Ministry of Justice

Ministry of Justice
AI and data science ethics framework

SAFE-D principles booklet

The Alan Turing Institute

2

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 are ethical principles?

Ethical principles are not to be confused with rules.

Ethical principles, like SAFE-D, should be thought of as **goals** and **objectives** for ethical discussion and decision-making.

They guide and lead choices and actions, but do not restrict or limit them. They are not **absolute**.

Rules are **fixed** and place **harder limits** on choice and action

Some examples of rules are legal requirements and organisational policies.

SAFE-D is a set of principles rather than rules.

What are the SAFE-D principles?

SAFE-D stands for:

Sustainability
Accountability
Fairness
Explainability
Data responsibility

Each SAFE-D principle includes core attributes that explain their role in designing, developing, and deploying data-driven technologies in the criminal justice system.

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

Before diving into the SAFE-D principles, it is important to understand how to move from principles to practice. This is covered in the next section.

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.

How do I use the SAFE-D principles?

The SAFE-D principles are goals to consider throughout the project lifecycle. They serve as a foundation for reflection and discussion about ethics and provide a shared language for everyone involved.

However, these principles alone do not ensure ethical actions; they don't dictate specific rules or actions to follow. Instead, they initiate a process of ethical assurance.

To turn principles into practice, follow this method:

- **Identify:** determine which principles apply to your project to highlight their relevance.
- **Weigh:** assess which SAFE-D principles are most important based on how they align with project goals.
- **Specify:** use the core attributes to clarify what each principle means in the context of your project.
- **Revise:** repeat the process as needed, incorporating ongoing engagement with stakeholders whenever possible.
- **Implement:** carry out the actions identified by applying the SAFE-D principles to your project activities. Continuously monitor and evaluate their effectiveness.

Sustainability principle

Sustainability principle

Sustainability can mean many things.

From a technical perspective, sustainability requires the outputs of a project to be safe, secure, robust, and reliable.

In criminal justice, safety is particularly important in situations where the effects of AI systems have implications for users of the justice system and members of society in general.

In the context of responsible data science and AI, sustainability also requires a project's practices to be informed by ongoing consideration of the risk of exposing individuals to harms even well after the system has been deployed and the project completed – a long-term (or sustainable) safety.

Sustainability: core attributes

Safety

Safety involves more than just how a system operates; it also includes its long-term use and impact, as well as the resources needed to keep it safe over time. Since some safety issues might not be clear to developers, involving users and stakeholders in the design and evaluation can help address potential impacts on human rights and freedoms. In the criminal justice context, ensuring the safety of users and the public is especially important.

Security

A secure system protects its information by preventing unauthorised changes or damage to its components. It must also remain functional and accessible to authorised users while keeping confidential information safe, even in challenging conditions. In the criminal justice context, there are specific security needs, especially regarding the risk of harm to vulnerable individuals from unauthorised access to information.

Robustness

The goal of robustness is for an AI system to operate reliably and accurately, even in difficult or uncertain conditions. These conditions can involve adversarial attacks, user errors, or issues with how the system learns from its environment. Proper documentation is crucial to ensure that data, models, and systems remain robust, especially when personnel changes occur.

Reliability

The goal of reliability is for an AI system to perform exactly as its designers intended. A reliable system follows the specifications it was programmed with. Reliability measures consistency and helps build confidence in the system's safety. Additionally, having access to relevant and high-quality data is crucial for maintaining accuracy and continuity.

Accuracy and performance

The accuracy of a model is the percentage of cases where it gives a correct output. This can also be described as the error rate, or the percentage of incorrect outputs. Setting a reasonable accuracy level might involve adjusting how accuracy is measured. For example, if some errors are more costly than others, a cost metric can be added to weigh different types of errors against each other. It's important to communicate and address margins of error. Additionally, accuracy and performance can vary with scale, so teams should be prepared for these changes.

Accountability principle

Accountability principle

Accountability involves being transparent about processes and outcomes, along with clear communication that helps stakeholders understand how a project was carried out and why decisions were made.

It also requires defining clear roles and responsibilities to ensure responsible governance. Having a single point of contact or ownership for a project helps maintain accountability.

In coding, using formal version control practices is essential for accountability. In the criminal justice context, teams may be familiar with analytical quality assurance standards for ensuring data quality.

Accountability principle: core attributes

Traceability

Traceability is the process of documenting all stages of the data lifecycle so that it is clear and easy to understand. This documentation should be accessible not only to the project team but also to individuals who use or are affected by the system.

Answerability

Answerability relies on a clear chain of human responsibility, helping to identify who is accountable for automated outcomes. There should be a designated point of contact responsible for the project. Clear communication and transparency with stakeholders are crucial for effective reporting. It's also important to have a handover procedure when responsibilities change.

Auditability

Auditability addresses how designers and implementers of systems are held accountable for their actions and decisions during the project lifecycle. It involves showing evidence of responsible design and usage practices, as well as justifying the outcomes achieved.

Accessibility

Accessibility means making information about the design, development, and deployment processes of a system easy to understand for everyone. This includes using clear language and suitable delivery methods. It's important to consider various audiences and adapt the information about the system to meet their needs.

Clear data provenance and lineage

Clear data provenance and lineage involve keeping accessible records that explain how data was collected and how it has been used and changed throughout the processes of pre-processing, modelling, training, testing, and deployment. This can include using version control or tracking changes to preserve different versions over time, showing how certain statistics were produced and the review processes involved.

Fairness principle

Fairness principle

To determine if the design, development, and deployment of data-driven technologies is fair, it's essential to recognise all rights and interests that may be impacted by the system.

From a legal or technical standpoint, outcomes should **avoid discrimination** (like profiling based on protected traits) and should **not negatively impact social equality**.

Additionally, consider broader justice issues, such as whether the technology is seen as disproportionately harmful by affected communities.

In criminal justice, **fairness requires balancing the rights and interests of everyone involved** – victims, defendants, prisoners, families, and staff – while also ensuring public safety and protecting vulnerable individuals.

While statistical measures can aid in this, social awareness and stakeholder input are also crucial.

Fairness principle: core attributes

Bias mitigation

While it's impossible to completely eliminate bias, effective processes can reduce its negative effects. These processes help address systematic issues, distortions, or unfair outcomes caused by governance problems or a lack of consideration for historical discrimination.

Diversity and inclusivity

A key part of fairness-aware design is including diverse voices and opinions in the development process by collaborating with a range of stakeholders. This involves ensuring that values like civic participation, inclusion, and diversity are considered when defining the project's purpose and goals. Consulting internal stakeholders helps enhance the project's openness and inclusivity. It's also important to engage with external stakeholders – such as civil society groups, NGOs, and affected communities – to foster collaboration and ensure relevant voices are heard.

Non-discrimination

A system or model should not create situations where members of protected groups are treated unfairly or less favourably than other groups based on their protected characteristics.

Equality

The outcome of a system should ensure that every individual has equal rights and freedoms. This includes equal treatment under the law, equal access to justice, and equal opportunities for any benefits the system provides.

Explainability principle

Explainability principle

Explainability is the ability of a technology, like an AI system, to **help users understand how it works**. It differs from interpretability, which focuses on how easy it is to grasp the system's inner workings.

For example, some algorithms, like decision trees, are easier to understand than others, like complex neural networks. However, understanding how an algorithm operates also depends on the overall system it is part of, and the expertise of the people using it. Sometimes, even the creators of a model may struggle to fully understand it, which may require additional tools like dashboards to help.

Additionally, striving for explainability can sometimes conflict with other important goals, such as keeping data confidential or ensuring safety, so it's essential to find a balance.

Explainability principle: core attributes

Interpretability

Interpretability is about understanding how and why a model made a particular decision in a specific situation. It involves grasping the reasoning behind its actions. In teams with diverse expertise, it's possible that some members may not fully understand the model, so it's important to seek input from various experts to fill any knowledge gaps.

Responsible model selection

Systems should be clear and easy to understand in their specific areas of use. When using opaque algorithms, it's crucial to consider the potential risks to individuals' rights and well-being. More interpretable models may be necessary to mitigate these risks. Implementing responsible explanation methods, like simpler surrogate models or sensitivity analyses, is essential when opting for opaque models. Organisations need the resources and capacity to provide these supplementary explanations to ensure transparency and accountability.

Accessible rationale explanation

Decisions, especially automated ones, should be explained clearly and without technical language. Where possible, this includes mitigating jargon and/or providing a glossary of terms to remove any assumptions regarding definitions. Ongoing communication of decisions is also important.

Implementation and user training

Training users to use the AI system should include:

a) basic information about machine learning

b) an explanation of the system's limitations

c) awareness of AI biases, like decision-automation bias or automation-distrust bias

d) a focus on how these systems can assist human decision-making rather than replace it

Reproducibility

Reproducibility means others should be able to follow your project steps to achieve the same results. If needed, they should be able to replicate the outcomes by using the same methods.

Data responsibility principle

Data responsibility principle

The principle of data responsibility emphasises the ethical aspects of the data used in AI and machine learning projects.

Data quality refers to the basic characteristics of data, such as:

a) being relevant and representative of the intended context

b) being balanced and complete to reflect the data collection process

c) being accurate and up-to-date as needed for the project

Data integrity involves how data is managed over the project's lifecycle. It requires:

a) keeping initial records (like logs and research statements)

b) ensuring consistent and verifiable data processing during development

c) creating accessible records at the end of the project

Other important factors include **legal obligations**, like following data protection laws and **human rights regulations**, as well as rules related to law enforcement data.

Data responsibility principle: core attributes

Responsible data management

Responsible data management involves training the team to handle data securely and responsibly. This includes identifying potential risks and assigning roles for addressing them. The team should discuss and document policies on data storage and how to share results with the public, involving stakeholders in these discussions.

Adequacy of quantity and quality

This attribute looks at whether the available data is sufficient to solve the problem based on the system's use case, domain, function, and purpose. It should consider factors like sample size, how well the data represents the situation, its relevance to the context, and the availability of important features.

Source integrity and measurement accuracy

Effective bias mitigation starts right from data extraction and collection. Both the sources and measurement tools can introduce bias into the dataset. If biased human decisions – like unfair scoring or evaluations – are included in the training data, they will influence the model and perpetuate that bias in the system's outputs. To prevent harm and ensure data integrity, it's essential to use reliable and unbiased sources and sound collection methods.

Timeliness and recency

Using outdated data can negatively impact how well a trained model performs because it may not reflect current conditions. If changes in data reflect shifts in social relationships or group dynamics, this can lead to inaccuracies and introduce bias into the system. To avoid discriminatory outcomes, it's important to regularly check that all data elements are up-to-date and relevant.

Legal and organisational compliance

Project teams must follow laws and regulations related to data responsibility, including data protection and privacy standards. In criminal justice, additional analysis may be needed to identify what qualifies as 'law-enforcement' data, which has different legal requirements. Organisational policies may restrict data access for certain parties (like offenders) while allowing broader access for others (like judges and attorneys). Teams might need to seek outside guidance to ensure their data practices meet all internal and external obligations.

Contact us:

DataEthics@justice.gov.uk