make decisions about people. Or perhaps some of the explanations we identify are not particularly relevant to your organisation and the people you make decisions about.

That's absolutely fine. The explanation types 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 you are deploying your AI decision-support system in. Considering the domain you work in, the particular use case and possible impacts of your system on individuals and wider society 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 you should discard the remaining explanations.

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 to make all the explanations you have identified as relevant 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 and pre-process your data in an explanation-aware manner**

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. This task therefore emphasises some of the things you should think about when you are at these stages of your design process, and how this can contribute to the information you provide to individuals for each explanation type.

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

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 for 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 easily understandable language to justify the outcome.

For example, if 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.

You should then translate these factors 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 how to build and present your explanation**

Finally, you should think about how you will build and present your explanation to an individual, whether you are doing this through a website or app, in writing or in person.

Considerations of context should be the cornerstone of building and presenting your explanation. You should consider contextual factors (domain, impact, data, urgency, audience) to help you decide how you should deliver appropriate information to the individual.

Considering context should also help you to customise what sort of (and how much) information to provide to decision recipients. The way you explain the results of your model to decision subjects may be quite different from how you provide information to other relevant audiences like auditors, who may also need explanations, though with different degrees of depth and technical detail. Differentiating the way you are providing information in an audience-responsive manner can help you avoid creating explanation fatigue in your customers (by saying too much) and at the same time allow you to protect your intellectual property and safeguard your system from being gamed.

When delivering an explanation to decision recipients, 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.

# Task 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. This means you should prioritise the rationale and responsibility explanation types.
- The setting and sector you are working in is important in working 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 decide 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 to selecting priority explanations](#)
- [Familiarise yourself with the different types of explanation](#)
- [Prioritise rationale and responsibility explanation](#)
- [Consider domain or sector context and use case](#)
- [Consider potential impacts](#)

- [Examples for choosing suitable explanation types](#)

## 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 you can provide clear and accessible process-based explanations when you design and deploy your AI system; 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.

Considering how to address these explanation types at the beginning of your process should provide you with a reasonable understanding of how your system works and who is responsible at each stage of the process. This will also mean the information is available for decision recipients when an explanation is provided.

Please note that although we recommend you prioritise documenting information about the rationale and responsibility explanations at the start of the process, you may not wish to provide this in the first layer of explanation to the decision subject (Task 6). However, we do recommend that this information is provided as part of the explanation where practical.

## 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 it will be used in.

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.

By understanding your AI application's domain context and setting, you may also gain insight into people's expectations of the content and scope of similar explanations previously offered. Doing due diligence and researching these sorts of sector-specific expectations will help you to draw on background knowledge as you decide 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 you should include as part of the 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 complainants 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.

### Further reading

See more on the types of explanation in the link for ['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 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. An accurate and reliable rationale explanation will also better support the evidence-based judgement of the medical professionals involved.
- **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. Developers should consult domain-specific requirements and standards to determine the scope, depth, and types of explanation that are reasonable expected.
- **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. Depending on the specific use case, the fairness explanation might also be important to prioritise, for instance, to demonstrate that the data collected is of sufficient quality, quantity, and representativeness for the system to perform accurately for different demographic groups.



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

## At a glance

- The data that you collect and pre-process before inputting it into your system has an important role to play in your ability to derive each explanation type.
- Careful labelling and selection of input data can help provide information for your rationale explanation.
- To be more transparent you may wish to provide details about who is responsible at each stage of data collection and pre-processing. You could provide this as part of your responsibility explanation.
- To aid your data explanation, you could include details on:
  - the source of the training data;
  - how it was collected;
  - assessments about its quality; and
  - steps taken to address quality issues, such as completing, augmenting, or removing data.
- You should check the data used within your model to ensure it is sufficiently representative of those you are making decisions about. You should also consider whether pre-processing techniques, such as re-weighting, are required. These will help your fairness explanation.
- You should ensure that the modelling, testing and monitoring stages of your system development lead to accurate results to aid your safety and performance explanation.
- Documenting your impact and risk assessment, and steps taken throughout the model development to implement these assessments, will help in your impact explanation.

## Checklist

- ☐ Our data is representative of those we 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, 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

- [Introduction to collection and pre-processing the data](#)
- [Rationale explanation](#)
- [Responsibility explanation](#)
- [Data explanation](#)
- [Fairness explanation](#)
- [Safety and performance explanation](#)
- [Impact explanation](#)

### Introduction

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. This task therefore emphasises some of the things you should think about when you are at these stages of your design process, and how this can contribute to the information you provide to individuals for each explanation type.

### Rationale explanation

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 as interpretable as possible. 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 this 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 task to assess how best to obtain an explanation of the rationale through appropriate model selection and approaches to explanation extraction.

## Responsibility explanation

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 Task 6 for more on layering explanations.

## Data explanation

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 (less directly linked to key areas of interest such as fairness and accuracy) there is still value in providing this information. 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 explanation

Fairness explanations are about giving people information on the steps you've 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. You may use pre-processing techniques such as re-weighting, up-weighting, masking, or even excluding features 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 your fairness explanations. Ensure that you appropriately document what you do at these early stages so you can reflect this in your explanation.

## Safety and performance explanation

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 explanation is mainly about 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 performance metrics chosen for a model, and about the various performance related measures you used to ensure optimal results.

## Impact explanation

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 individually and as a whole. It also involves demonstrating that you have 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 and risk assessment (eg a data protection impact assessment). But also 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.

You may have covered such steps in your fairness and safety and performance explanations (eg ensuring the collection of representative and up-to-date datasets). However, 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).

Example method for pre-processing data


## A provenance-based approach to pre-processing data

One approach to pre-processing data for the purposes of explanation is based on provenance. All the information, data dependencies and processes underpinning a decision are collectively known as the **provenance of the decision**. The PROV data model [PROV-DM] is a vocabulary for provenance, which was standardised at the World Wide Web Consortium. Organisations can use PROV to uniformly represent and link relevant information about the processes running around the AI model, and to seamlessly query it, in order to construct relevant explanations. In addition, for organisations that depend on **external** data for their decisions, PROV allows for the provenance of data to be linked up across organisation boundaries.

PROV is a standard vocabulary to encode provenance – a form of knowledge graph providing an account of what a system performed, including references to: people, data sets, and organisations involved in decisions; attribution of data; and data derivations. The provenance of a decision enables you to trace back an AI decision to its input data, and to identify the responsibility for each of the activities found along the way. It allows for an explicit record of where data comes from, who in the organisation was associated with data collection and processing, and what data was used to train the AI system. Such provenance information provides the foundations to generate explanations for an AI decision, as well as for making the processes that surround an AI decision model more transparent and accountable.


When PROV is adopted as a way of uniformly encoding the provenance of a decision within or across organisations, it becomes possible to perform a range of tasks. This includes being able to computationally query the knowledge graph capturing the information, data dependencies and processes underpinning the decision. You can then extract the relevant information to construct the desired explanation. Therefore, the approach will help automate the process of extracting explanations about the pipeline around an AI model. Those include explanations on the processes that led to the decision being made, who was responsible for what step in these processes, whether the AI model was solely responsible for the decision, what data from which source influenced the decision etc. However, currently, no work has yet addressed the ability to build explanations for the AI model itself from provenance, so you will need to couple it with another approach (such as the ones presented in Task 3).


This technique can be used to help provide information for the data explanation. It can also provide details for the responsibility, and safety and performance explanations.

There is an online demonstrator illustrating the provenance-based approach described above using a loan decision scenario at: <https://explain.openprovenance.org/loan/> .

### Further reading

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.

You can read more about the provenance-based approach to pre-processing data at <https://explain.openprovenance.org/report/> .

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

## 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 whether:
  - there are costs and benefits to replacing your current system with a newer and potentially less explainable AI model;
  - the data you use requires a more or less explainable system;
  - your use case and domain context encourage choosing an inherently interpretable system;
  - your processing needs lead you to select a 'black box' model; and
  - the supplementary interpretability tools that help you to explain a 'black box' model (if chosen) 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 it.
- ☐ 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 and documented these mitigation processes.
- ☐ We have made it clear how the model has been tested, including which parts of the data have been used to train the model, 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 an 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, accurate, 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 and have made sure to clearly convey this limitation to implementers and decision recipients.
- ☐ 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 for individual cases.

## In more detail

- [Introduction to building your system](#)
- [Select an appropriately explainable model](#)
- [Tools for extracting a rationale explanation](#)

### 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.

The selection of an appropriate model can also help you provide information about the safety and performance and, fairness explanations. Therefore, it is important that you carefully consider how you select your model. We would also recommend that you document your decision making process, so you can evidence that you have considered the impact your model selection will have on the decision recipient.

### Select an appropriately explainable model

Selecting an appropriate model is important whether you are procuring a model from an external vendor, or looking to build a bespoke system in-house. In both cases, you need to consider the following factors to ensure you select the most appropriate model for your needs. Where you are procuring a system, you may wish to ask the vendor about some or all of these elements, as you will need to understand the system in order to provide an appropriate explanation.

## Where do we start?

Before you consider the technical factors, you should consider:

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

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 or vendor 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 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.

### Further reading

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.

It may be helpful to group the kinds of data that you may use in your AI system into two categories:

- 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 you are processing social or demographic data (group i. above) 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 you are processing biological or physical data (group ii. above), only for the purposes of gaining scientific insight (eg predicting protein structures in genomics research), 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 these models are trained and tested on.
- In cases where you are processing both these groups of data 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 you are using unconventional data to support decisions that affect individuals, you should bear the following in mind:

- you can consider this data to be of the same type as group i. data above, and treat it 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; and
- 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 or generate inferences that meet reasonable expectations.

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 algorithmic techniques that are as interpretable as possible.

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 in Annexe 2) 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. As a result, the advantages gained by the combination of their improved performance and their transparency may outweigh those of less transparent approaches.

**'Black box' AI systems:** when you consider using opaque algorithmic techniques, make sure that the supplementary interpretability tools that you will use 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 are likely to be opaque when you are using AI applications to classify images, recognise speech, or detect 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 in Annexe 2.

You should only use 'black box' models if you have thoroughly considered their potential impacts and risks in advance. The members of your team should also 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. This needs to be reasonably sufficient to mitigate the potential risks of the system and provide 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 and in Annexe 3.

As part of the process-based aspect of the rationale explanation, you should document and keep a record of any deliberations that cover how you selected a 'black box' model.

**Hybrid methods – use of 'challenger' models:** when you select an interpretable model to ensure explainable data processing, you should only carry out parallel use of opaque 'challenger' models for purposes of feature engineering/selection, insight, or comparison if you do so in a transparent, responsible, and lawful manner.

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

'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, you should make the process explicit and document it. 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 you use challenger models to process the data of affected decision recipients – even for benchmarking purposes – you should properly record and document them. You should treat them as core production models, document them, and hold them to the same explainability standards, if you incorporate the insights from this challenger model's processing into any dimension of actual decision-making. For example, the comparative benchmarking results are shared with implementers/ users, who are making decisions.

## **What types of models are we choosing between?**

To help you get a better picture of the spectrum of algorithmic techniques, Annexe 2 lays out some of the basic properties, potential uses, and interpretability characteristics of the most widely used algorithms at present. These techniques are also listed in the table below.

The 11 techniques listed in the left column 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 four techniques in the right column are more or less considered to be 'black box' algorithms.

Broadly interpretable systems	Broadly “black box” systems
Linear regression (LR)	Ensemble methods
Logistic regression	Random Forest
Generalised linear model (GLM)	Support vector machines (SVM)
Generalised additive model (GAM)	Artificial neural net (ANN)
Regularised regression (LASSO and Ridge)	
Rule/decision lists and sets	
Decision tree (DT)	
Supersparse linear integer model (SLIM)	
K-nearest neighbour (KNN)	
Naïve Bayes	
Case-based reasoning (CBR)/ Prototype and criticism	

## Further Reading

 [Further reading on algorithm types](#)  
For organisations

## 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 to

convey your model's results reliably, clearly, and in a way that enables 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 as a whole 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 intelligible the way its components and relationships function. 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 insights 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 you can draw internal explanations 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. It draws 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. 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 building interpretable

AI models such as those listed above. These techniques not only make the rationale behind AI models readily understandable; 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 it may be misleading to overly rely on supplementary tools.

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.

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 to setting up an accurate local explanation. It will also be essential to securing your AI system's fairness, safety and optimal performance. This will help you provide decision recipients with the fairness explanation and safety and performance explanation.

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 you can properly address concerns raised by such global insights.

In Annexe 3 we provide you with a table containing details of some of the more widely used supplementary explanation strategies and tools, and we highlight some of their strengths and weaknesses. 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 need to move beyond the basic information we are offering there. In Annexe 3 we cover the following supplementary explanation strategies:

<hr/>	
Local supplementary explanation strategies	Global supplementary explanation strategies
Individual Conditional Expectations Plot (ICE)	Partial Dependence Plot (PDP)
<hr/>	

Sensitivity Analysis and Layer-Wise Relevance Propagation (LRP)	Accumulated Local Effects Plot (ALE)
Local Interpretable Model-Agnostic Explanation (LIME) and anchors	Global Variable Importance
Shapley Additive ExPlanations (SHAP)	Global Variable Interaction
Counterfactual Explanation	
Surrogate models (SM) (Could also be used for global explanation)	
Self-Explaining and Attention-Based Systems (Could also be used for global explanation)	

## Further Reading

 [Further reading on supplementary techniques](#)  
For organisations

### Combining and integrating supplementary explanation strategies

The main purpose of using supplementary explanation tools is to make the underlying rationale of the results 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 together. 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 you could combine these different strategies into a portfolio of tools for explanation extraction.

Keeping in mind the various strategies we have introduced in the table in Annexe 3, there are three significant layers of technical rationale to include in your portfolio:

- visualise how the model works;
- understand the role of variables and variable interactions; and
- understand how the behaviours or circumstances that influence an AI-assisted decision would need to be changed to change that decision.

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

#### • Visualise 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 of 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?
- **Understand 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?
  - **Understand 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?

# Task 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 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.

## In more detail

- [Introduction to translating the rationale of your system's results](#)
- [Understand the statistical rationale](#)
- [Sense-check correlations and identify legitimate determining factors in a case-by-case manner](#)
- [Integrate your chosen correlations and outcome determinants into your reasoning](#)

## 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 your 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. This is outlined in more detail in Task 5.

## 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 phrase 'correlation does not imply causation'.

You need to take further steps to assess the role that these statistical associations should play in a reasonable explanation, given the particulars of the case you are considering.

## 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 this case. The challenge 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.

Taking this context sensitive approach should help you do two things:

- Sense-checking 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 AI-generated recommendation were weighed as part of a decision-support process or if an automated outcome based on this result were challenged, the model's implementer or reviewer 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. Taking a context sensitive approach should help you pick out the features and interactions that could reasonably make a real-world difference to the outcome. This is because 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, your 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. You could do this through a plain-language explanation, or any other format they may need to be able to make sense of the decision.

# Task 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 using these systems to assist decision-making, particularly how they help humans come to judgements rather than replacing that judgement; and
  - how to manage cognitive biases, including both decision-automation bias and automation-distrust bias.

## 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.
- ☐ Recognise situations where decision-automation bias and automation-distrust bias can occur and mitigate against this.
- ☐ Understand the strengths and limitations of the system.

## In more detail

- [Introduction to preparing implementers to deploy your AI system](#)



- [Basics of implementer training](#)

## 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 you have appropriately trained and prepared them to use your model's results responsibly and fairly.

Your 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. Your training should avoid any anthropomorphic (or human-like) portrayals of AI systems. You 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, your 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 decision-automation bias), or where under-reliance or under-compliance with the results can occur (automation-distrust 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. These can also include situations where the implementer may disregard the outcome of the system due to scepticism or distrust of the technology.

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

## Basics of implementer training

### Educate implementers about cognitive biases

Good implementer preparation begins with anticipating the pitfalls of potential bias-in-use which AI decision support systems tend to give rise to.

Your training about responsible implementation should therefore start with educating users about the two main types of AI-related bias: decision-automation bias and automation-distrust bias.

- **Decision-automation bias:** Users of AI decision-support systems may become hampered in their critical judgment and situational awareness as a result of an overconfidence in the objectivity, or certainty of the AI system.

This may lead to an over-reliance on the automated systems results. Implementers may lose the capacity to identify and respond to the faults, errors, or deficiencies, because they become complacent and defer to its directions and cues.

Decision-automation bias may also lead to a tendency to over-comply with the system's results. Implementers may defer to the perceived infallibility of the system and become unable to detect problems emerging from its use because they fail to hold the results against available information. This

may be exacerbated by underlying fears or concerns about how 'disagreeing with' or 'going against' a system's results might create accountability or liability issues in wider organisational or legal contexts.

Both over-reliance and over-compliance may lead to what is known as 'out-of-loop syndrome'. This is where the degradation of the role of human reason and the deskilling of critical thinking hampers the user's ability to complete the tasks that have been automated. This may reduce their ability to respond to system failure and may lead both to safety hazards and dangers of discriminatory harm.

- **Automation-Distrust Bias:** At the other extreme, users of an automated decision-support system may tend to disregard its contributions to evidence-based reasoning as a result of their distrust or scepticism about AI technologies in general. They may also over-prioritise the importance of prudence, common sense, and human expertise, failing to see how AI-decision support may help them to reduce implicit cognitive biases and understand complex patterns in data otherwise unavailable to human-scale reasoning.

If users have an aversion to the non-human and amoral character of automated systems, this could also lead decision subjects to mistrust these technologies in high impact contexts such as healthcare, transportation, and law.

To combat risks of decision-automation bias and automation-distrust bias, there are certain actions you should take:

- Build a comprehensive training and preparation program for all implementers and users that explores **both** AI-related judgment biases (decision-automation and automation-distrust biases) and human-based cognitive biases.
- Educate on this spectrum of biases including examples of particular misjudgements that may occur when people weigh statistical evidence. Examples of the latter may include:
  - overconfidence in prediction based on the historical consistency of data;
  - illusions that any clustering of data points necessarily indicates significant insights; and
  - discounting of societal patterns that exist beyond the statistical results.
- Make explicit and operationalise strong regimes of accountability when systems are deployed, in order to steer human decision-makers to act on the basis of good reasons, solid inferences, and critical judgment, even as they are supported by AI-generated results.

## **Educate implementers about the strengths and limitations of AI decision-support systems**

Your training about responsible deployment should also include a balanced and comprehensive technical view of the possible limitations and advantages of AI decision-support systems.

This means, first and foremost, giving users a working knowledge of the statistical and probabilistic methods behind the operation of these systems. Continuing education and professional development in this area is crucial to ensure that people using and implementing these systems have sufficient understanding. It is also crucial for providing users with a realistic and demystified picture of what these computation-based models are and what they can and cannot do.

A central component of this training should be to identify the limitations of statistical and probabilistic generalisation. Your training materials and trainers should stress the aspect of uncertainty that underlies all statistical and probabilistic reasoning. This will help users and implementers to approach AI-generated

results with an appropriately critical eye and a clear understanding of indicators of uncertainty like confidence intervals and error bars.

Your training should also stress the variety of performance and error metrics available to measure statistical results and the ways that these metrics may sometimes conflict or be at cross-purposes with each other, depending on the metrics chosen.

When educating users about the advantages of these AI systems, your training should involve example-based demonstrations of the capacities of applied data-science. This will show how useful and informative patterns and inferences can be drawn from large amounts of data, that may have otherwise escaped human insight, given time pressures as well as sensory and cognitive limitations. Educating users on the advantages of AI systems should also involve example-based demonstrations of how responsible and bias-aware model design can support and improve the objectivity of human decision-making through equitable information processing.

### **Train implementers to use statistical results as support for evidence-based reasoning**

Another crucial part of your training about responsible implementation is preparing users to be able to see the results of AI decision-support as assisting evidence-based reasoning rather than replacing it. As a general rule, we use the results of statistical and probabilistic analysis to help guide our actions. When done properly, this kind of analysis offers a solid basis of empirically derived evidence that assists us in exercising sound and well-supported judgment about the matters it informs.

Having a good understanding of the factors that produce the result of a particular AI decision-support system means that we are able to see how these factors (for instance, input features that weigh heavily in determining a given algorithmically generated output) mean that the result is rationally acceptable.

You should train your implementers to understand how the output of a particular AI system can support their reasoning. You should train them to grasp how they can optimally draw on the determining factors that lie behind the logic of this output to exercise sound judgment about the instance under consideration. Your training should emphasise the critical function played by rational justification in meeting the reasonable expectations of decision recipients who desire, or require, explanations. Carrying out this function demands that users and implementers offer well-founded arguments justifying the outcome of concern. Arguments that make sense, are expressed in clear and understandable terms, and are accessible enough to be rationally assessed by all affected individuals, especially the most vulnerable or disadvantaged.

### **Train implementers to think contextually and holistically**

The results of AI decision-support systems are based on population-level correlations that are derived from training data and that therefore do not refer specifically to the actual circumstances, background, and abilities of the individual decision recipient. They are statistical generalisations that have picked up relationships between the decision recipient's input data and patterns or trends that the AI model has extracted from the underlying distribution of that model's original dataset. For this reason, you should train your implementers to think contextually and holistically about how these statistical generalisations apply to the specific situation of the decision recipient.

This training should involve preparing implementers to work with an active awareness of the socio-technical aspect of implementing AI decision-assistance technologies from an integrative and human-centred point of

view. You should train implementers to apply the statistical results to each particular case with appropriate context-sensitivity and ‘big picture’ sensibility. This means that the dignity they show to decision subjects can be supported by interpretive understanding, reasonableness, and empathy.

Example-based training materials should illustrate how applying contextually-sensitive judgment can help implementers weigh the AI system’s results against the unique circumstances of the decision recipient’s life situation. In this way, your implementers can integrate the translation the system’s rationale into useable and easily understandable reasons (Task 4) into more holistic considerations about how those reasons actually apply to a particular decision subject. Training implementers to integrate Task 4 into context-sensitive reasoning will enable them to treat the inferences drawn from the results of the model’s computation as evidence that supports a broader, more rounded, and coherent understanding of the individual situations of the decision subject and other affected individuals.

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

## At a glance

- To build an explanation, you should start by gathering together the information gained when implementing Tasks 1-4. You should review the information, and determine how this provides an evidence base for process-based or outcome-based explanations.
- You should then revisit the contextual factors to establish which explanation types you should prioritise.
- How you present your explanation depends on the way 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, rather than 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 gathered the information collected in Tasks 1-4 and reviewed how these fit within the process-based and outcome-based explanations introduced in Part 1.
- ☐ We have considered the contextual factors and how this will impact the order in which we deliver the explanation types, and how this will affect our delivery method.
- ☐ 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 to considering how to build and present your explanation](#)
- [Gather relevant information for each explanation type](#)
- [Consider contextual factors in delivering an explanation](#)
- [Layer explanations](#)
- [Explanation as a dialogue](#)
- [Explanation timing](#)
- [Proactive engagement](#)

## Introduction

Before you are able to provide an explanation to an individual, you need to consider how to build and present the information in a clear, accessible manner.

You should start by considering the information you obtained when completing Tasks 1-4, and determine how much of this information is required by the decision recipient. You should consider both process-based and outcome-based explanations as part of this step. You should also consider what explanation types you should prioritise. You could revisit the contextual factors introduced in Part 1 of this guidance to help you with this.

You should determine the most appropriate method of delivery based on the way 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 with 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.

## Gather relevant information for each explanation type

When completing Tasks 1-4 above, you should have documented the steps you took and any other information you require to deliver an explanation to a decision recipient. You can then use this information to build your explanation ready for delivery.

Under Task 1 you should have identified the most relevant explanation types for your use case, taking account of the contextual factors. Part 1 of this guidance sets out the kinds of information that you need to extract to support each explanation type, including information about the process (eg the data used to train the model) and the outcome (eg how other people in a similar position were treated in comparison to the decision recipient).

Task 2 discusses how to take account of explanation types in collecting and pre-processing the data.

Task 3 sets out key issues related to how to extract the information needed for the relevant explanation types (especially rationale explanation) from your AI model.

Task 4 focusses on translating the rationale of the AI system into plain-language, context-sensitive, and understandable terms but this can also yield information to support other explanation types.

Having followed these tasks you should then be ready to consider how to present your explanation to the decision recipient.

## Consider contextual factors in delivering an explanation

When building an explanation, you should revisit the contextual factors introduced in [Part 1 of this guidance](#):

- Domain factor
- Impact factor
- Data factor
- Urgency factor
- Audience factor

Although these are relevant throughout the design, development and deployment of the system, you should consider them in detail when you are deciding how to build and present your explanation.

The domain factor is important because the domain, or sector that you are operating in will affect the type of explanation your decision recipients want to receive. There also may be legislation specific to your sector that dictates how you deliver an explanation.

The impact factor is important, because the impact a decision has on an individual or society will determine the level of information required in an explanation. For example, if the decision has a significant impact on an individual, you may need to provide a more detailed explanation than if the impact was low.

The data factor helps you and the decision recipient understand both how the model has been trained and what data has been used to make a decision. The type of data that was processed may affect how you deliver the explanation, as there may be some circumstances where the explanation provided gives the decision recipient information on how to affect the outcome in the future.

The urgency factor helps you determine how quickly to provide an explanation, and in what order you should be providing the different explanation types.

The audience factor is about the level of understanding you would expect the decision recipient to have about what the decision is about. This will affect the type of language you use when delivering your explanation.

## Layer 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 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. Instead, 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 it as a conversation as opposed to a one-way process. By providing the priority explanations, you are then 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 we provide in advance?**

In Part 1 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 will 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 we 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 we 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. For example, 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.

Further Reading

Part 1 The basics of explaining AI  
For organisations

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 (in this case, this may include measures taken to mitigate unbalanced, unrepresentative datasets or possible selection biases).  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 about the rationale of the AI system, and the outcome-based explanation about the AI system's impact on the individual, fulfil this requirement of the GDPR.

It is up to you 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.

#### Further reading

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 should outline 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 cover 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 you 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 do we use AI for decisions?

This should clearly and comprehensively explain 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 individuals 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 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.

### **Further reading**

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 \(GDPR\)](#)

### **Example**

In [Annexe 1](#) we provide an example showing how all of the above tasks could relate to a particular case in the health sector.

# Part 3: What explaining AI means for your organisation

## About this guidance

### What is the purpose of this guidance?

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

### How should we use this guidance?

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

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

### What is the status of this guidance?

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

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

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

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

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



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

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



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

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

# Organisational roles and functions for explaining AI

## At a glance

- Anyone involved in the decision-making pipeline has a role to play in contributing to an explanation of a decision supported by an AI model's result.
- This includes what we have called the AI development team, as well as those responsible for how decision-making is governed in your organisation.
- We recognise that every organisation has different structures for their AI development and governance teams, and in smaller organisations several of the functions we outline will be covered by one person.
- Many organisations will outsource the development of their AI system. In this case, you as the data controller have the primary responsibility for ensuring that the AI system you use is capable of producing an explanation for the decision recipient.

## Checklist

- ☐ We have identified the people who are in key roles across the decision-making pipeline and how they are responsible for contributing to an explanation of the AI system.
- ☐ We have ensured that different people along the decision-making pipeline are able to carry out their role in producing and delivering explanations, particularly those in AI development teams, those giving explanations to decision recipients, and our DPO and compliance teams.
- ☐ If we are buying the AI system from a third party, we know we have the primary responsibility for ensuring that the AI system is capable of producing explanations.

## In more detail

- [Who should participate in explanation extraction and delivery?](#)
- [What if we use an AI system supplied by a third party?](#)

### Who should participate in explanation extraction and delivery?

People involved in every part of the decision-making pipeline, including the AI model's design and implementation processes, have a role to play in providing explanations to individuals who receive a decision supported by an AI model's result.

In this section, we will describe the various roles in the end-to-end process of providing an explanation. In some cases, part of this process may not sit within your organisation, for example if you have procured the system from an external vendor. More information on this process is provided later in Part 3.



The roles discussed range from those involved in the initial decision to use an AI system to solve a problem and the teams building the system, to those using the output of the system to inform the final decision and those who govern how decision-making is done in your organisation. Depending on your organisation, the roles outlined below might be configured in different ways, or concentrated in just one or two people.

Please note that this is not an exhaustive list of all individuals that may be involved in contributing to an explanation for a decision made by an AI system. There may be other roles unique to your organisation or sector that are not outlined here. The roles listed below are the main ones we feel every organisation should consider when implementing an AI system to make decisions about individuals.

## Overview of the roles involved in providing an explanation

**Product manager:** defines the product requirements for the AI system and determines how it should be managed, including the explanation requirements and potential impacts of the system's use on affected individuals. The product manager is also responsible throughout the AI system's lifecycle. They are responsible for ensuring it is properly maintained, and that improvements are made where relevant. They also need to ensure that the system is procured and retired in compliance with all relevant legislation, including GDPR and DPA 2018.

**AI development team:** The AI development team performs several functions, including:

- collecting, procuring and analysing the data that you input into your AI system, which must be representative, reliable, relevant, and up-to-date;
- bringing in domain expertise to ensure the AI system is capable of delivering the types of explanations required. Domain experts could, for example, be doctors, lawyers, economists or engineers;
- building and maintaining the data architecture and infrastructure that ensure the system performs as intended and that explanations can be extracted;
- building, training and optimising the models you deploy in your AI system, prioritising interpretable methods;
- testing the model, deploying it, and extracting explanations from it; and
- supporting implementers in deploying the AI system in practice.

Please note that the AI development team may sit within your organisation, or be part of another organisation if you purchased the system from a third party. If you procure a system from a third party, you still need to ensure that you understand how the system works and how you can extract the meaningful information necessary to provide an appropriate explanation.

**Implementer:** where there is a human in the loop (ie the decision is not fully automated) the implementer relies on the model developed to supplement or complete a task in their everyday work life. In order to extract an explanation, implementers either directly use the model, if it is inherently interpretable and simple, or use supplementary tools and methods that enable explanation, if it is not. The tools and methods provide implementers with information that represents components of the rationale behind the model's results, such as relative feature importance. Implementers take this information and consider it together with other evidence to make a decision on how to proceed.

Where a system is developed by a third party vendor, you should ensure that they provide sufficient training and support so that your implementers are able to understand the model you are using. If you do not have this support in place, your implementers may not have the skills and knowledge to deploy the

system responsibly and to provide accurate and context sensitive explanations to your decision recipients.

**Compliance teams, including DPO:** these ensure that the development and use of the AI system comply with regulations and the your own policies and governance procedures. This includes compliance with data protection law, such as the expectation that AI-assisted decisions are explained to individuals affected by those decisions.

**Senior management:** this is the team with overall responsibility for ensuring the AI system that is developed and used within your organisation, or you procure from a third party, is appropriately explainable to the decision recipient.

We suggest that both the compliance teams, including DPO and senior management should expect assurances from the product manager that the system you are using provides the appropriate level of explanation to decision recipients. These assurances should give these roles a high level understanding of the systems and types of explanations they should and do produce. Additionally, there may be occasions when your DPO and/ or compliance teams interact directly with decision recipients. For example, if a complaint has been made. In these cases, they will need a more detailed understanding of how a decision has been reached, and they will need to be trained on how to convey this information appropriately to affected individuals.

Your AI system may also be subject to external audit, for example by the Information Commissioner's Office (ICO) to assess whether your organisation is complying with data protection law. Data protection includes the expectation that decisions made with AI are explained to individuals affected by those decisions. During an audit you will need to produce all the documentation you've prepared and the testing you've undertaken to ensure that the AI system is able to provide the different types of explanation required.

As the focus of this guidance is on providing explanations to decision recipients, we will not go into detail about this here. However, if you would like more information on the documentation you are required to provide if you are subject to a GDPR audit, please read our Auditing Framework.

As you move along the decision-making pipeline, through the roles identified above, there will be a certain amount of translation and exposition required. That is, you will need to translate the reasoning behind the statistical outputs of the AI system for different audiences. Likewise, you will have to organise the documented innovation processes, which have ensured that your system is accountable, safe, ethical, and fair and make them accessible to these different audiences. Internally to your organisation this means to the implementer, DPO and compliance team, and to senior management. Externally this means translating what you have performed from a technical point of view into language and reasoning that is clear and easily understandable to the decision recipient and the external auditor.

**Other roles:** the roles we've identified are generic ones. If they don't fit in your particular case or you have others, then you should consider how your roles relate to the decision making pipeline, and therefore to the task of providing an explanation.

## Further Reading

 [AI Auditing Framework](#) 

About the ICO

## What if we use an AI system supplied by a third party?

If you are sourcing your AI system, or significant parts of it, from a third party supplier, the functions and responsibilities may look different. Whether you do this or build it yourself, you as the data controller have the primary responsibility for ensuring that the AI system you use is capable of producing an appropriate explanation for the decision recipient. If you procure the system from a third party supplier that is off the shelf and does not contain inherent explainability, you may need another model alongside it.

### Further reading

More information on supplementary models and techniques is in '[Explaining AI in Practice](#)'.

# Policies and procedures

## At a glance

- Whether you create new policies and procedures or update existing ones, they should cover all the 'explainability' considerations and actions that you require from your employees from concept to deployment of AI decision-support systems.
- Your policies should set out what the rules are, why they are in place, and who they apply to.
- Your procedures should then provide directions on how to implement the rules set out in the policies.

## Checklist

- ☐ Our policies and procedures cover all the explainability considerations and actions we require from our employees from concept to deployment of AI systems.
- ☐ Our policies make clear what the rules are around explaining AI-assisted decisions to individuals, why those rules are in place, and who they apply to.
- ☐ Our procedures give directions on how to implement the rules set out in the policies.

## In more detail

- [Why do we need policies and procedures for explaining AI?](#)
- [What should our policies and procedures cover?](#)

### Why do we need policies and procedures for explaining AI?

Your policies and procedures are important for several reasons, they:

- help ensure consistency and standardisation;
- clearly set out rules and responsibilities; and
- support the creation/ adoption of your organisational culture.

These are all highly desirable for your approach to explaining AI-assisted decisions to individuals.

You may want to create new policies and procedures, or it might make more sense to adapt and extend those that already exist, such as data protection and information management policies, or broader information governance and accountability frameworks.

How you choose to do this depends on the unique set up of your organisation. What matters is that there is a clear and explicit focus on explaining AI-assisted decisions to individuals, why this is necessary and how it is done. This will help to embed explainability as a core requirement of your use of AI. It may be that you

have several current policies you can add AI into. If this is the case, you should document your processes as accurately as possible, and cross reference these policies where necessary.

## What should our policies and procedures cover?

Both your policies and procedures should cover all the explainability considerations and actions that you require from your employees from concept to deployment of AI decision-support systems. In short, they should codify what's in the different parts of this guidance for your organisation.

Your policies should set out the what, why and who of explaining AI-assisted decisions to individuals, ie they should make clear what the rules are, why they are in place, and who they apply to. Your procedures should set out how you explain AI-assisted decisions to individuals, ie directions on how to implement the rules set out in the policies.

The table below summarises the key areas to cover in your policies and indicates where it is beneficial to have an accompanying procedure.

This is only a guide. Depending on what you already have in place, you may find it is more important to provide more (or less) detail in certain areas than others. There may also be additional aspects, not covered in this table, that you wish to cover in your policies and procedures.

As such, if there are policies and procedures listed here that you feel do not apply to your organisation, sector, or the type of AI system you have implemented, you may not need to include them. The level of detail is likely to be proportionate to the level of risk. The more impactful and the less expected the processing is, the more detail you are likely to need in your policies, procedures and documentation.

If you procure a system from another organisation, and do not develop it in-house, there may be some policies and procedures that are required by the vendor and not you. However, as you are still the data controller for the decisions made by the system, it is your responsibility to ensure the vendor has taken the necessary steps outlined in the table.

You should consult relevant staff when drafting your policies and procedures to ensure that they make sense and will work in practice.

	Policy	Procedure
<b>Policy objective</b>	Explain what the policy seeks to achieve – the provision of appropriate explanations of AI-assisted decisions to individuals. Even if you incorporate your requirements around explaining AI-assisted decisions into an existing policy or framework, you should ensure that this particular objective is explicitly stated.	N/A
<b>Policy rationale</b>	Outline why the policy is necessary. You should cover the broad legal requirements, and any legal	N/A

requirements specific to your organisation or sector. You should also cover the benefits for your organisation, and, where relevant, link the rationale to your organisation's broader values and goals.

### **Policy scope**

Set out what the policy covers. Start by clarifying what types of decision-making and AI systems are in scope. Say which departments or parts of your organisation the policy applies to. Where necessary, explain how this policy links to other relevant policies your organisation has, signposting other organisational requirements around the use of AI systems that are not within this policy's scope.

N/A

### **Policy ownership**

Make clear who (or which role) within your organisation has ownership of this policy, and overarching responsibility for the 'explainability' of AI decision-support systems. Explain that they will monitor and enforce the policy.

Detail the steps that the policy owner should take to monitor and enforce its use. Set out what checks to make, how often to make them, and how to record and sign-off this work.

### **Roles**

Set out the specific roles in your organisation that have a stake or influence in the explainability of your AI decision-systems. Describe the responsibilities of each role (in connection with explaining AI-assisted decisions to individuals) and detail the required interaction between the different roles (and departments) and the appropriate reporting lines, all the way up to the policy owner.

N/A

### **Impact assessment**

Explain the requirement for explainability to be embedded within your organisation's impact assessment methodology. This is likely to be a legally required assessment such as a Data Protection Impact Assessment, but it could also form part of broader risk or ethical assessments.

Explicitly state the need to conduct the assessment (including considering explainability) before work begins on an AI decision-support system. Describe how to make the necessary explainability assessments including consideration of the most relevant explanation types, and the most suitable AI model for the context within which you will be making AI-assisted

		decisions about individuals.
<b>Awareness raising</b>	Explain the importance of raising awareness about your use of AI-assisted decisions with your customers or service users. Set out the information you should communicate including why you use AI for decision-making, where you do this, and simple detail on how it works.	Detail the specific approach your organisation takes to raising awareness. Clarify where you host information for your customers, how you will communicate it or make it available (eg which channels and how often), and which roles or departments are responsible for this.
<b>Data collection</b>	Underline the need to consider explanations from the earliest stages of your AI model development, including the data collection, procurement and preparation phases. Explain why this is important with reference to the benefits of interpretable and well-labelled training data.	Set out the steps to take at the data collection stage to enhance the explainability of your AI model, including assessing data quality, structure, feature interpretability, and your approach to labelling.
<b>Model selection</b>	Explain how considerations of explainability factored in to the selection of your AI model in its development stage and how the algorithmic techniques you chose to use are appropriate for both the system's use case and its potential impacts.	Set out the steps you took to weigh model types against the priority of the interpretability of your AI system. Signpost how you ensured that the selected model is appropriate to fulfil that priority.
<b>Explanation extraction</b>	Set out the different types of explanation and outline the requirements to obtain information relevant to each one.	Signpost the various technical procedures your organisation uses to extract the rationale explanation from your AI models (eg how to use a local explanation tool such as LIME) or how you use visualisation or counterfactual methods. Describe how to obtain information on the other explanations, who to obtain this from, and how to record it.
<b>Explanation delivery</b>	Explain the need to build and deliver explanations in the way that is most meaningful for the individuals your AI-assisted decisions are about.	Detail how to prioritise the explanation types, how to translate technical terminology into plain language, the format in which to present the explanation (eg in layers), and how to assess appropriate timing of delivery (eg before or after a decision).
<b>Documentation</b>	Clearly state the necessity to document the justifications and	Set out the standardised method by which all stakeholders can record

choices made through the whole process of developing/ acquiring and deploying an AI decision-support system. Outline the requirement to document the provision of explanations to individuals, and clarify what information to record (eg URN, time stamp, explanation URL).

their justifications and choices. Explain how your organisation keeps an audit trail of explanations provided to individuals, including how this can be accessed and checked.

### **Training**

Set requirements for general staff training on explaining AI decisions to individuals, covering why it's necessary, and how it's done. Identify roles that require more in-depth training on specific aspects of explaining AI-assisted decisions, such as preparing training data or extracting rationale explanations.

N/A



# Documentation

## At a glance

- It is essential to document each stage of the process behind the design and deployment of an AI decision-support system in order to provide a full explanation for how you made a decision.
- In the case of explaining AI-assisted decisions, this includes both documenting the processes behind the design and implementation of the AI system and documenting the actual explanation of its outcome.
- The suggested areas for documentation may not apply to all organisations, but are intended to give you an indication of what might help you provide the evidence to establish how a decision was made.
- The key objective is to provide good documentation that can be understood by people with varying levels of technical knowledge and that covers the whole process from designing your AI system to the decision you make at the end.

## Checklist

- ☐ We have documented what we are required to do under the GDPR.
- ☐ We have documented how each stage of our use of AI contributes to building an explanation, from concept to deployment.
- ☐ Our documentation provides an audit trail about who we give explanations to, and how we provide them.
- ☐ We have considered how best to organise our documentation so that relevant information can be easily accessed and understood by those providing explanations to decision recipients.

## In more detail

- [What documentation is legally required under the GDPR?](#)
- [What documentation should we provide to demonstrate the explainability of our AI system?](#)
- [How should we organise this documentation?](#)

### What documentation is legally required under the GDPR?

**Article 5 of the GDPR** says that “The controller shall be responsible for, and able to demonstrate compliance with, paragraph 1 (‘accountability’).”

**Article 12 of the GDPR** requires you to provide information to the data subject in “concise, transparent, intelligible and easily accessible form, using clear and plain language...”. It also states that you can provide the information “in combination with standardised icons in order to give in an easily visible, intelligible and

clearly legible manner a meaningful overview of the intended processing.”

**Article 13 of the GDPR** requires you to provide your DPO’s contact details, which aligns with the responsibility explanation; the purpose for which you are processing the data subject’s personal data, as well as the legal basis for that processing, which in many cases should form part of your explanation; and the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject. You must document all of this to ensure you remain accountable.

**Article 14 of the GDPR** applies to cases where you have not obtained personal data from the data subject directly. You should provide data subjects with the following information in addition to that required under Article 13, within a reasonable period after obtaining the personal data, but at the latest within one month, having regard to the specific circumstances in which the personal data are processed:

- what categories of personal data you are processing; and
- the source from which you obtained their personal data, and if applicable, whether it came from publicly accessible sources.

See Article 14 of the GDPR for further information on when it is not required to provide this information to the data subject. This includes when you are processing personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes. This is subject to certain conditions and safeguards.

**Article 15 of the GDPR** gives data subjects an additional right of access to the personal data that you hold on them. This means you should document how you will provide them with a copy of the personal data that you process.

**Article 21 of the GDPR** gives data subjects the right to object at any time, on grounds relating to their particular situation, to processing of personal data concerning them, including profiling. This means you should document how you ensure data subjects are aware of this right, and how you record if they have exercised this right.

**Article 22 of the GDPR** gives individuals the right not to be subject to a solely automated decision producing legal or similarly significant effects, unless certain conditions apply. It obliges you to adopt suitable measures to safeguard individuals, including the right to obtain human intervention, to express their view, and to contest the decision. This means you need to document how you will do this.

**Article 30 of the GDPR** helps you to fulfil the accountability principle. It states that an organisation shall “...maintain a record of processing activities under its responsibility.”

**Article 35 of the GDPR** requires organisations to carry out a Data Protection Impact Assessment (DPIA) when they are doing something with personal data, particularly when using new technologies, which is likely to have high risks for individuals. A DPIA is always required for any systematic and extensive profiling or other automated evaluation of individuals’ personal aspects which are used for decisions that produce legal or similarly significant effects.

You can reconcile some of these documentation requirements through your **privacy notice**, which must contain certain elements:

- The lawful basis for the processing – one or more of the bases laid out in Article 6(1) of the GDPR.
- If applicable, the legitimate interests for the processing – these are the interests pursued by you or a third party if you are relying on the lawful basis for processing under Article 6(1)(f) of the GDPR. You could also include a link to the record of your assessment of whether legitimate interests apply to the particular processing purpose.
- The rights available to individuals regarding the processing – eg access, rectification, erasure, restriction, data portability, and objection. The rights vary depending on the lawful basis for processing. Your documentation can reflect these differences.
- If applicable, the existence of automated decision-making, including profiling. In certain circumstances you will need to tell people about the logic involved and the envisaged consequences.
- If applicable, the source of the personal data. This is relevant when you didn't obtain personal data directly from an individual.

You should be aware that different documentation requirements may apply for law enforcement processing under Part 3 of the DPA 2018 and for intelligence services processing under Part 4 of the Act.

While this guidance focusses on documentation required to support explanations, the auditing framework covers other aspects of an AI system. In particular, the framework details those that require documentation for data protection compliance and good information governance.

## Further Reading



[Documentation](#)

For organisations



[Auditing framework](#)

External link

## What documentation can help us to demonstrate the explainability of our AI system?

The list below should support you to provide an explanation to the decision recipient, and maintain an audit trail about who you give explanations to, and how you provide them.

You may not need to provide or document all of this information, and you may have to obtain some information from your vendor if you procure a system. It is up to you to decide what information is required, and how the documented information can help you provide an explanation to the decision recipient. As with the policies and procedures section, a risk-based approach can help. For example, an AI system that recommends which groceries to buy or which films to watch will require less detail than those in recruitment.

## Decision to use an AI system

- What the system is intended to be used for, so you can explain to the decision recipient why you are planning to use it.
- Who the ultimate decision recipient will be.
- What the AI system you have chosen will do from a technical perspective, in a way that the decision

recipient can also understand.

- How the specifications of the system were determined, and by whom – as well as alternative specifications that were considered, and why they were not chosen.
- If your AI system has been procured from a third party or outsourced, how you can change or retool its specifications to meet changing performance and explainability needs over time.
- What trade-offs are involved for the data subject whose data will be used in the model and who will often also be the decision recipient. For example, the data subject's personal data may be used in training the AI system to produce a highly accurate model, but this use of data may not be in the data subject's interest.
- What the demographics and background of the development team are, in order to be aware of the diversity within the team responsible for designing, deploying and maintaining the system, and how that may impact on the decision for the decision recipient.
- What domain you will be using the AI system in, and how this system has been tested and validated in that domain.
- What other impact assessment are relevant to your domain, in addition to the DPIA mentioned.
- Which people within the organisation have responsibility for providing explanations along the design and implementation pipeline of your AI system.

Explanation types this supports: rationale, responsibility, fairness, safety and performance, impact.

## **Scoping and selecting explanation types**

- What processes you have set up to optimise the end-to-end accountability of your AI model.
- What setting or sector your AI model will be used in and the bearing this has on the types of explanation you will offer.
- Why you have prioritised certain explanation type(s), based on your AI system's potential impact.
- Why you have chosen to handle the remaining explanation types that will not be prioritised in a certain way.
- How you have set up process-based and outcome-based aspects of the explanations types that you will offer.
- Why you have provided the depth or comprehensiveness of the explanation, given the potential impacts of the system. This includes the general risks of deploying the system, and the risks for the specific person receiving the AI-assisted decision.
- Who within your organisation is responsible for selecting the appropriate type(s) of explanation.

Explanation types this supports: rationale, responsibility, data, fairness, safety and performance, impact.

## **Data collection and procurement**

- Where the data came from, for what purpose the data was originally collected, and for whom – this will help you explain to the decision recipient how relevant the data you have used is to the decision the AI system has made about them.
- What the components of the dataset are together with a brief summary of why each element is being included.
- How the data is representative of the people that will be subject to the AI decisions you will make – for

example through consideration by a domain expert.

- How you have made sure that the data is reliable, accurately measured and obtained from a source of integrity.
- How you have examined your datasets for any potential inherent bias.
- Whether the data is recent, up-to-date and appropriately timely given the rate of change in the underlying distribution you are modelling. This will demonstrate that you have accounted for concept drift and data fluctuation from the start. The rate of change will depend on the domain you are operating in, and the specific case you are considering.
- If you use synthetic data, provide documentation on when and how it was created, and its properties – this helps you explain and justify to the decision recipient why created data has been used in the training of the model and why this is appropriate.
- What the risks associated with using the data are, and the risks to those whose data is included.
- How individuals can opt out of being included in the data used either to train or run the AI system.

Explanation types this supports: data, fairness, safety and performance, impact.

## **Data pre-processing**

- How, especially in cases where social and demographic data is involved, you have ensured that the pre-processing of your data has produced a feature space which includes variables that are understandable, relevant and reasonable and that does not include variables that are opaque or difficult to understand about your model's target variable.
- How you have labelled the data and why you have labelled it in that way. This should include tagging and annotating what a piece of data is, and the reasons for that tag.
- How you have mitigated any bias in your data through pre-processing techniques such as re-weighting, up-weighting, masking, or excluding features and their proxies.
- If you are using 'raw', observed, or unconventional data, documentation of what interpretively significant feature such data is supposed to indicate about the individual whose data is being processed and evidence that this has been included in the metadata.
- Who within your organisation is responsible for data collection and pre-processing.

Explanation types this supports: responsibility, data, fairness, safety and performance, impact.

## **Model selection**

- What the specific interpretability or transparency standards, conventions and requirements of the domain in which your AI system will be applied are.
- How the specific type of application and the impact on individuals informs the type of model you choose.
- How the types of data you are using, for example social or demographic data, or biophysical data, have influenced your model selection regarding its interpretability.
- Whether your use case enables you to use maximally interpretable algorithmic techniques, and if not, why not.
- When using 'black box' models, what the risks are of using them, and how you will provide supporting evidence that your team has determined that your use case and your organisational capacities and resources support the responsible design and implementation of these systems.

- When using opaque algorithmic techniques such as ‘black boxes’, how the supplementary tools that you will use to explain the model provide a domain-appropriate level of explainability. Your documentation should demonstrate how the supplementary tool will mitigate the potential risks of using a ‘black box’ system, and how the use of the tool will help you to provide meaningful information about the rationale of any given outcome.
- If you use ‘challenger’ models alongside more interpretable models, what is the purpose of these models and how will you use them.

Explanation types this supports: rationale, responsibility, data, fairness, safety and performance, impact.

## **Model building, testing and monitoring**

- What the accuracy rate and other performance metrics you have chosen for the model are, as well as any tuning of cost ratios to constrain error allocation, and how and why you have selected these. You should be able to explain to the decision recipient how this choice may affect the decision that you have made about them.
- If relevant, what are the group-specific error rates and how the model has been tuned to redress any significant imbalances.
- How you have monitored and assessed potential for biases in the model design and what measures you have taken to mitigate those you have identified.
- How you have tested the model, including test results and which portions of the data you have used to train, test the model and holdout data.
- How frequently you will monitor, update and re-examine the model after it is deployed in the real world.
- How often you will update the training data after model production and deployment. You should also document what you have put in place to establish the appropriate frequency of updates.
- How you will track each time the model has been updated, and how each version has changed, so that you can explain to the decision recipient how that particular version of the model came to the decision, and why this might differ from the output of a subsequent or prior model.

Explanation types this supports: rationale, data, fairness, safety and performance.

## **Tools for extracting an explanation**

- When using more inherently interpretable models, what measures you have taken to ensure optimal explainability, for example the sparsity constraints placed on the feature space so that explanations can remain human understandable.
- When using supplementary interpretability tools for ‘black box’ models, an outline of what local and global techniques you have used to provide explanations. This may be in the form of detailed specifications of the supplementary tools you have used.
- How you plan to combine these different explanation tools to produce meaningful information about the rationale of the system’s results.
- Who is responsible for ensuring that the explanations generated by the supplementary tools are accessible to the people they are intended to inform.
- How you will translate the statistical output of your model and supplementary tools into a plain-language explanation, for example by establishing and documenting appropriate implementer training and providing users with comprehensive guidelines for responsible implementation.

Explanation types this supports: rationale, responsibility.

## Explanation delivery

- Why you will prioritise certain explanation types when you deliver the explanation to the affected individual, given the contextual factors you determine to be relevant in the particular case you are considering.
- How and why you have prioritised the remaining explanation types.
- What training you have provided to implementers to enable them to use the model's results responsibly and fairly.
- How the implementer will be presented with the model's result, including:
  - how you present performance metrics and error rates for the model as a whole and for sub-groups if appropriate;
  - how you present uncertainty measures like error bars and confidence intervals;
  - how you use visualisation tools and present indicators of relative variable important or variable interactions; and
  - in the case of 'black box' models, how you present information from supplementary tools as well as indicators of the limitations and uncertainty levels of these tools.
- What reasonable adjustments you will make for the form in which you deliver the explanation, as required under the Equality Act 2010.
- What information you will proactively share with your customers and stakeholders, so that they are able to make informed choices in advance of engaging with the decision-making process.
- Who decision recipients can contact to query a decision.

Explanation types this supports: rationale, responsibility, data, fairness, safety and performance, impact.

## How should we organise this documentation?

In each part of this guidance, we have emphasised that preparing your organisation to explain AI-assisted decisions is a holistic and end-to-end activity. It involves both demonstrating that you have undertaken the processes behind the design, development, and deployment of your AI system responsibly and clarifying the outcomes of that system's decision-support in a clear, understandable, and context-sensitive way. We have called these aspects of explaining AI process-based and outcome-based explanations.

Whether you are a developer, who is building and supplying AI applications, or an organisation developing your own AI systems in-house, one of the challenges you may face is figuring out how best to organise the documentation of your innovation practices to help with your process-based explanations.

This may, at first, seem like a daunting task, because it involves:

- documenting diverging governance activities across the AI design and deployment lifecycle;
- consolidating this information to easily convey it to a diverse range of stakeholders with varying needs and levels of technical and domain expertise; and
- differentially organising how to provide the information, so that different stakeholders receive the appropriate kinds and quantities of information (for example, ensuring that decision recipients are not overwhelmed by technical details and vast amounts of text or provided with commercially sensitive

information).


One method for organising the documentation for process-based explanations is building **argument-based assurance cases** for those high-level properties of your AI model (like safety or fairness). You can find details about how to do this in Annexe 5.

However you choose to organise your documentation, you should do it in a way that:

- allows you to easily access the relevant information required for each explanation type;
- is supported by your current document management system; and
- is accessible to those within your organisation that provide explanations to decision recipients.

If you plan to procure a system, you should ensure that the process you choose allows you to communicate with your vendor in a way that mutually manages expectations. If your vendor is able to offer evidence of their compliance and ability to explain decisions (through justification, evidence, and documentation), you will be able to better provide this information to your decision recipients. You will also be able to assess whether the model offered by the vendor meets the acceptable criteria and standards you have set for an AI system.

### Further reading

For further guidance on procuring systems, you may wish to read the World Economic Forum's [AI Government Procurement Guidelines](#) . Although this guidance is primarily aimed at public sector organisations, a large number of the principles contained within this guidance are not sector-specific. As such you should be able to apply them to your organisation.

## Further Reading



[ICO and The Turing consultation on Explaining AI decisions guidance - a summary of responses](#) 

About the ICO  
PDF (346.58K)



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

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

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

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

### Priority explanation types:

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

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

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

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

### Other explanation types:

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

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

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

### Information governance policy

Use of AI

Explaining AI decisions to patients

Types of explanations:

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

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

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

### **Rationale**

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

### **Responsibility**

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

### **Data**

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

### **Safety and performance**

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

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

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

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

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

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

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

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

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

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

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

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

Implementer training covers:

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

This includes any limitations in using the system.

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

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

### **Rationale**

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

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

### **Responsibility**

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

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

### **Safety and performance**

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

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

### **Impact**

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

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

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

## **Contextual factors**

Domain – regulated, safety testing...

Data – biophysical...

Urgency – if cancer, urgent...

Impact – high, safety-critical...

Audience – mostly non-expert...

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

### **Layer 1**

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

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

### **Layer 2**

- Data explanation
- Fairness explanation

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

## Annexe 2: Algorithmic techniques

Algorithm type	Basic description	Possible uses	Interpretability
<b>Linear regression (LR)</b>	Makes predictions about a target variable by summing weighted input/predictor variables.	Advantageous in highly regulated sectors like finance (eg credit scoring) and healthcare (predict disease risk given eg lifestyle and existing health conditions) because it's simpler to calculate and have oversight over.	High level of interpretability because of linearity and monotonicity. Can become less interpretable with increased number of features (ie high dimensionality).
<b>Logistic regression</b>	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.
<b>Regularised regression (LASSO and Ridge)</b>	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.
<b>Generalised linear model (GLM)</b>	To model relationships between features and target variables that do not follow normal (Gaussian) distributions a GLM introduces a link function that allows for the extension of LR to non-normal distributions.	This extension of LR is applicable to use cases where target variables have constraints that require the exponential family set of distributions (for instance, if a target variable involves number of people, units of time or probabilities of outcome, the result has to have a non-negative value).	Good level of interpretability that tracks the advantages of LR while also introducing more flexibility. Because of the link function, determining feature importance may be less straightforward than with the additive character of simple LR, a degree of transparency may be lost.

<b>Generalised additive model (GAM)</b>	To model non-linear relationships between features and target variables (not captured by LR), a GAM sums non-parametric functions of predictor variables (like splines or tree-based fitting) rather than simple weighted features.	This extension of LR is applicable to use cases where the relationship between predictor and response variables is not linear (ie where the input-output relationship changes at different rates at different times) but optimal interpretability is desired.	Good level of interpretability because, even in the presence of non-linear relationships, the GAM allows for clear graphical representation of the effects of predictor variables on response variables.
<b>Decision tree (DT)</b>	A model that uses inductive branching methods to split data into interrelated decision nodes which terminate in classifications or predictions. DT's moves from starting 'root' nodes to terminal 'leaf' nodes, following a logical decision path that is determined by Boolean-like 'if-then' operators that are weighted through training.	Because the step-by-step logic that produces DT outcomes is easily understandable to non-technical users (depending on number of nodes/ features), this method may be used in high-stakes and safety-critical decision-support situations that require transparency as well as many other use cases where volume of relevant features is reasonably low.	High level of interpretability if the DT is kept manageably small, so that the logic can be followed end-to-end. The advantage of DT's over LR is that the former can accommodate non-linearity and variable interaction while remaining interpretable.
<b>Rule/decision lists and sets</b>	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.
<b>Case-based reasoning (CBR)/</b>	Using exemplars drawn from prior human knowledge, CBR	CBR is applicable in any domain where experience-based	CBR is interpretable-by-design. It uses examples drawn from

<b>Prototype and criticism</b>	<p>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.</p>	<p>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.</p>	<p>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.</p>
<b>Supersparse linear integer model (SLIM)</b>	<p>SLIM utilises data-driven learning to generate a simple scoring system that only requires users to add, subtract, and multiply a few numbers in order to make a prediction. Because SLIM produces such a sparse and accessible model, it can be implemented quickly and efficiently by non-technical users, who need no special training to deploy the system.</p>	<p>SLIM has been used in medical applications that require quick and streamlined but optimally accurate clinical decision-making. A version called Risk-Calibrated SLIM (RiskSLIM) has been applied to the criminal justice sector to show that its sparse linear methods are as effective for recidivism prediction as some opaque models that are in use.</p>	<p>Because of its sparse and easily understandable character, SLIM offers optimal interpretability for human-centred decision-support. As a manually completed scoring system, it also ensures the active engagement of the interpreter-user, who implements it.</p>
<b>Naïve Bayes</b>	<p>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</p>	<p>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</p>	<p>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</p>



membership of that feature by multiplying the prior probability of the class with the class conditional probability of the feature.

feature spaces. Common applications include spam filtering, recommender systems, and sentiment analysis.

also a weakness, because feature interactions are not considered.

### **K-nearest neighbour (KNN)**

Used to group data into clusters for purposes of either classification or prediction, this technique identifies a neighbourhood of nearest neighbours around a data point of concern and either finds the mean outcome of them for prediction or the most common class among them for classification.

KNN is a simple, intuitive, versatile technique that has wide applications but works best with smaller datasets. Because it is non-parametric (makes no assumptions about the underlying data distribution), it is effective for non-linear data without losing interpretability. Common applications include recommender systems, image recognition, and customer rating and sorting.

KNN works off the assumption that classes or outcomes can be predicted by looking at the proximity of the data points upon which they depend to data points that yielded similar classes and outcomes. This intuition about the importance of nearness/proximity is the explanation of all KNN results. Such an explanation is more convincing when the feature space remains small, so that similarity between instances remains accessible.

### **Support vector machines (SVM)**

Uses a special type of mapping function to build a divider between two sets of features in a high dimensional feature space. An SVM therefore sorts two classes by maximising the margin of the decision boundary between them.

SVM's are extremely versatile for complex sorting tasks. They can be used to detect the presence of objects in images (face/no face; cat/no cat), to classify text types (sports article/arts article), and to identify genes of interest in bioinformatics.

Low level of interpretability that depends on the dimensionality of the feature space. In context-determined cases, the use of SVM's should be supplemented by secondary explanation tools.

### **Artificial neural net (ANN)**

Family of non-linear statistical techniques (including recurrent, convolutional, and deep neural nets) that build complex mapping functions to predict or classify data by employing the feedforward—and sometimes feedback—of input variables through

ANN's are best suited to complete a wide range of classification and prediction tasks for high dimensional feature spaces—ie cases where there are very large input vectors. Their uses may range from computer vision, image recognition, sales and weather forecasting,

The tendencies towards curviness (extreme non-linearity) and high-dimensionality of input variables produce very low-levels of interpretability in ANN's. They are considered to be the epitome of 'black box' techniques. Where appropriate, the use of ANN's should be

trained networks of interconnected and multi-layered operations.

pharmaceutical discovery, and stock prediction to machine translation, disease diagnosis, and fraud detection.

supplemented by secondary explanation tools.

### **Random Forest**

Builds a predictive model by combining and averaging the results from multiple (sometimes thousands) of decision trees that are trained on random subsets of shared features and training data.

Random forests are often used to effectively boost the performance of individual decisions trees, to improve their error rates, and to mitigate overfitting. They are very popular in high-dimensional problem areas like genomic medicine and have also been used extensively in computational linguistics, econometrics, and predictive risk modelling.

Very low levels of interpretability may result from the method of training these ensembles of decision trees on bagged data and randomised features, the number of trees in a given forest, and the possibility that individual trees may have hundreds or even thousands of nodes.

### **Ensemble methods**

As their name suggests, ensemble methods are a diverse class of meta-techniques that combines different 'learner' models (of the same or different type) into one bigger model (predictive or classificatory) in order to decrease the statistical bias, lessen the variance, or improve the performance of any one of the sub-models taken separately.

Ensemble methods have a wide range of applications that tracks the potential uses of their constituent learner models (these may include DT's, KNN's, Random Forests, Naïve Bayes, etc.).

The interpretability of Ensemble Methods varies depending upon what kinds of methods are used. For instance, the rationale of a model that uses bagging techniques, which average together multiple estimates from learners trained on random subsets of data, may be difficult to explain. Explanation needs of these kinds of techniques should be thought through on a case-by-case basis.

## Annexe 3: Supplementary models

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

associations with the output at different points. The PDP may flatten out these heterogeneities to the mean.

**Global/local?**  
**Internal/post-hoc?**

PDP's are global post-hoc explainers that can also allow deeper causal understandings of the behaviour of an opaque model through visualisation. These insights are, however, very partial and incomplete both because PDP's are unable to represent feature interactions and heterogenous effects, and because they are unable to graphically represent more than a couple of features at a time (human spatial thinking is limited to a few dimensions, so only two variables in 3D space are easily graspable).

**Individual**  
**Conditional**  
**Expectations Plot**  
**(ICE)**

Refining and extending PDP's, ICE plots graph the functional relationship between a single feature and the predicted response for an individual instance. Holding all features constant except the feature of interest, ICE plots represent how, for each observation, a given prediction changes as the values of that feature vary. Significantly, ICE plots therefore disaggregate or break down the averaging of partial feature effects generated in a PDP by showing changes in the feature-output relationship for each specific instance, ie observation-by-observation. This means that it can both detect interactions and account for uneven associations of predictor and response variables.

When used in combination with PDP's, ICE plots can provide local information about feature behaviour that enhances the coarser global explanations offered by PDP's. Most importantly, ICE plots are able to detect the interaction effects and heterogeneity in features that remain hidden from PDP's in virtue of the way they compute the partial dependence of outputs on features of interest by averaging out the effect of the other predictor variables. Still, although ICE plots can identify interactions, they are also liable to missing significant correlations between features and become misleading in some instances.

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).

<b>Global/local? Internal/post-hoc?</b>	ICE plots offer a local and post-hoc form of supplementary explanation.	
<b>Accumulated Local Effects Plots (ALE)</b>	<p>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.</p> <p>ALE plots are also more computationally tractable than PDP's because they are able to use techniques to compute effects in smaller intervals and chunks of observations.</p>	<p>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.</p> <p>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.</p>
<b>Global/local? Internal/post-hoc?</b>	ALE plots are a global and post-hoc form of supplementary explanation.	
<b>Global Variable Importance</b>	The global variable importance strategy calculates the contribution of each input feature to model output across the dataset by permuting the	While permuting variables to measure their relative importance, to some extent, accounts for interaction effects, there is still a

feature of interest and measuring changes in the prediction error: if changing the value of the permuted feature increases the model error, then that feature is considered to be important. Utilising global variable importance to understand the relative influence of features on the performance of the model can provide significant insight into the logic underlying the model's behaviour. This method also provides valuable understanding about non-linearities in the complex model that is being explained.

high degree of imprecision in the method with regard to which variables are interacting and how much these interactions are impacting the performance of the model.

A bigger picture limitation of global variable importance comes from what is known as the 'Rashomon effect'. This refers to the variety of different models that may fit the same data distribution equally well. These models may have very different sets of significant features. Because the permutation-based technique can only provide explanatory insight with regard to a single model's performance, it is unable to address this wider problem of the variety of effective explanation schemes.

**Global/local?  
Internal/post-hoc?**

Global variable importance is a form of global and post-hoc explanation.

**Global Variable  
Interaction**

The global variable interaction strategy computes the importance of variable interactions across the dataset by measuring the variance in the model's prediction when potentially interacting variables are assumed to be independent. This is primarily done by calculating an 'H-statistic' where a no-interaction partial dependence function is subtracted from an observed partial dependence function in order to compute the variance in the prediction. This is a versatile explanation strategy, which has been employed to calculate interaction effects in many types of complex models including ANN's and Random Forests. It can be used to calculate

While the basic capacity to identify interaction effects in complex models is a positive contribution of global variable interaction as a supplementary explanatory strategy, there are a couple of potential drawbacks to which you may want to pay attention.

First, there is no established metric in this method to determine the quantitative threshold across which measured interactions become significant. The relative significance of interactions is useful information as such, but there is no way to know at which point interactions are strong enough to exercise effects.



interactions between two or more variables and also between variables and the response function as a whole. It has been effectively used, for example, in biological research to identify interaction effects among genes.

Second, the computational burden of this explanation strategy is very high, because interaction effects are being calculated combinatorially across all the data points. This means that as the number of data points increase, the number of necessary computations increase exponentially.

**Global/local?  
Internal/post-hoc?**

Global variable interaction is a form of global and post-hoc explanation.

**Sensitivity Analysis  
and Layer-Wise  
Relevance  
Propagation (LRP)**

Sensitivity analysis and LRP are supplementary explanation tools used for artificial neural networks. Sensitivity analysis identifies the most relevant features of an input vector by calculating local gradients to determine how a data point has to be moved to change the output label. Here, an output's sensitivity to such changes in input values identifies the most relevant features. LRP is another method to identify feature relevance that is downstream from sensitivity analysis. It uses a strategy of moving backward through the layers of a neural net graph to map patterns of high activation in the nodes and ultimately generates interpretable groupings of salient input variables that can be visually represented in a heat or pixel attribution map.

Both sensitivity analysis and LRP identify important variables in the vastly large feature spaces of neural nets. These explanatory techniques find visually informative patterns by mathematically piecing together the values of individual nodes in the network. As a consequence of this piecemeal approach, they offer very little by way of an account of the reasoning or logic behind the results of an ANNs' data processing.

Recently, more and more research has focused on attention-based methods of identifying the higher-order representations that are guiding the mapping functions of these kinds of models as well as on interpretable CBR methods that are integrated into ANN architectures and that analyse images by identifying prototypical parts and combining them into a representational wholes. These newer techniques are showing that some significant progress is being made in uncovering the underlying logic of some ANN's.

**Global/local?**  
**Internal/post-hoc?**

Sensitivity analysis and salience mapping are forms of local and post-hoc explanation, although the recent incorporation of CBR techniques is moving neural net explanations toward a more internal basis of interpretation.

**Local Interpretable  
Model-Agnostic  
Explanation (LIME)  
and anchors**

LIME works by fitting an interpretable model to a specific prediction or classification produced by an opaque system. It does this by sampling data points at random around the target prediction or classification and then using them to build a local approximation of the decision boundary that can account for the features which figure prominently in the specific prediction or classification under scrutiny.

LIME does this by generating a simple linear regression model by weighting the values of the data points, which were produced by randomly perturbing the opaque model, according to their proximity to the original prediction or classification. The closest of these values to the instance being explained are weighted the heaviest, so that the supplemental model can produce an explanation of feature importance that is locally faithful to that instance. Note that other interpretive models like decision trees may be used as well.

While LIME appears to be a step in the right direction, in its versatility and in the availability of many iterations in very useable software, a host of issues that present challenges to the approach remains unresolved.

For instance, the crucial aspect of how to properly define the proximity measure for the 'neighbourhood' or 'local region' where the explanation applies remains unclear, and small changes in the scale of the chosen measure can lead to greatly diverging explanations. Likewise, the explanation produced by the supplemental linear model can quickly become unreliable, even with small and virtually unnoticeable perturbations of the system it is attempting to approximate. This challenges the basic assumption that there is always some simplified interpretable model that successfully approximates the underlying model reasonably well near any given data point.

LIME's creators have largely acknowledged these shortcomings and have recently offered a new explanatory approach that they call 'anchors'. These 'high precision rules' incorporate into their formal structures 'reasonable patterns' that are operating within the underlying model (such as the implicit linguistic conventions that are at work in a



sentiment prediction model), so that they can establish suitable and faithful boundaries of their explanatory coverage of its predictions or classifications.

**Global/local?  
Internal/post-hoc?**

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

**Shapley Additive  
ExPlanations  
(SHAP)**

SHAP uses concepts from cooperative game theory to define a 'Shapley value' for a feature of concern that provides a measurement of its influence on the underlying model's prediction.

Broadly, this value is calculated by averaging the feature's marginal contribution to every possible prediction for the instance under consideration. The way SHAP computes marginal contributions is by constructing two instances: the first instance includes the feature being measured, while the second leaves it out by substituting a randomly selected stand-in variable for it. After calculating the prediction for each of these instances by plugging their values into the original model, the result of the second is subtracted from that of the first to determine the marginal contribution of the feature. This procedure is then repeated for all possible combinations of features so that the weighted average of all of the marginal contributions of the feature of concern can be

Of the several drawbacks of SHAP, the most practical one is that such a procedure is computationally burdensome and becomes intractable beyond a certain threshold.

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 when conditional contributions are approximated. The result is the introduction of uncertainty into the explanation that is produced, because the

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.

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

While counterfactual explanation offers a useful way to contrastively explore how feature importance may influence an outcome, it has limitations that originate in the variety of possible features that may be included when considering alternative outcomes. In certain cases, the sheer number of potentially significant features that could be at play in counterfactual explanations of a given result can make a clear and direct explanation difficult to obtain and selected sets of possible explanations seem potentially arbitrary.

decisions about changeable human actions (like loan decisions or credit scoring), incorporating counterfactual explanation into the development and testing phases of model development may allow the incorporation of actionable variables, ie input variables that will afford decision subjects with concise options for making practical changes that would improve their chances of obtaining the desired outcome.

In this way, counterfactual explanatory strategies can be used as way to incorporate reasonableness and the encouragement of agency into the design and implementation of AI systems.

Moreover, there are as yet limitations on the types of datasets and functions to which these kinds of explanations are applicable.

Finally, because this kind of explanation concedes the opacity of the algorithmic model outright, it is less able to address concerns about potentially harmful feature interactions and questionable covariate relationships that may be buried deep within the model's architecture. It is a good idea to use counterfactual explanations in concert with other supplementary explanation strategies—that is, as one component of a more comprehensive explanation portfolio.

**Global/local?  
Internal/post-hoc?**

Counterfactual explanations are a local and post-hoc form of supplementary explanation strategy.

**Self-Explaining and  
Attention-Based  
Systems**

Self-explaining and attention-based systems actually integrate secondary explanation tools into the opaque systems so that they can offer runtime explanations of their own behaviours. For instance, an image recognition system could have a primary component, like a convolutional neural net, that extracts features from its inputs and classifies them while a secondary component, like a built-in recurrent neural net with an 'attention-directing' mechanism translates the extracted features into a natural language representation that

Automating explanations through self-explaining systems is a promising approach for applications where users benefit from gaining real-time insights about the rationale of the complex systems they are operating. However, regardless of their practical utility, these kinds of secondary tools will only work as well as the explanatory infrastructure that is actually unpacking their underlying logics. This explanatory layer must remain accessible to human evaluators and be understandable to affected individuals. Self-explaining systems, in other words, should themselves remain optimally interpretable. The

produces a sentence-long explanation of the result to the user.

Research into integrating 'attention-based' interfaces is continuing to advance toward potentially making their implementations more sensitive to user needs, explanation-forward, and humanly understandable. Moreover, the incorporation of domain knowledge and logic-based or convention-based structures into the architectures of complex models are increasingly allowing for better and more user-friendly representations and prototypes to be built into them.

task of formulating a primary strategy of supplementary explanation is still part of the process of building out a system with self-explaining capacity.

Another potential pitfall to consider for self-explaining systems is their ability to mislead or to provide false reassurance to users, especially when humanlike qualities are incorporated into their delivery method. This can be avoided by not designing anthropomorphic qualities into their user interface and by making uncertainty and error metrics explicit in the explanation as it is delivered.

**Global/local?**  
**Internal/post-hoc?**


Because self-explaining and attention-based systems are secondary tools that can utilise many different methods of explanation, they may be global or local, internal or post-hoc, or a combination of any of them.

# Annexe 4: Further reading

## PROV provenance standard


Moreau, L. & Missier, P. (2013). *PROV-DM: The PROV Data Model*. W3C Recommendation.


URL: <https://www.w3.org/TR/2013/REC-prov-dm-20130430/> 


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

## Resources for exploring algorithm types


### General


Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2), 83-85. [http://thuvien.thanglong.edu.vn:8081/dspace/bitstream/DHTL\\_123456789/4053/1/%5BSpringer%20Series%20in%20Statistics-1.pdf](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](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](http://www.dphu.org/uploads/attachements/books/books_4200_0.pdf)

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

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

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

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



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

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

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

## Supersparse linear integer model (SLIM)

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

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

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

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

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

Resources for exploring supplementary explanation strategies

## Surrogate models (SM)

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

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

Van Assche, A., & Blockeel, H. (2007). Seeing the forest through the trees: Learning a comprehensible model from an ensemble. In *European Conference on Machine Learning* (pp. 418-429). Springer, Berlin, Heidelberg. [https://link.springer.com/content/pdf/10.1007/978-3-540-74958-5\\_39.pdf](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](https://projecteuclid.org/download/pdf_1/euclid.aos/1013203451) ↗

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

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

## Individual Conditional Expectations Plot (ICE)

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

For the software in R see:

<https://cran.r-project.org/web/packages/ICEbox/index.html>

<https://cran.r-project.org/web/packages/ICEbox/ICEbox.pdf>

## Accumulated Local Effects Plots (ALE)

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

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

## Global variable importance

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

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

Fisher, A., Rudin, C., & Dominici, F. (2018). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. [arXiv:1801.01489](https://arxiv.org/abs/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](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](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.fu.ru/files/Roth2.pdf#page=39>

## Counterfactual explanation

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

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

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

## Secondary explainer 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/](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](https://github.com/pbiecek/xai_resources)

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

# Annexe 5: Argument-based assurance cases

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

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

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

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

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

## Main components of argument-based assurance cases

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

### Top-level normative goals

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

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

## Claims

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

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

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

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

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

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

## Arguments

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

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

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

## Evidence

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

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

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

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

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

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

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

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

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


### Further reading

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

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

[Fact sheets](#) 

[Datasheets for datasets](#) 

[Model cards for model reporting](#) 

[AI auditing framework blog](#) 

[Understanding Artificial Intelligence ethics and safety](#) 

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

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

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

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

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

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

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


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

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

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


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

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

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

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

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

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

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

## Background readings on methods of argument-based assurance

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

- Ninth IEEE International Symposium on High-Assurance Systems Engineering (HASE'05)* (pp. 99-108).
- Ashmore, R., Calinescu, R., & Paterson, C. (2019). Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *arXiv preprint arXiv:1905.04223*.
- Barry, M. R. (2011). CertWare: A workbench for safety case production and analysis. In *2011 Aerospace conference* (pp. 1-10). IEEE.
- Bloomfield, R., & Netkachova, K. (2014). Building blocks for assurance cases. In *2014 IEEE International Symposium on Software Reliability Engineering Workshops*(pp. 186-191). IEEE.
- Bloomfield, R., & Bishop, P. (2010). Safety and assurance cases: Past, present and possible future—an Adelard perspective. In *Making Systems Safer* (pp. 51-67). Springer, London.
- Cârlan, C., Barner, S., Diewald, A., Tsalidis, A., & Voss, S. (2017). ExplicitCase: integrated model-based development of system and safety cases. *International Conference on Computer Safety, Reliability, and Security* (pp. 52-63). Springer.
- Denney, E., & Pai, G. (2013). A formal basis for safety case patterns. In *International Conference on Computer Safety, Reliability, and Security* (pp. 21-32). Springer.
- Denney, E., Pai, G., & Habli, I. (2015). Dynamic safety cases for through-life safety assurance. *IEEE/ACM 37th IEEE International Conference on Software Engineering* (Vol. 2, pp. 587-590). IEEE.
- Denney, E., & Pai, G. (2018). Tool support for assurance case development. *Automated Software Engineering*, 25(3), 435-499.
- Despotou, G. (2004). Extending the safety case concept to address dependability. *Proceedings of the 22nd International System Safety Conference*.
- Gacek, A., Backes, J., Cofer, D., Slind, K., & Whalen, M. (2014). Resolute: an assurance case language for architecture models. *ACM SIGAda Ada Letters*, 34(3), 19-28.
- Ge, X., Rijs, R., Paige, R. F., Kelly, T. P., & McDermid, J. A. (2012). Introducing goal structuring notation to explain decisions in clinical practice. *Procedia Technology*, 5, 686-695.
- Gleirscher, M., & Kugele, S. (2019). Assurance of System Safety: A Survey of Design and Argument Patterns. *arXiv preprint arXiv:1902.05537*.
- Górski, J., Jarzębowicz, A., Miler, J., Witkowicz, M., Czyżnikiewicz, J., & Jar, P. (2012). Supporting assurance by evidence-based argument services. *International Conference on Computer Safety, Reliability, and Security* (pp. 417-426). Springer, Berlin, Heidelberg.
- Habli, I., & Kelly, T. (2014, July). Balancing the formal and informal in safety case arguments. In *VeriSure: Verification and Assurance Workshop, colocated with Computer-Aided Verification (CAV)*.
- Hawkins, R., Habli, I., Kolovos, D., Paige, R., & Kelly, T. (2015). Weaving an assurance case from design: a model-based approach. *2015 IEEE 16th International Symposium on High Assurance Systems Engineering* (pp. 110-117). IEEE.
- Health Foundation (2012). Evidence: Using safety cases in industry and healthcare.



- Kelly, T. (1998) *Arguing Safety: A Systematic Approach to Managing Safety Cases*. Doctoral Thesis. University of York: Department of Computer Science.
- Kelly, T., & McDermid, J. (1998). Safety case patterns-reusing successful arguments. *IEEE Colloquium on Understanding Patterns and Their Application to System Engineering*, London.
- Kelly, T. (2003). *A Systematic Approach to Safety Case Management*. SAE International.
- Kelly, T., & Weaver, R. (2004). The goal structuring notation—a safety argument notation. In *Proceedings of the dependable systems and networks 2004 workshop on assurance cases*.
- Maksimov, M., Fung, N. L., Kokaly, S., & Chechik, M. (2018). Two decades of assurance case tools: a survey. *International Conference on Computer Safety, Reliability, and Security* (pp. 49-59). Springer.
- Nemouchi, Y., Foster, S., Gleirscher, M., & Kelly, T. (2019). Mechanised assurance cases with integrated formal methods in Isabelle. *arXiv preprint arXiv:1905.06192*.
- Netkachova, K., Netkachov, O., & Bloomfield, R. (2014). Tool support for assurance case building blocks. In *International Conference on Computer Safety, Reliability, and Security* (pp. 62-71). Springer.
- Picardi, C., Hawkins, R., Paterson, C., & Habli, I. (2019, September). A pattern for arguing the assurance of machine learning in medical diagnosis systems. In *International Conference on Computer Safety, Reliability, and Security* (pp. 165-179). Springer.
- Picardi, C., Paterson, C., Hawkins, R. D., Calinescu, R., & Habli, I. (2020). Assurance Argument Patterns and Processes for Machine Learning in Safety-Related Systems. *Proceedings of the Workshop on Artificial Intelligence Safety* (pp. 23-30). CEUR Workshop Proceedings.
- Rushby, J. (2015). The interpretation and evaluation of assurance cases. Comp. Science Laboratory, SRI International, Tech. Rep. SRI-CSL-15-01.
- Strunk, E. A., & Knight, J. C. (2008). The essential synthesis of problem frames and assurance cases. *Expert Systems*, 25(1), 9-27.