Explaining decisions made with AI: summary

Part 1

Definitions

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 a decision supported by an AI system should be able to hold someone accountable for it.

Legal framework

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.

Benefits and risks

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.

What goes into an explanation?

When providing an explanation, 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.

What are the contextual factors?

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.

The principles to follow

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.

Part 2

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

- Getting to know the different types of explanation will help you identify the dimensions of an explanation that decision recipients will find useful.

- In most cases, explaining AI-assisted decisions involves identifying what is happening in your AI system and who is responsible. That means you should prioritise the rationale and responsibility explanation types.

- The setting and sector you are working in is important in figuring out what kinds of explanation you should be able to provide. You should therefore consider domain context and use case.

- In addition, consider the potential impacts of your use of AI to determine which other types of explanation you should provide. This will also help you think about how much information is required, and how comprehensive it should be.

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

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

- The data that you collect and pre-process before inputting it into your system has an important role to play in the 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 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 aid in your impact explanation.

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

- 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 of 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.

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

- 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.

Task 5: Prepare implementers to deploy your AI system

- In cases where decisions are not fully automated, implementers need to be meaningfully involved.

- This means that they need to be appropriately trained to use the model's results responsibly and fairly.

- Their training should cover:
    - the basics of how machine learning works;
    - the limitations of AI and automated decision-support technologies;

  - the benefits and risks of deploying these systems 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.

Task 6: Consider how to build and present your explanation

- 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 the process-based or outcome-based explanations.

- You should then revisit the contextual factors to establish which explanation types should be prioritised.

- 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.

Part 3

Organisational roles and functions for explaining AI

- 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.

Policies and procedures

- 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 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.

Documentation

- 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 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.