Ministry
of Justice

Ministry of Justice
AI and data science ethics framework

Design phase activity booklet

The
Alan Turing
Institute

3

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

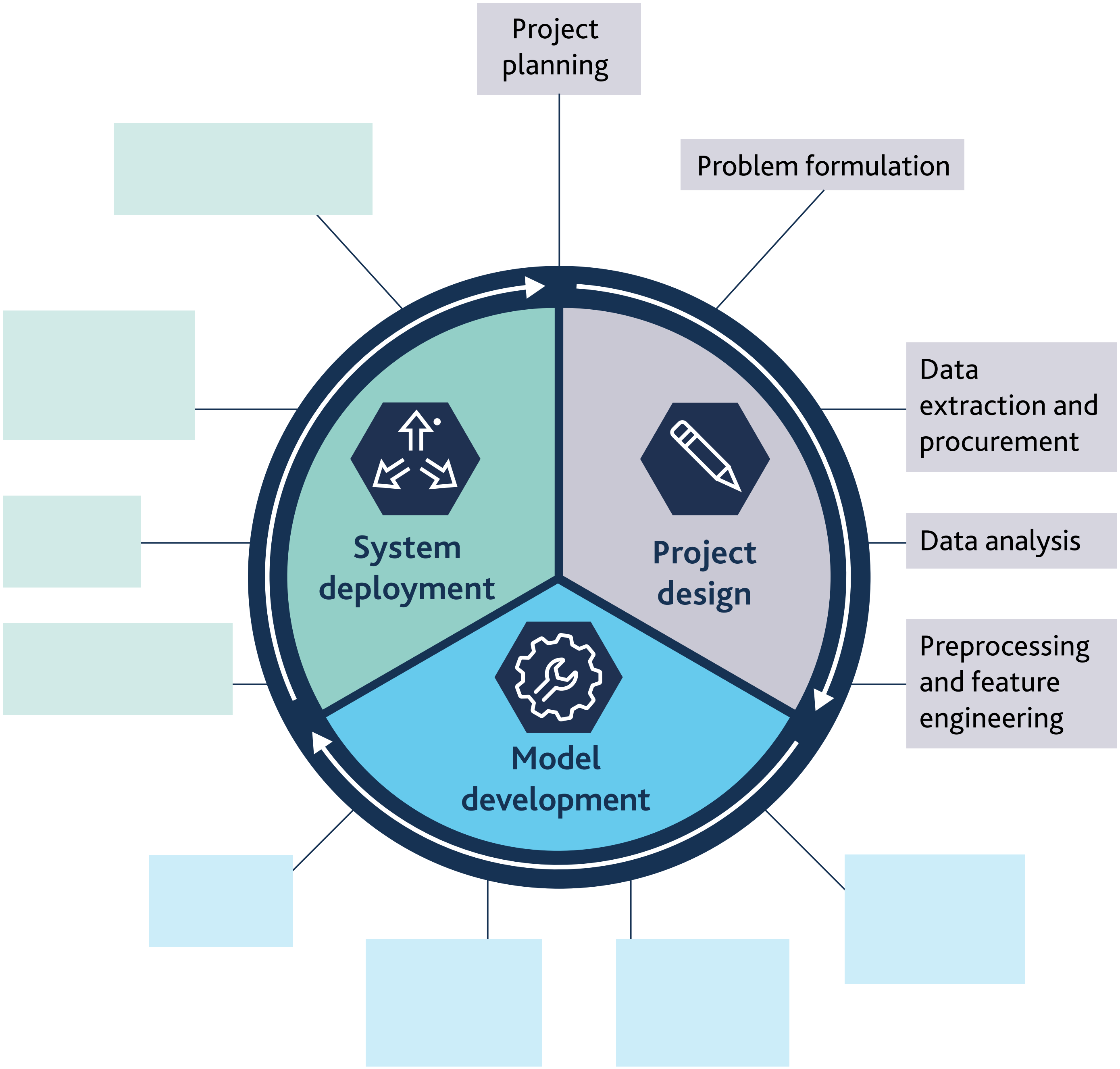
5 Deployment phase activity booklet

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

Project design phase

Project design

Preliminary tasks and activities that set the foundations for the development of the model and system.



Project design phase

Lower-level lifecycle stage	Ethical significance
Project planning Project planning includes initial activities to define the project's aims, objectives, scope, processes, and assess potential risks and benefits.	**Foundation for success:** encourages anticipatory and reflective activities, providing a stable base for the project. **Team agreements:** allows the team to establish 'red lines', identifying contexts where the system shouldn't be used and data types that are off-limits. **Setting milestones:** helps the team establish milestones and objectives to track progress and ensure goals are met throughout the project.
Problem formulation This task involves creating a clear statement about the main problem the project aims to solve, and a detailed description of the computational process involved.	**Understanding the problem:** this task helps clarify the validity of the project by defining the computational process and its higher-level goals. **Addressing broader issues:** it ensures the project's goal is valid and that the system can effectively tackle wider social or policy challenges. This clarity also aids in engaging stakeholders and enhancing project communication.

Lower-level lifecycle stage	Ethical significance
Data extraction (or procurement) Data extraction involves designing methods for gathering data based on earlier planning and problem formulation, as well as the actual collection and storage of new or existing data.	The principle of 'garbage in, garbage out' highlights the importance of this task. Machine learning and AI systems rely heavily on the quality of input data. **Responsible practices:** proper data extraction is crucial for creating accountable and trustworthy services, developing safe and fair algorithms, and ensuring sustainable, privacy-preserving systems.
Data analysis Data analysis is divided into two types: • **exploratory data analysis:** this helps analysts understand the dataset's structure and identify possible relationships between data types and variables • **confirmatory data analysis:** this evaluates initial hypotheses from the previous stage using statistical methods like significance testing	In responsible research and innovation, data analysis is crucial for identifying biases that could negatively affect the project. **Missing data:** recognising and addressing missing data is especially important. While engaging stakeholders early can reduce bias, understanding its extent and how to address it largely depends on the quality of the data analysis.

Lower-level lifecycle stage	Ethical significance
Preprocessing and feature engineering Data analysis can provide valuable insights, but the data structures may not be suitable for training machine learning algorithms. Thus, preprocessing and feature engineering are necessary to clean and transform data into the features used for model training and testing.	**Feature dependence:** features are derived from the raw data collected earlier. They can be created manually or using algorithms to enhance ML performance. **Impact on ethics:** the features chosen not only affect the model's accuracy but also its ethical implications, such as reducing explainability or leading to biased outcomes. **Complex selection process:** choosing the right features is crucial but often complex and time-consuming, requiring trade-offs between aspects like predictive power and interpretability.

SAFE-D workshop activity

Workshop exercise: SAFE-D identification

As a project team, and with additional appropriate stakeholders, we encourage you to complete a SAFE-D identification workshop.

This will help you to:

- **identify** potential ethical risks and issues upfront
- **assess** the relevance of the SAFE-D principles in your project context and **weigh** them accordingly
- **define** the SAFE-D principles in your project context, allowing for more targeted and proportionate ethical interventions

To get started, you should:

- nominate a workshop host
  - this could be the project lead or a volunteer/s
- send a calendar invite to ensure dedicated time is set aside for this activity
  - we suggest anywhere between 60–90 minutes to make sure you can facilitate productive discussions. You may want to split this across multiple sessions
- make sure you have access to a Miro board template for your team and that attendees are comfortable using Miro
- read through the activity instructions in this booklet and make sure you are comfortable with the content. Set up an agenda with timings that will work for your team

Activity overview

As a team, brainstorm and identify which of the SAFE-D principles might be impacted by your project. This is an opportunity to gather initial thoughts on ethical risks and issues during the design phase.

Steps:

1) Individual reflection (5–10 minutes)

- Take a moment to think about the SAFE-D principles and their core attributes.
- Write down your thoughts regarding potential ethical risks and issues related to our project on post-it notes under each of the SAFE-D principles.

2) Group discussion (10 mins)

- After individual brainstorming, come together as a team to discuss the post-it notes.
- Group similar ideas into key themes to identify overarching concerns.

Prompts for discussion:

- How might our project impact different stakeholders?
- Are there any potential biases we should address?
- What steps can we take to enhance accountability and transparency?
- How can we ensure the project remains sustainable and fair?

Remember:
This exercise is about open dialogue and gathering diverse perspectives, so feel free to share any initial thoughts, challenges or concerns you may have.

Workshop exercise: SAFE-D weighing

Activity overview

Weighing ethical principles involves evaluating which principles align most closely with the goals and objectives of your project. This process is crucial for addressing ethical risks and resolving potential conflicts between principles.

Steps:

1) Group discussion: evaluate the SAFE-D principles (10–20 minutes)

- It's essential to record and be transparent about how decisions were made regarding the prioritisation of principles.
- Use the prompt questions below to guide discussion.
- Use the post-it notes to document your discussion points.
- The aim of the exercise is to number the principles from high (1) through to low (5) priority.

Prompts for discussion

- Identify principles that are central to the project and will likely require ongoing attention.
- Consider areas that may need extra focus or investment to ensure ethical standards are met.
- Determine which principles are critical for the project's success and align closely with its objectives.
- Identify potential areas where principles may clash, necessitating careful consideration and resolution.

Importance of transparency

Ensure that this process is communicated clearly to all stakeholders, as it fosters trust and understanding regarding the ethical considerations guiding the project. This transparency can enhance stakeholder engagement and support collaborative efforts in addressing ethical challenges.

Workshop exercise: SAFE-D specification

Activity overview

The goal of this step is to refine each SAFE-D principle to reflect the specific nuances of your project context. This will help make the principles actionable and ensure ethical goals are met.

Steps:

1) Group discussion: SAFE-D principles

- As a group, use the previous activities as prompts to help you answer the following question: what does the [x] principle mean to you in the context of our project?
- Make sure everyone has the opportunity to contribute their thoughts. This will help surface a variety of perspectives and formulate a well-rounded definition for the principle.

2) Synthesise responses

- Group contributions into key themes.
- As a team, agree on a short objective or goal statement for each of the SAFE-D principles and document this on the board.

3) Defining the principles

- By narrowing the scope of each principle, you will make it more relevant and actionable in the context of your project. It will also allow your decision-making to be guided by tangible goals and objectives.

SAFE-D litmus test

SAFE-D litmus test

After finishing the SAFE-D principles workshop, you are ready to conduct a SAFE-D litmus test. This assessment will help identify key risk areas and ensure that ethical considerations have been adequately addressed during the design phase. You should leverage the insights and discussions from your SAFE-D principles workshop to inform your responses in the assessment.

Assign team members

Consider nominating different team members to collaborate on various sections of the assessment. This allows for diverse input and thorough exploration of each principle.

Complete the questionnaire

The SAFE-D litmus test consists of a series of questions organised around the SAFE-D principles. Each principle has 3–4 questions that require a brief answer. We encourage you to provide any additional justification or evidence. This can be used in your discussions with the data ethics team.

Justification and evidence

Ensure that for each answer, you provide sufficient justification or evidence. This could include references to workshop insights, project documentation, or stakeholder feedback.

Mitigate risks early

Remember that the purpose of this assessment is to help identify and mitigate risks early in the project lifecycle. We understand you may not have all the answers at this early stage, so just document as much as you can.

Document the assessment

After completing the assessment, compile and document the responses. This will serve as a useful reference for future stages of the project and facilitate transparency with stakeholders.

Sustainability

Question number	Core attribute(s)	Question
1.1a	Reliability	How does the project align to the long-term strategic objectives and the MoJ data strategy?
1.1b	Safety, robustness	How will the model be free of any long-term societal risks (e.g. worsening of existing inequalities or justice outcomes)?
1.1c	Security	Are there clearly defined roles and responsibilities within the project, especially between the technical development team, business team and governance team?
1.1d	Accuracy and performance	What are the safe and reliable outputs that fall within pre-defined thresholds (e.g. in accuracy, consistency and security) as part of the long-term delivery of this project?

Accountability

Question number	Core attribute(s)	Question
1.2a	Traceability, clear data provenance and lineage	How will you ensure that this project will be free from automated components that involve decisions that could negatively impact an individual (as either a member of the public or MoJ staff)?
1.2b	Answerability	Will there be frequent human oversight in the process/project?
1.2c	Answerability, auditability	Will there be an effective contest process or other means to ensure the ability of stakeholders beyond the development team to challenge the output?

Fairness

Question number	Core attribute(s)	Question
1.3a	Non-discrimination	Will the project be free from potential use or misuse of the data that could result in negative discrimination against people on the basis of any protected characteristics, especially those in vulnerable circumstances?
1.3b	Bias mitigation	Will the proposed system work the same for all groups of people? If not, why?
1.3c	Equality	How will you address any potential feedback loops that may worsen existing inequalities?
1.3d	Diversity and inclusivity	How will you involve diverse stakeholder perspectives in the project in a meaningful and impactful way?

Explainability

Question number	Core attribute(s)	Question
1.4a	Interpretability	How will the use of technology and/or data be effectively articulated to key stakeholders?
1.4b	Accessible rationale explanation	Will there be a clear way for an individual to understand and obtain an explanation regarding how data is being used in this project, either for public scrutiny or research?
1.4c	Responsible model selection	Is compliance with data protection principles for transparency enforced and demonstrated?

Data responsibility

Question number	Core attribute(s)	Question
1.5a	Legal and organisational compliance	Does data processing adhere to the principles of data protection? (lawfulness; fairness and transparency; purpose limitation; data minimisation; accuracy; storage limitation; integrity and confidentiality; accountability)
1.5b	Adequacy of quality and quantity, source integrity	Will any proposed data sharing, linking or enrichment be proportionate to, and justifiable against, the project objectives?
1.5c	Responsible data management	Will there be clear governance relating to which practices are acceptable and which are not? Who will enforce this?

Contact us:

DataEthics@justice.gov.uk