AI Explainability in Practice

What is the AI Ethics and Governance in Practice Programme?

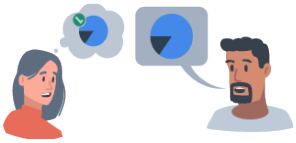
In 2021, the UK's National AI Strategy recommended that UK Government's official Public Sector Guidance on AI Ethics and Safety be transformed into a series of practice-based workbooks. The result is the AI Ethics and Governance in Practice Programme. This series of eight workbooks provides end-to-end guidance on how to apply principles of AI ethics and safety to the design, development, deployment, and maintenance of AI systems. It provides public sector organisations with a Process-Based Governance (PBG) Framework designed to assist AI project teams in ensuring that the AI technologies they build, procure, or use are ethical, safe, and responsible.

At a Glance

- Explores how, why, and when explanations of AI-supported or -generated outcomes need to be provided, and what impacted people's expectations are about what these explanations should include.
- Addresses the essential questions of *What do we need to explain? And who do we need to explain this to?*
- Discusses the tasks needed to help design and deploy appropriately transparent and explainable AI systems and to assist in providing clarification of the results these systems produce to a range of impacted stakeholders. This involves **Explainability Assurance Management**:
    - A systematic approach to selecting, extracting and delivering explanations that are differentiated according to the needs and skills of the different audiences they are directed at.

The Alan Turing Institute

Key Concepts



AI Explainability

The degree to which a system or a set of governance practices and tools support a person's ability to:

1. Explain and communicate the rationale underlying the behaviour of the system.

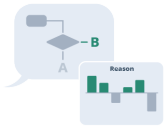
2. Demonstrate and convey that the processes behind its design, development, and deployment have been undertaken in ways that ensure its sustainability, safety, fairness, and accountability across its particular contexts of use and application.

AI Transparency

It encompasses two meanings:

1. Interpretability of an AI system or the ability to know how and why a model performed the way it did in a specific context and therefore the ability to understand the rationale behind its decision or behaviour. This sort of transparency is often referred to by way of the metaphor of 'opening the black box' of AI. It involves content clarification and intelligibility.

2. Transparent AI asks that the designers and developers of AI systems demonstrate that their processes and decision-making, in addition to system models and outputs, are sustainable, safe, fair, and driven by responsibly managed data.



Outcome-based explanations

Explanations that include the components and reasoning behind model outputs while delineating contextual and relational factors.

Process-based explanations

Explanations that demonstrate that the AI project team has followed good governance processes and best practices throughout the AI project lifecycle.

Workbook Summary

Understanding how, why, and when explanations of AI-supported or -generated outcomes need to be provided, and what impacted people's expectations are about what these explanations should include, is crucial to fostering responsible and ethical practices within your AI projects. To guide AI project teams through this process, we first define AI Explainability, as well as outcomes- and process-based explanations. We then explore the maxims that provide a broad steer on what to think about when explaining AI/machine learning-assisted decisions to individuals and the high-level considerations for project teams to consider so as to achieve higher degrees of explainability of the model and improved interpretability of outputs to wide and diverse audiences. We then take a deeper dive into different types of explanations and explore the means to identify when and how to employ them effectively.

To put the principle of AI Explainability into practice across the AI project workflow, we then explore the tasks to ensure that the design, development, and deployment of AI projects is done in a transparent and explanation-aware fashion and to facilitate the selection, extraction and delivery of explanations that are differentiated according to the needs and skills of the different audiences they are directed at. We then present the Explainability Assurance Management template, which aims to help AI project teams to accomplish these six tasks.

Six Types of Explanations

Rationale Explanation

Helps people understand the reasons that led to a decision outcome.

Impact Explanation

Helps people understand the considerations taken about the effects that the AI decision-support system may have on an individual and society.

Responsability Explanation

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.

Fairness Explanation

Helps people understand the steps taken to ensure AI decisions are generally unbiased and equitable, and whether or not they have been treated equitably themselves.

Data Explanation

Helps people understand what data about them, and what other sources of data, were used in a particular AI decision, as well as the data used to train and test the AI model.

Safety Explanation

Helps people understand the measures that are in place and the steps taken to maximise the performance, reliability, security, and robustness of the AI outcomes, and the justification for the chosen type of AI system.

Putting AI Explainability into Practice

Explainability Assurance Management

There are a number of tasks both to help you design and deploy appropriately transparent and explainable AI systems and to assist you in providing clarification of the results these systems produce to a range of impacted stakeholders (from operators, implementers, and auditors to decision recipients).

These tasks make up Explainability Assurance Management for AI projects:

- **Task 1.** Select priority explanations by considering the domain, use case, and impact on the individual.
- **Task 2.** Collect and pre-process your data in an explanation-aware manner.
- **Task 3.** Build your system to ensure you are able to extract relevant information for a range of explanation types.
- **Task 4.** Translate the rationale of your system's results into useable and easily understandable reasons.
- **Task 5.** User training.
- **Task 6.** Consider how to build and present your explanation.

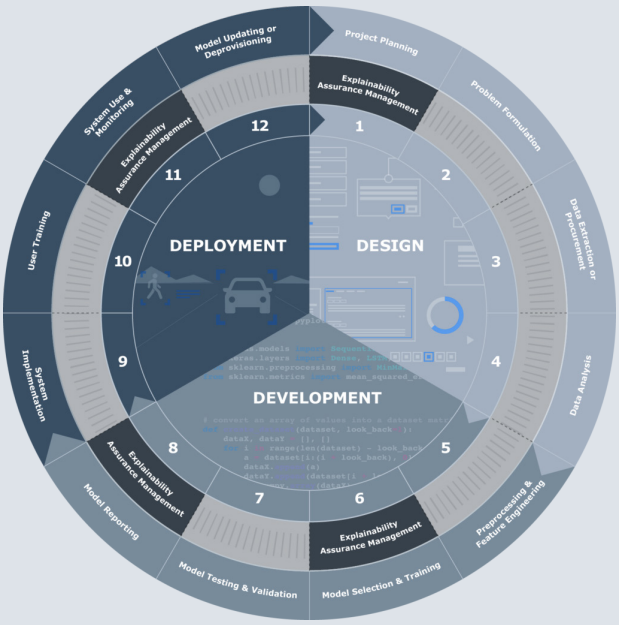


Figure 1: The Explainability governance actions within the Process-Based Governance (PBG) Framework.

For detailed information about authorships, acknowledgements, and references, please consult the **AI Explainability in Practice** workbook. The workbook series of the AI Ethics and Governance in Practice programme is available at: aiethics.turing.ac.uk