Project ExplAIn

Explaining decisions made with AI

Draft guidance for consultation

Part 1:

The basics of explaining AI

ico.
Information Commissioner's Office

The Alan Turing Institute

About this guidance

What is the purpose of this guidance?

This guidance outlines the definitions, the legal basis for explaining AI and the benefits and risks of doing so, as well as the explanation types 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.

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 and the benefits and risks of explaining AI systems to the individuals affected by their use.

What is the status of this guidance?

This guidance is issued in response to the commitment in the Government's AI Sector Deal, but it is not a statutory code of practice under the Data Protection Act 2018.

This is practical guidance that sets out good practice for explaining decisions to individuals that have been made using AI systems processing personal data. It clarifies the application of data protection provisions associated with explaining AI decisions, as well as highlighting other relevant legal regimes outside the ICO's remit.

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) is an umbrella term for a range of algorithm-based technologies that often try to mimic human thought to solve complex tasks. 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 a decision supported by an AI system 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.

There are several ways of building 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 large amounts of expert input and rules. But recently more sophisticated techniques, such as unsupervised and reinforcement machine learning, have enabled the creation of AI models through automated discovery of patterns and correlations in data without extensive programming. This guidance deals with supervised as well as unsupervised and reinforcement learning systems.

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. It is therefore important to consider how and why these systems create the outputs they do.

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:

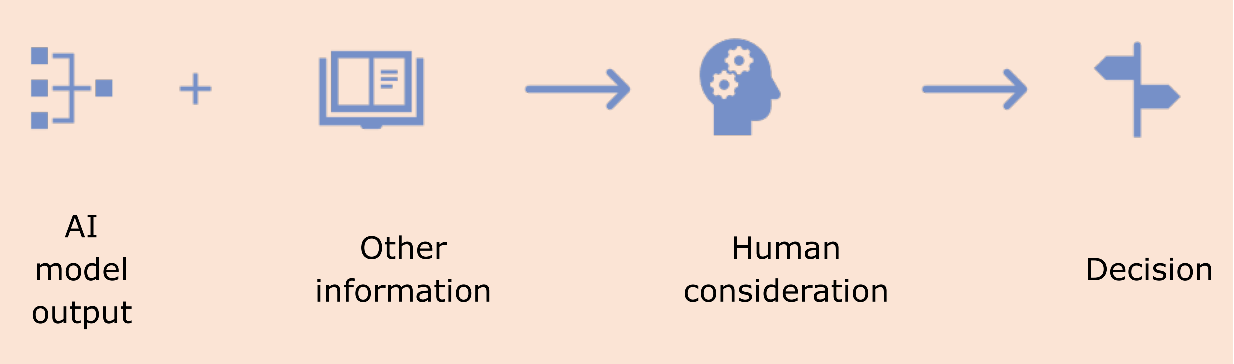
- a prediction, eg you will not default on a loan;
- a recommendation, eg you would like this news article; or
- a classification, eg this email is spam.

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.

AI model output

Decision

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’.



For more information on what constitutes meaningful human involvement in an AI-assisted decision process, you can read our guidance on automated decision-making and profiling in the Guide to the GDPR, and a blogpost on this topic in our AI auditing framework blog.

Guide to the GDPR
AI auditing framework: the role of meaningful human reviews blog

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.

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

One of the key differences between a decision that has been made by an AI system, and one where no AI system has been used, 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.

Individuals should not lose 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.

Checklist

☐ We are compliant with the GDPR where it refers to explanations.

☐ We have completed a Data Protection Impact Assessment.

☐ We are aware of the other legal requirements with which we must comply.

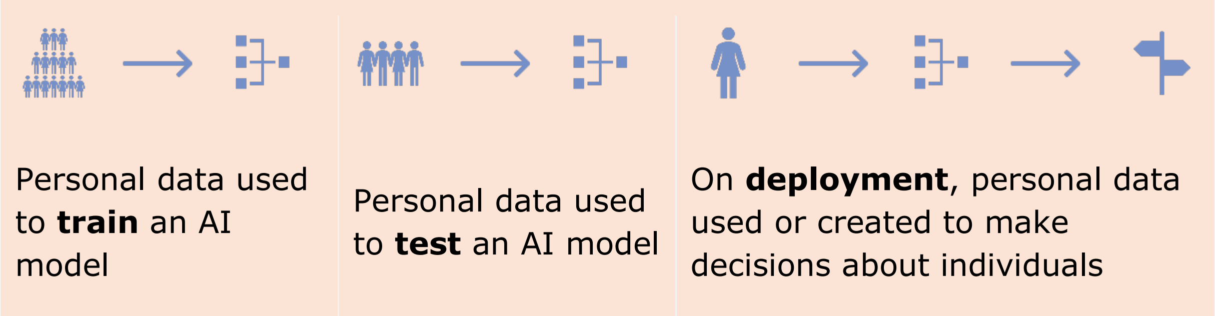
In more detail

- 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.

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.



Personal data used to **train** an AI model

Personal data used to **test** an AI model

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 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. It 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, to express their view, and to 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) 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.

So, many of the rights it gives to individuals, and the obligations it places on organisations, are directly relevant to the use of AI.

The ICO has published additional guidance on DPIAs, including a list of processing operations which require a DPIA. The list mentions AI, machine learning and deep learning.

> Guide to Data Protection Impact Assessments

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.

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, for example:

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

> Equality and Human Rights Commission
> Reasonable adjustments

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'.

> Administrative Law and the Machines of Government
> Algorithm-assisted decision-making in the public sector

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 may not be 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, your organisation 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 organisations 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 about your explanations of AI-assisted decisions disclosing 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 such as feature weightings or importance), you should justify and document your reasons for this.

Third-party personal data

Due to the way in which you train your AI model, or input data for particular decisions, you may be concerned about the inappropriate disclosure of the 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 relate to more than one person).

You should assess this risk as part of a data protection impact assessment, 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 the purpose of the AI-assisted decisions you make is 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 in these circumstances.

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. 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 determine this is necessary. 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 to regulatory intervention or action. The ICO utilises 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 of 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/or citizens.

Disengaged public

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

What goes into an explanation?

At a glance

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; what data has been used to train and test the AI model 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:** the impact 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 you 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 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.

We 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, as well as the right to object and the right to obtain human intervention.

As a result of our research and engagement we identified six different types of explanation, which you can combine into your explanation in various ways depending on the decision at hand.

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 can gain an understanding of 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 so, knowing its reasoning supports them to formulate a coherent argument for why they think this is the case if they wish to challenge the decision.

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 can still be useful so that they may validate their belief as to why this was the case, or adjust it if the reasons for the favourable outcome were different to those they expected.

How can the guidance help me with this?

See Explaining AI in practice for more information on extracting the technical rationale and translating this into understandable 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.

How can the guidance help me with this?

See What explaining AI means for your organisation for more information on identifying the roles involved in explaining an AI-assisted decision. See Explaining AI in practice for more information on the information you need to provide for this explanation.

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 that the AI decisions you make are generally unbiased and fair. 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 meaningful 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.

How can the guidance help me with this?

See Explaining AI in practice for more information on building fairness into the design and deployment of your AI model. See also 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.

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.

How can the guidance help me with this?

See Explaining AI in practice for more information on ensuring the accuracy, reliability, security and robustness of your AI system. See also 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. Impact explanations are therefore often well suited to delivery before an AI-assisted decision has been made. See Step 7 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.

How can the guidance help me with this?

See Explaining AI in practice for more information on considering impact into how you select an appropriately explainable AI model. See also What explaining AI means for your organisation for information on how to document this.

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. It is important that 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 are a complement to the data protection principles outlined under the GDPR.

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 the decision being made.

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 in Step 7 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 rationale explanation in Step 3 of Explaining AI in Practice.
- Selecting your priority explanations in Step 1 of Explaining AI in practice.
- Explanation timing in Step 7 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 steps 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 AI-assisted decision process, from concept to deployment and presentation of the explanation to the decision recipient.

There are therefore multiple types of context that we address in this guidance:

- **Sector context**, which refers to the domain-specific expectations, requirements and conventions of the sector within which you are setting up and using your AI system.
- **Use case context**, which refers to the specific objectives, settings and risks of your AI application.
- **Decision recipient's individual circumstances**, which are an important consideration for how your organisation should apply the output of an AI system.
- **Contextual factors** that influence how best to deliver your explanation to the decision recipient, based on what that individual will find most useful.

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.

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 principle 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.

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

Be transparent	**Rationale** • What is the technical logic or reasoning behind the model’s output? • Which input features, parameters and correlations played a significant role in the calculation of the model’s result and how? • How can the technical rationale underlying the model’s output be rendered into easily understandable and plain-language reasons that may be subjected to rational evaluation by affected individuals or their representatives? • How can the statistical results be applied to the specific circumstances of the individual receiving the decision? **Data** • What data did you use to train the model? • Where did the data come from? • How did you ensure the quality of the data you used?
Be accountable	**Responsibility** • Who is to be held 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? • What are the mechanisms they will be held accountable by? • How have design and implementation processes been made traceable and auditable across the entire project?
Consider context	See Step 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 Step 6 of Explaining AI in practice to see how contextual factors 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 the objectives of preventing discrimination and of mitigating bias been sufficiently integrated into the design and implementation of the system? • Have formal criteria of fairness that determine the distribution of outcomes been incorporated into the system and made 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? • Has the model been designed, verified, and validated in a way that sufficiently ensures that it is secure, accurate, reliable, and robust? • Have sufficient measures been taken to ensure that the system dependably operates in accordance with its designers' expectations when confronted with unexpected changes, anomalies, and perturbations? **Impact** • Have impacts on the wellbeing of affected individuals and communities been sufficiently

	considered from start to finish of the AI model's design and deployment?