AI Fairness 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 concepts of fairness are applied in the field of AI ethics and governance.
- Provides an overview of how different social, technical, and sociotechnical contexts of the AI project lifecycle give rise to different fairness concerns.
- Discusses actions needed to identify and mitigate unfair bias and discrimination across the AI project workflow, including:
  - **Bias Self-Assessment and Bias Risk Management**, which facilitates the iterative identification and documentation of risks of bias across the lifecycle and assurance actions implemented to address these.
  - **Fairness Position Statement**, which documents the metric-based fairness criteria for individual AI projects, providing an explanation in plain and nontechnical language.

The Alan Turing Institute

Key Concepts

Fairness

General ethical and legal concepts of fairness are founded on core beliefs in the equal moral status of all human beings and the right of all human beings to equal respect, concern, protection, and regard before the law. Fairness, on this view, has to do with the moral duty to treat others as moral equals and to secure the membership of all in a 'moral community' where every person can regard themself as having equal value. Wrongful discrimination occurs when decisions, actions, institutional dynamics, or social structures do not respect the equal moral standing of individual persons.

Public Sector Equality Duty (PSED)

Under the PSED, if your organisation is planning to use a specific AI technology to deliver one of its functions or services, it must:

1. think about its potential impact (negative and positive) on people with protected characteristics under the Equality Act 2010 before going ahead with it; and

2. monitor its actual impact during and after implementation. The latter is essential to satisfy the ongoing nature of the PSED.

Principle of Discriminatory Non-Harm

This principle states that the producers and users of AI systems should prioritise the identification and mitigation of biases and discriminatory influences, which could lead to **direct** or **indirect discrimination** or **discriminatory harassment.**

- In **direct discrimination**, individuals are treated adversely based on their membership in some protected class. It involves instances where otherwise similarly positioned individuals receive different and more-or-less favourable treatment on the basis of differences between their respective protected characteristics.
- In **indirect discrimination**, existing provisions, criteria, policies, arrangements, or practices—which could appear on their face to be neutral—disparately harm or unfairly disadvantage members of some protected class in comparison with others who are not members of that group.
- In **discriminatory harassment**, unwanted or abusive behaviour linked to a protected characteristic violates someone's dignity, degrades their identity, or creates an offensive environment for them.

Workbook Summary

Reaching consensus on a commonly accepted definition of AI Fairness has long been a central challenge in AI ethics and governance. There is a broad spectrum of views across society on what the concept of fairness means and how it should best be put to practice. In this workbook, we tackle this challenge by exploring how a context-based and society-centred approach to understanding AI Fairness can help project teams better identify, mitigate, and manage the many ways that unfair bias and discrimination can crop up across the AI project workflow.

We begin by exploring how, despite the plurality of understandings about the meaning of fairness, priorities of equality and non-discrimination have come to constitute the broadly accepted core of its application as a practical principle. We focus on how these priorities manifest in the form of equal protection from direct and indirect discrimination and from discriminatory harassment. These elements form ethical and legal criteria based upon which instances of unfair bias and discrimination can be identified and mitigated across the AI project workflow.

We then take a deeper dive into how the different contexts of the AI project lifecycle give rise to different fairness concerns. This allows us to identify several types of AI Fairness (Data Fairness, Application Fairness, Model Design and Development Fairness, Metric-Based Fairness, System Implementation Fairness, and Ecosystem Fairness) that form the basis of a multi-lens approach to bias identification, mitigation, and management. Building on this, we discuss how to put the principle of AI Fairness into practice across the AI project workflow through Bias Self-Assessment and Bias Risk Management as well as through the documentation of metric-based fairness criteria in a Fairness Position Statement.

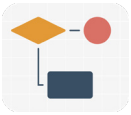
Six Types of Fairness

Data Fairness

The AI system is trained and tested on datasets that are properly representative, fit-for-purpose, relevant, accurately measured, and generalisable.

Application Fairness

The policy objectives and agenda-setting priorities of an AI project are non-discriminatory and are acceptable to and line up with the aims, expectations, and sense of justice possessed by impacted people.



Model Design and Development Fairness

The AI system has a model architecture that does not include target variables, features, processes, or analytical structures which are discriminatory, unreasonable, morally objectionable, or unjustifiable or that encode social and historical patterns of discrimination.

Metric-Based Fairness

Lawful, clearly defined, and justifiable formal metrics of fairness have been operationalised in the AI system. They have been made transparently accessible to relevant stakeholders and impacted people.

System Implementation Fairness

The AI system is deployed by users sufficiently trained to implement it. They have an appropriate understanding of its limitations and strengths and deploy it in a bias-aware manner that gives due regard to the unique circumstances of affected individuals.

Ecosystem Fairness

The economic, legal, cultural, and political structures or institutions in which the AI project lifecycle is embedded do not steer AI research and innovation agendas in ways that entrench or amplify asymmetrical and discriminatory power dynamics or that generate inequitable outcomes for protected, marginalised, vulnerable, or disadvantaged social groups.

Putting AI Fairness into Practice

Bias Self-Assessment and Bias Risk Management

At each stage of the AI project lifecycle, the project team should:

- reflect on how the AI project might be vulnerable to biases that may arise at each stage;
- identify biases that may be present across the project workflow; and
- determine and document bias risk mitigation actions to correct any identified biases and strengthen specific stages in the workflow that have possible discriminatory consequences.

Fairness Position Statement

When decisions about fairness criteria have been finalised, the project team should prepare a statement in which the metric-based fairness criteria being employed in the model is made explicit and explained in plain and nontechnical language. The statement should be made publicly available for review by all affected stakeholders.

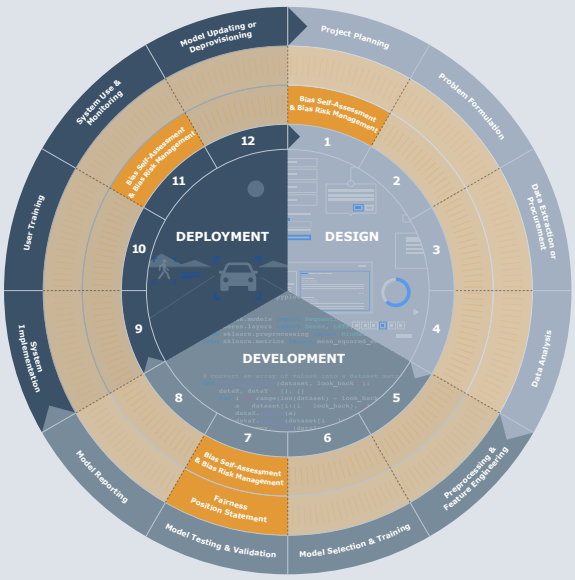


Figure 1: The Fairness governance actions within the Process-Based Governance (PBG) Framework.

For detailed information about authorships, acknowledgements, and references, please consult the **AI Fairness in Practice** workbook.