Ministry of Justice

Ministry of Justice
AI and data science ethics framework

Deployment phase activity booklet

The Alan Turing Institute

5

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

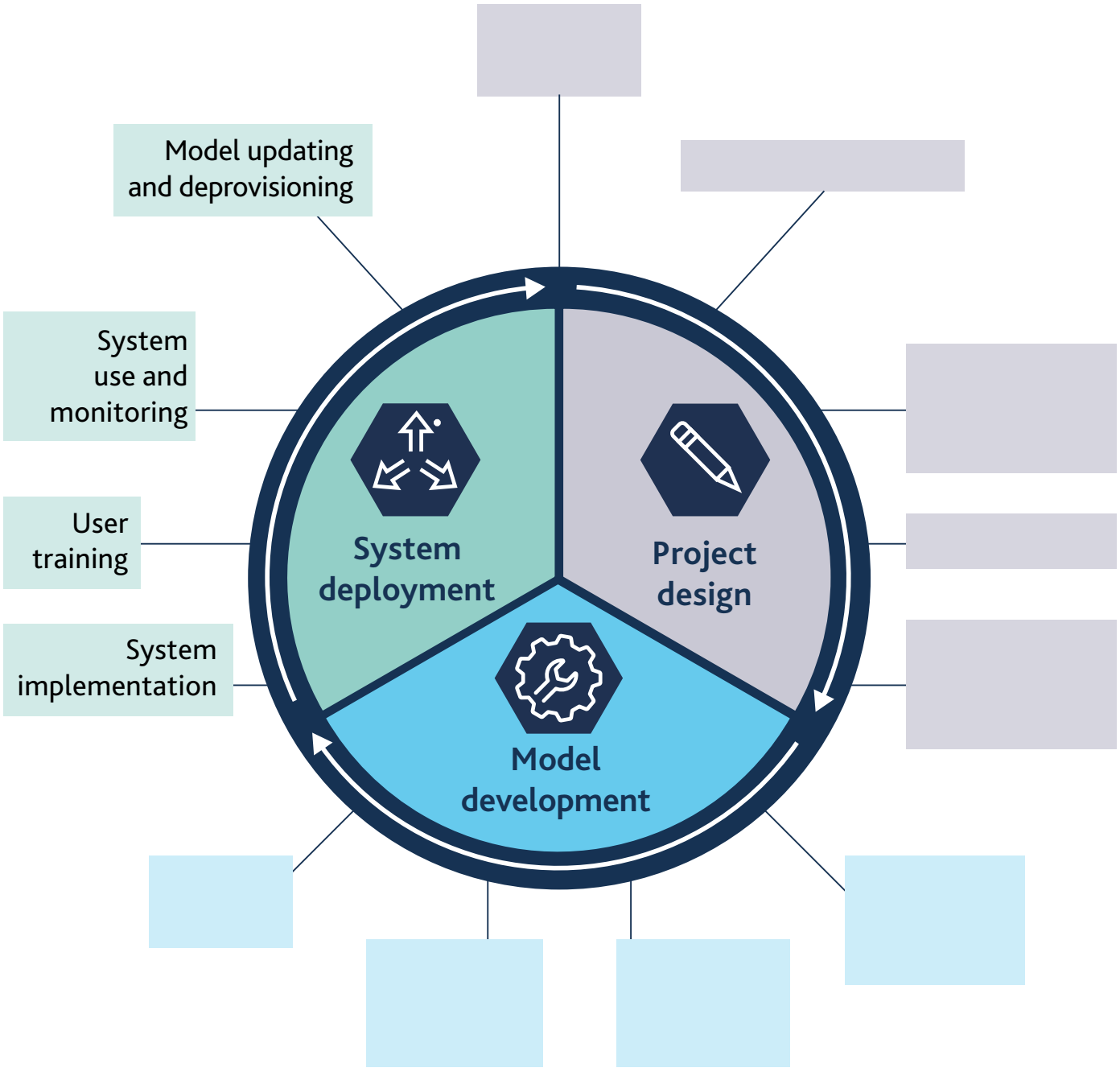
5 Deployment phase activity booklet

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

System deployment phase

System deployment

Focuses on the safe implementation and use of the system, including ongoing monitoring and updates.



System deployment phase

Lower-level lifecycle stage	Ethical significance
System design and implementation This process involves putting a model into action and integrating it into an operational setting, allowing users to interact with it.	The effectiveness of the model depends on how well the system is implemented. There are two key types of implementation: • **technical implementation:** this involves creating the hardware and software needed to support the model (like servers and interfaces). It's crucial to ensure the system is secure, efficient, and accessible • **social or organisational implementation:** this focuses on how the technical system fits within wider social and organisational practices, ensuring that users are informed and the system aligns with existing practices

Lower-level lifecycle stage	Ethical significance
User training This refers to any support or skill-building provided to individuals or groups who need to operate a data-driven technology, especially in safety-critical situations.	User training is usually not done by the same team that created the system. While developers might provide documentation, it's often not enough; more formal training sessions may be needed, especially for complex systems. Poor training can lead to issues like algorithmic aversion, where users distrust a reliable system or trust an unreliable one.
System use and monitoring Typically, metrics and evaluation methods are used to track the system's performance, ensuring it maintains or improves upon its initial performance.	Depending on its design, an AI system can allow for continuous feedback and learning when deployed in a physical or virtual environment (like robotic systems using reinforcement learning or digital twins linked to real counterparts). Because machine learning models and AI systems can behave dynamically and unpredictably, ongoing monitoring and feedback are crucial to detect issues like model drift that could harm individuals or groups.

Lower-level lifecycle stage	Ethical significance
Model updating and decommissioning If monitoring a model or system reveals vulnerabilities or poor performance, it may be necessary to update the model through retraining or to remove the system if it no longer works effectively.	An algorithmic model that changes over time might need updates or removal from use. While improvements to the system's infrastructure (like speed or security) are important, the key focus should be on the model itself (such as its parameters and features). A significant issue to consider is model drift, which can happen when the data used to train the model changes (like fluctuating house prices) or when the meaning of features shifts due to changes in societal practices or norms.

Development phase SAFE-D reflection

Workshop exercise: SAFE-D deployment checkpoint

Activity overview

Now that you're in the system deployment phase, it's important to reflect on the SAFE-D principles could be impacted as you get ready to deploy your system. Use the following steps and prompts to guide your discussion and documentation.

Steps:

1) Review previous work

- Look back at the answers and action strategies you developed during previous activities to refresh your memory.

2) Use the Miro board

- Track any changes, progress, and new questions that have arisen in the deployment phase.
- You can create a new section on the board, or add post-its in a different colour to your original answers.

3) Evaluate SAFE-D goals and objectives

- Revisit the objectives you set during the SAFE-D specification exercise. Assess whether these strategies are working effectively or if they need adjustments.

Prompt questions for discussion

- What changes have occurred since the development phase that impact the SAFE-D principles?
- How has the project progressed in relation to the ethical principles? Are there any areas where you feel you have made significant progress?
- Have any new questions or concerns emerged regarding the SAFE-D principles as the project develops?
- Are the strategies you implemented in the design phase proving effective? What challenges have you encountered?
- Have you received any feedback from stakeholders that may influence your assessment of the SAFE-D principles?

Documentation

Make sure to document your reflections and discussions in a clear format. This will not only provide a record of your progress but also facilitate ongoing communication among team members and stakeholders.

Importance of reflection

Reflecting on the SAFE-D principles at this stage is crucial for maintaining ethical integrity throughout the project. By actively assessing your project's alignment with these principles, you can identify potential risks and ensure that ethical considerations are prioritised as you move forward.

Deployment phase questionnaire

Deployment phase questionnaire

The deployment phase questionnaire is the next formal checkpoint in the ethics process.

Steps:

Use previous insights
- Leverage the insights and discussions from your previous SAFE-D workshops and assessments 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 deployment phase questionnaire consists of a series of questions organised around the SAFE-D principles. Each principle has 3–4 questions that require a simple Yes/No answer, along with justification or evidence.

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.

Track changes and decisions
- Remember that the purpose of this assessment is not to hinder innovation but to help identify and mitigate risks early in the project lifecycle.

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
3.1a	Robustness	Is the team confident in its ability to monitor the model's performance in case of any external changes?
3.1b	Safety	Are there monitoring mechanisms in place for all target objectives, including ethical objectives?
3.1c	Security, accuracy and performance	Is the system designed in a way that is at a suitable level of security, performance and accessibility for the intended usage?

Accountability

Question number	Core attribute(s)	Question
3.2a	Traceability	Have any gaps in the human feedback mechanism for errors been recorded? (e.g. can high cost errors missed by humans be caught)?
3.2b	Answerability	Is there an accountable party in charge of monitoring for any changes in key performance indicators?
3.2b	Auditability	Is there an escalation and intervention process in place if key indicators fall outside of a pre-defined threshold?
3.2d	Accessibility	Is there appropriate support or upskilling available where required for users of the system?

Fairness

Question number	Core attribute(s)	Question
3.3a	Bias mitigation	If there is a feedback loop, is it free from reinforcing any existing biases? Is there any user interaction with the output?
3.3b	Equality	Is the model being monitored for any shifts in demographic characteristics?
3.3c	Non-discrimination	Is performance variation between demographic groups being tracked?

Explainability

Question number	Core attribute(s)	Question
3.4a	Interpretability	After deployment, can explanations be produced as required by both global and local levels?
3.4b	Responsible model selection	Are the explanations stable over model re-training cycles?
3.4c	Implementation and user training	Are explanations being reviewed before sharing with appropriate stakeholders?

Data responsibility

Question number	Core attribute(s)	Question
3.5a	Responsible data management, legal and organisational compliance	Are appropriate measures in place to protect personal and sensitive data in live environments?
3.5b	Adequacy of quantity and quality	Are data quality issues in any incoming datasets being monitoring for potential impact on the model?
3.5c	Legal and organisational compliance	Are there regular refreshes to legal and compliance assessments?
3.5d	Responsible data management	Is there a provision for taking the model offline or de-commissioning it, and a fall-back plan if the model is failing to meet its requirements?

Contact us:

DataEthics@justice.gov.uk