Ministry of Justice

Ministry of Justice
AI and data science ethics framework

Development phase activity booklet

The Alan Turing Institute

4

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

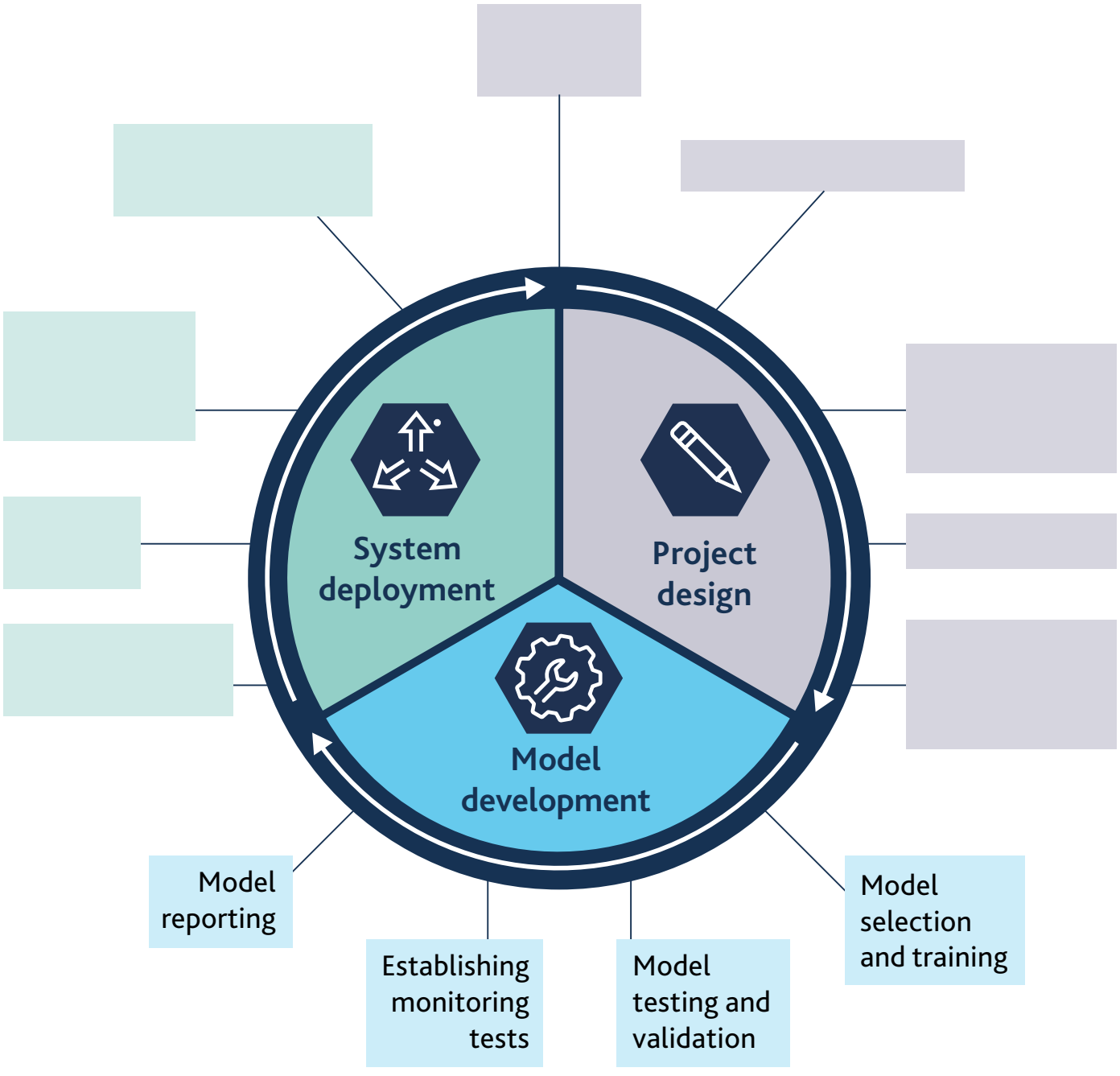
5 Deployment phase activity booklet

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

Model development phase

Model development

Covers technical tasks such as training, testing, and validating the machine learning model to ensure it's suitable for its purpose.



Model development phase

Lower-level lifecycle stage	Ethical significance
Model selection and training This stage involves choosing one or more algorithms for training a model. Factors influencing this decision include: • access to computational resources • predictive performance of the model • characteristics of the data	While there are technical reasons for responsibly selecting and training a model, a key aspect is the model's interpretability and explainability. Generally, more complex models (like convolutional neural networks) are harder to interpret compared to simpler ones (like linear regression). The choice of algorithm should align with the model's intended use and the system it will be part of.
Model testing and validation Model testing and validation involves evaluating a model using various metrics, including its accuracy on new data that wasn't part of the original training set.	When a dataset is divided into training and testing sets (internal validation) or when a model is assessed with completely new data (external validation), there are opportunities to evaluate more than just performance. Testing a model's generalisability to new contexts helps ensure it is both sustainable and fair, meaning it should maintain similar accuracy and performance levels even when validated externally.

Lower-level lifecycle stage	Ethical significance
Establishing monitoring tests Effective monitoring of a model's performance in its runtime environment relies on setting up metrics to check if the model is functioning within its intended parameters.	While tests typically measure accuracy and global interpretability, it's also important to assess system-level performance, such as efficiency and resource usage. The team responsible for the model will need their expertise for this stage, along with input from systems engineers, software engineers, and end users. This stage is crucial for establishing collective responsibility throughout the project's lifecycle. It occurs at an important point in the project and relies on clear and accessible communication among all team members.
Model documentation This task involves documenting both formal and informal aspects of the model and its development process, including: • data sources and summary statistics • the model used • evaluation metrics	Clear and accessible documentation is vital for responsible project governance for several reasons: • it ensures results can be reproduced and supports values like public accessibility in open research • it promotes accountability and transparency in decision-making • it helps individuals affected by the technology seek redress for any harms that may occur during its design, development, or deployment

Development phase SAFE-D reflection

Workshop exercise: SAFE-D development checkpoint

Activity overview

Now that you're in the project development phase, it's important to reflect on the SAFE-D principles and assess how your project aligns with them. 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 the design phase workshop activities to refresh your memory.

2) Use the Miro board

- Track any changes, progress, and new questions that have arisen in the development 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 design 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.

Development phase questionnaire

Development phase questionnaire

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

Steps:

Use previous insights
- Leverage the insights and discussions from your SAFE-D workshops and litmus test 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 development 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
2.1a	Safety	Is the use case free from any risks to fundamental rights and freedoms?
2.1b	Security	Are there sufficient security measures and access controls in place?
2.1c	Robustness	Have you tested the model for robustness to data drifts and adversarial attacks? Please answer N/A if this is a one-off analysis or will not be used with new data
2.1d	Reliability	Are the outputs of the model consistent and aligned to expectations?
2.1e	Accuracy and performance	Is the performance within acceptable thresholds?

Accountability

Question number	Core attribute(s)	Question
2.2a	Answerability	Are the escalation pathways and approval pathways clear for this project? i.e. who is approving this model?
2.2b	Answerability	Are the roles and responsibilities of the project team and wider stakeholders clearly defined?
2.2c	Clear data provenance and lineage	Have you liaised with the data owner(s) on any limitations, quality issues or bias in the data set?
2.2d	Clear data provenance and lineage	Do you know who the data owners are and have clear data documentation of all features?
2.2e	Accessibility	Is the documentation on the model accessible and easy to understand?
2.2f	Auditability	Have you checked that any engineering or selected features are appropriate with the relevant domain experts?

Question number	Core attribute(s)	Question
2.2g	Auditability	Have you signed off the performance and error rates with the appropriate stakeholder(s)?
2.2h	Auditability	Are all decisions made documented with justifications? (For example, any data input methods and why they were chosen)
2.2i	Traceability	Are all relevant actions and analyses traceable and reproducible?

Fairness

Question number	Core attribute(s)	Question
2.3a	Non-discrimination	Is the project data collection strategy returning a representative sample of the population?
2.3b	Non-discrimination	Is there a sufficient sample in any subgroup of interest for this analysis?
2.3c	Diversity and inclusivity	Have you considered including diverse voices and opinions in the design and development process, e.g. through engaging with civil society and affected communities?
2.3d	Bias mitigation	Have any features added by developers that could be affected by personal judgement been documented and mitigated?
2.3e	Bias mitigation	Have you identified and mitigated potential proxies in your data, i.e., features in your data that may be associated with unjustifiable sources of differences in outcome, e.g., postcode with race, income with gender?

Question number	Core attribute(s)	Question
2.3f	Bias mitigation	Has the model been calibrated across different metrics for each group and mechanism?
2.3g	Equality	Is the model free of all potential feedback loops that may worsen existing inequalities?

Explainability

Question number	Core attribute(s)	Question
2.4a	Implementation and user training	Have you considered all the stakeholders that might need an explanation, either global (model level) or local explanation (individual level)?
2.4b	Implementation and user training	Are the users sufficiently informed and knowledgeable to understand the explanation?
2.4c	Interpretability	Can you clearly communicate how each feature was calculated or collected?
2.4d	Interpretability	Are you able to create clear global and local explanations for each prediction?
2.4e	Interpretability	Has the explainability of the model remained consistent as the size of the data set and the number of features has increased?
2.4f	Responsible model selection	Where a complex model is selected, is the rationale documented as to why a simpler model cannot be used?

Question number	Core attribute(s)	Question
2.4g	Accessible rationale explanation	Is the explanation accessible and easy to understand for each stakeholder?
2.4h	Reproducibility	Is the explanation reproducible?

Data responsibility

Question number	Core attribute(s)	Question
2.5a	Timeliness and recency	If using historical data, are all data sets being used timely and applicable for deriving future insights?
2.5b	Legal and organisational compliance	Have you ensured adherence to existing laws and regulations, including data protection law, privacy standards, and public sector duties?
2.5c	Legal and organisational compliance	Have all the potential risks, assumptions, and limitations of this use-case been reviewed and signed-off?
2.5d	Source integrity and measurement accuracy	Are any of the recorded features free from human judgement? If no, have they been recorded and mitigated?
2.5e	Responsible data management	Are there clear policies in place for data storage, sharing, documentation, and processing?

Question number	Core attribute(s)	Question
2.5f	Responsible data management	If any data sets are linked and shared, have data protection principles been considered?
2.5g	Adequacy of quality and quantity	Do the available data set(s) have the adequate sample size, representativeness, and contextual relevance for the use case at hand? If not, are the limitations documented?

Contact us:

DataEthics@justice.gov.uk