Ministry of Justice

Ministry of Justice
AI and data science ethics framework

Introduction to the SAFE-D framework

The Alan Turing Institute

1

These materials were produced in collaboration with the Alan Turing Institute through an extended engagement with personnel at the UK Ministry of Justice.

Ethics process

1 Introduction to the SAFE-D framework

Read the introduction booklet and familiarise yourself with the project lifecycle model.

2 SAFE-D principles booklet

Read the SAFE-D principles booklet and familiarise yourself with the principles and their core attributes.

3 Design phase activity booklet

- SAFE-D identification workshop exercise
- litmus test
- stakeholder engagement worksheet
- additional activities where relevant

4 Development phase activity booklet

- SAFE-D reflection workshop exercise
- development phase questionnaire
- stakeholder engagement worksheet
- additional activities where relevant

5 Deployment phase activity booklet

- SAFE-D assurance workshop exercise
- deployment phase questionnaire
- stakeholder engagement worksheet
- additional activities where relevant

What is the MoJ AI and data science ethics framework?

What is the framework?

The MoJ AI and data science ethics framework is a comprehensive collection of processes, tools, and guidance aimed at fostering a responsible ecosystem for data-driven technologies.

It promotes ethical reflection and deliberation throughout all stages of a project's lifecycle, helping teams integrate ethical considerations into their work. This framework seeks to maximise the benefits of a project while minimising risks and harms.

The Framework is centred around the 'SAFE-D principles'. SAFE-D stands for Sustainability, Accountability, Fairness, Explainability and Data responsibility.

What is the purpose of the framework?

The main purpose is to guide project teams in embedding ethical considerations into their work, ensuring that technological advancements are responsible and beneficial to society.

What are the SAFE-D principles?

SAFE-D stands for:

Sustainability
Accountability
Fairness
Explainability
Data responsibility

Each SAFE-D principle includes core attributes that define what the principle means when creating data technologies for the criminal justice system.

These core attributes clarify and narrow down the principle, making it easier to apply in practice. You can use them to check how well you are upholding each SAFE-D principle in your work.

You can find out more in the **SAFE-D principles booklet**.

What is data ethics?

What is meant by 'ethics' in the context of this framework?

Data ethics refers to the principles and practices that guide the responsible and ethical use of AI and data in research, technology, and business.

Within the framework context, ethics draws from various philosophical traditions, such as egalitarianism, virtue ethics, care ethics, and liberal political theory. It also incorporates legal principles concerning data protection, privacy, and human rights.

Why is AI and data ethics important?

Ethical decision-making ensures fair and humane treatment, helping to protect individual rights and maintain dignity and respect. This is particularly important in the context of the criminal justice system.

AI/data ethics is fundamental for ensuring that technology serves humanity in a way that is fair, safe, and beneficial to society.

By addressing concerns like fairness, transparency, accountability, and security, ethics helps prevent harm, mitigate risks, and build systems that reflect shared human values.

SAFE-D principles

S	**Sustainability** Ensure safe and reliable outputs and practices to ensure mitigation of long-term risk. Establish clear roles and responsibilities across projects to have a single point of contact.
A	**Accountability** Implement transparent processes, outcomes and communications channels to inform relevant stakeholders effectively.
F	**Fairness** Ensure prevention of discrimination resulting from project outcomes. Recognise rights and interests that a project may affect, balancing interests of all parties.
E	**Explainability** Assess and support ability for someone to explain the behaviour of a data-driven technology within a system. This principle is driven by expertise and skills.
D	**Data responsibility** Consider aspects of data and datasets including data quality, relevance to the project and data integrity. This also includes considerations around legal and policy obligations.

Introduction to the project lifecycle model

Project lifecycle model

To create a safe and ethical system, a project team needs a flexible design, the right tools, and sufficient skills and resources.

The project lifecycle model acts as a framework to help teams identify where these elements belong.

It supports teams in:

- reflecting on necessary tasks at each stage of the project
- considering how these tasks might affect project goals, like fairness in classification
- making ongoing decisions and documenting actions as the project progresses

The model outlines the typical stages involved in the design, development, and deployment of data-driven technologies, forming the basis for ethical activities throughout the project lifecycle.

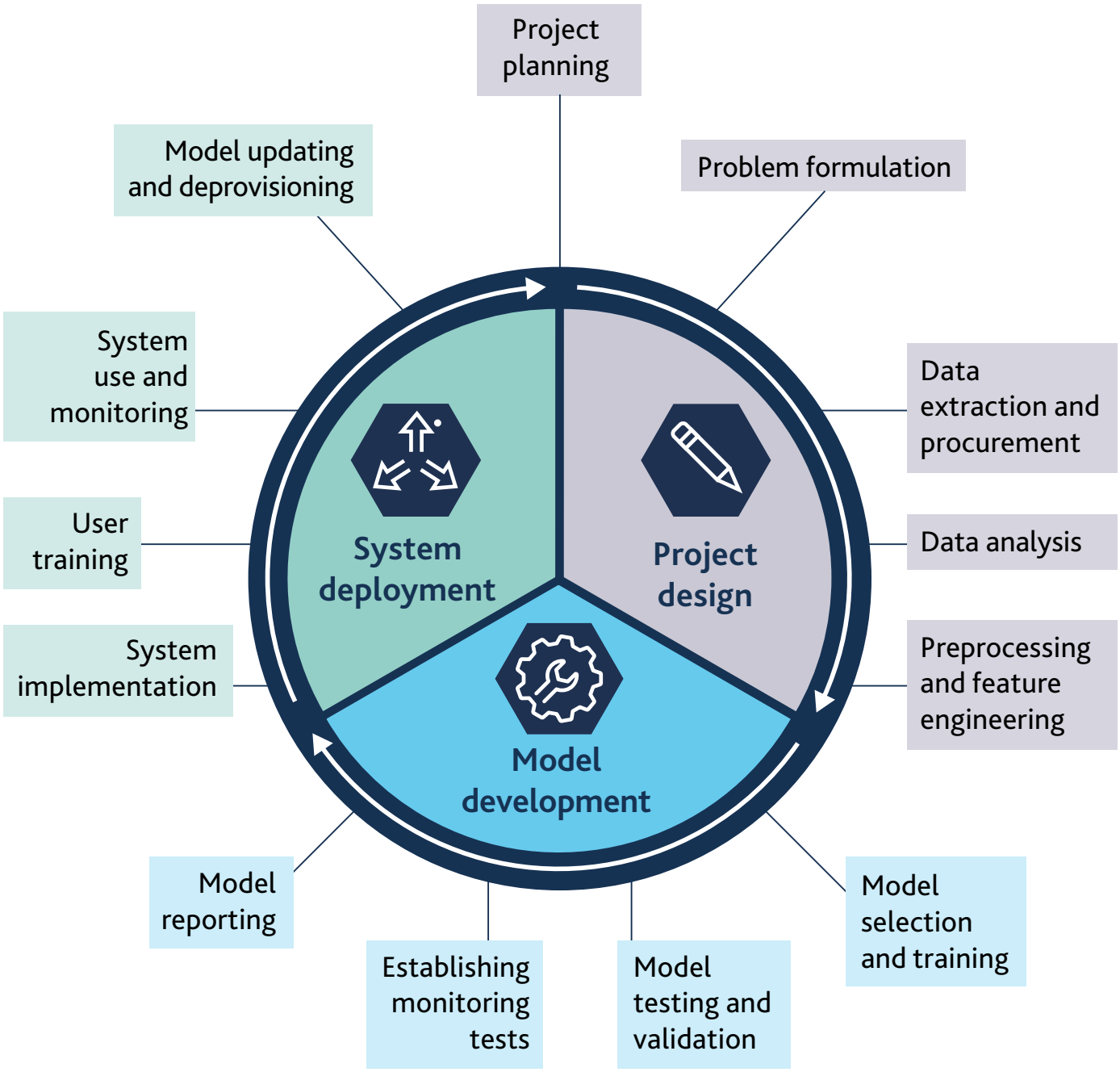
Key components of the model

Three overarching stages:

- **Project design:** preliminary tasks and activities that set the foundations for the development of the model and system.
- **Model development:** covers technical tasks such as training, testing, and validating the machine learning model to ensure it's suitable for its purpose.
- **System deployment:** focuses on the safe implementation and use of the system, including ongoing monitoring and updates.

Thirteen lower-level stages:

- These stages provide more detailed activities under each overarching stage to support ethical deliberation throughout the project.



The model outlines the typical stages of a project involving the design, development, and deployment of data-driven technology.

Detailed explanations of the lower-level stages can be found in the activity booklets for each of the three main sections.

Getting started

Framework FAQs

Who can use the framework?

The MoJ AI and data science ethics framework is intended for anyone involved in the design, development, and deployment of data-driven technologies. It is particularly relevant for those in roles defined by the government digital and data profession capability framework.

When should you use the framework?

The framework should be used from the initial stage of projects that involve designing, developing, and deploying AI models and systems.

Who is responsible for completing the framework?

The responsibility for completing the framework lies with the project team. Team members can take on different roles to manage various sections collaboratively.

What should you know before using the framework?

The framework provides explanations of all terms and concepts necessary for its effective implementation.

When to revisit the framework?

Teams should determine how often to revisit the framework to ensure that the information gathered is accurate and up-to-date.

What do I do next?

The MoJ AI and data science ethics framework combines the project lifecycle model and the SAFE-D principles to help teams identify and address ethical risks throughout the project lifecycle.

It is important to remember that:

- **ethics is iterative:** it's not a linear process or a simple checklist; the approach may vary for different projects and you'll need to revisit this framework at various points
- **you can ask for help:** the data ethics team and champions network can provide support and guidance to help you apply the framework to your project
- **legal and organisational requirements:** ethical assessments do not replace existing legal and organisational processes, such as data protection impact assessments (DPIAs) or assurance quality assessments (AQAs). These should still be conducted as usual

Contact us:

DataEthics@justice.gov.uk