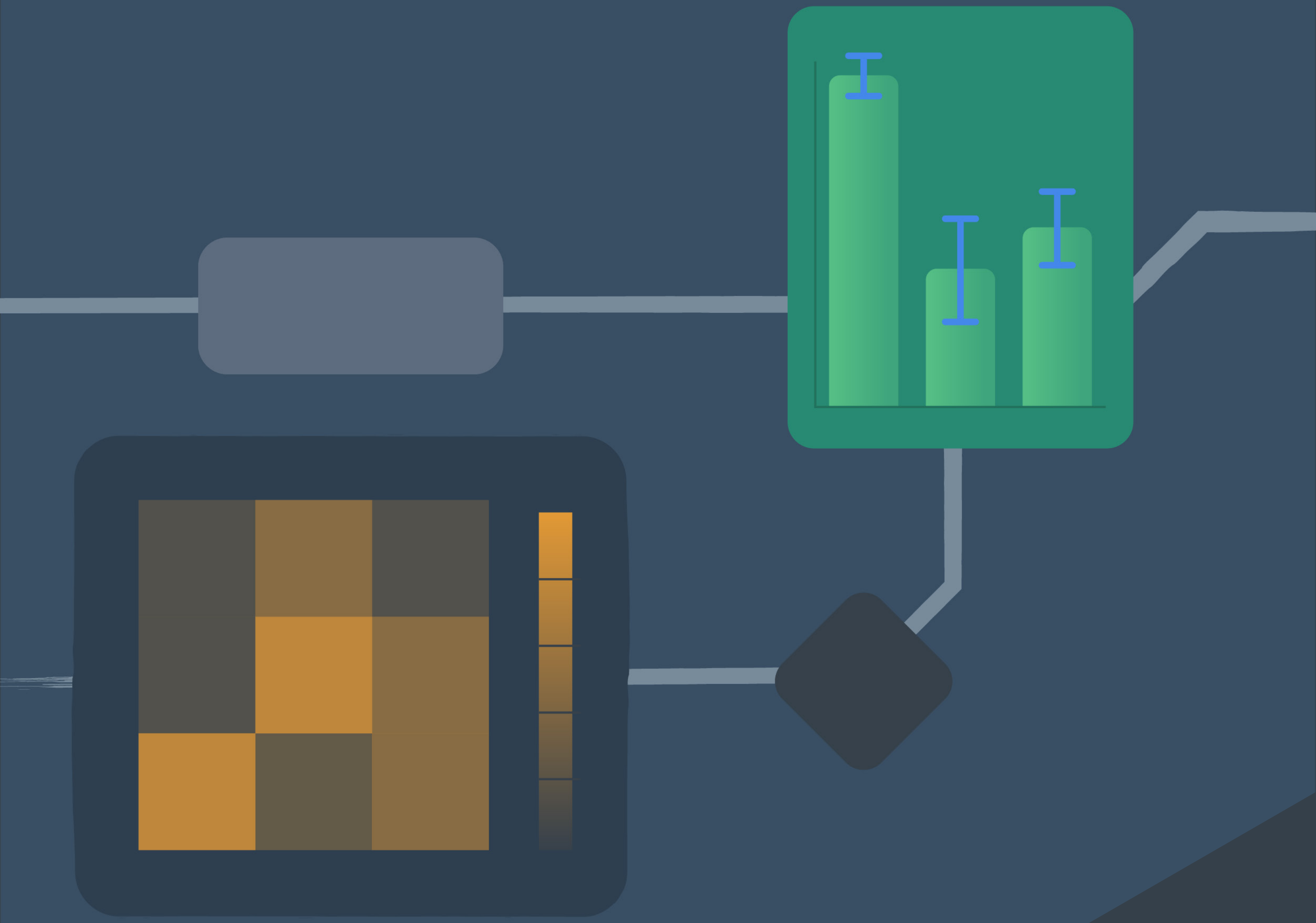
AI Ethics and Governance in Practice Programme

AI Safety in Practice



Facilitator Workbook

Annotated to support facilitators in delivering the accompanying activities.

The Alan Turing Institute

Acknowledgements

This workbook was written by David Leslie, Cami Rincón, Morgan Briggs, Antonella Maia Perini, Smera Jayadeva, Ann Borda, SJ Bennett, Christopher Burr, and Claudia Fischer.

The creation of this workbook would not have been possible without the support and efforts of various partners and collaborators. As ever, all members of our brilliant team of researchers in the Ethics Theme of the Public Policy Programme at The Alan Turing Institute have been crucial and inimitable supports of this project from its inception several years ago, as have our Public Policy Programme Co-Directors, Helen Margetts and Cosmina Dorobantu. We are deeply thankful to Conor Rigby, who led the design of this workbook and provided extraordinary feedback across its iterations. We also want to acknowledge Johnny Lighthands, who created various illustrations for this document, and Alex Krook and John Gilbert, whose input and insights helped get the workbook over the finish line. Special thanks must be given to the Digital Office for Scottish Local Government, The National Institute for Health and Care Excellence (NICE), and Carolyn Ashurst, Kamalaruban Parameswaran, Ruth Drysdale, Ryan Burnell, Elhassan Mohamed (The Alan Turing Institute) for helping us test the activities and review the content included in this workbook. Lastly, we want to thank Madeleine Waller (Kings College London) for her meticulous peer review and timely feedback, which greatly enriched this document.

This work was supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/W006022/1, particularly the Public Policy Programme theme within that grant & The Alan Turing Institute; Towards Turing 2.0 under the EPSRC Grant EP/W037211/1 & The Alan Turing Institute; and the Ecosystem Leadership Award under the EPSRC Grant EP/X03870X/1 & The Alan Turing Institute.

Cite this work as: Leslie, D., Rincón, C., Briggs, M., Perini, A., Jayadeva, S., Borda, A., Bennett, SJ., Burr, C., and Fischer, C. (2024). *AI Safety in Practice.* The Alan Turing Institute.

Contents

About the Workbook Series

Key Concepts

Activities

Further Readings

About the AI Ethics and Governance in Practice Workbook Series

Who We Are

The Public Policy Programme at The Alan Turing Institute was set up in May 2018 with the aim of developing research, tools, and techniques that help governments innovate with data-intensive technologies and improve the quality of people's lives. We work alongside policymakers to explore how data science and artificial intelligence can inform public policy and improve the provision of public services. We believe that governments can reap the benefits of these technologies only if they make considerations of ethics and safety a first priority.

Origins of the Workbook Series

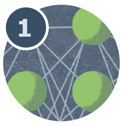
In 2019, The Alan Turing Institute's Public Policy Programme, in collaboration with the UK's Office for Artificial Intelligence and the Government Digital Service, published the UK Government's official Public Sector Guidance on AI Ethics and Safety. This document provides end-to-end guidance on how to apply principles of AI ethics and safety to the design, development, and implementation of algorithmic systems in the public sector. It provides a governance framework designed to assist AI project teams in ensuring that the AI technologies they build, procure, or use are ethical, safe, and responsible.

In 2021, the UK's National AI Strategy recommended as a 'key action' the update and expansion of this original guidance. From 2021 to 2023, with the support of funding from the Office for AI and the Engineering and Physical Sciences Research Council as well as with the assistance of several public sector bodies, we undertook this updating and expansion. The result is the AI Ethics and Governance in Practice Programme, a bespoke series of eight workbooks and a digital platform designed to equip the public sector with tools, training, and support for adopting what we call a Process-Based Governance (PBG) Framework to carry out projects in line with state-of-the-art practices in responsible and trustworthy AI innovation.

About the Workbooks

The AI Ethics and Governance in Practice Programme curriculum is composed of a series of eight workbooks. Each of the workbooks in the series covers how to implement a key component of the PBG Framework. These include Sustainability, Safety, Accountability, Fairness, Explainability, and Data Stewardship. Each of the workbooks also focuses on a specific domain, so that case studies can be used to promote ethical reflection and animate the Key Concepts.

Programme Curriculum: AI Ethics and Governance in Practice Workbook Series



1

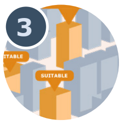
AI Ethics and Governance in Practice: An Introduction

Multiple Domains

2

AI Sustainability in Practice Part One

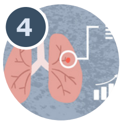
AI in Urban Planning



3

AI Sustainability in Practice Part Two

AI in Urban Planning



4

AI Fairness in Practice

AI in Healthcare

5

Responsible Data Stewardship in Practice

AI in Policing and Criminal Justice

6

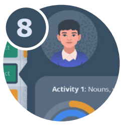
AI Safety in Practice

AI in Transport

7

AI Explainability in Practice

AI in Social Care



8

AI Accountability in Practice

AI in Education

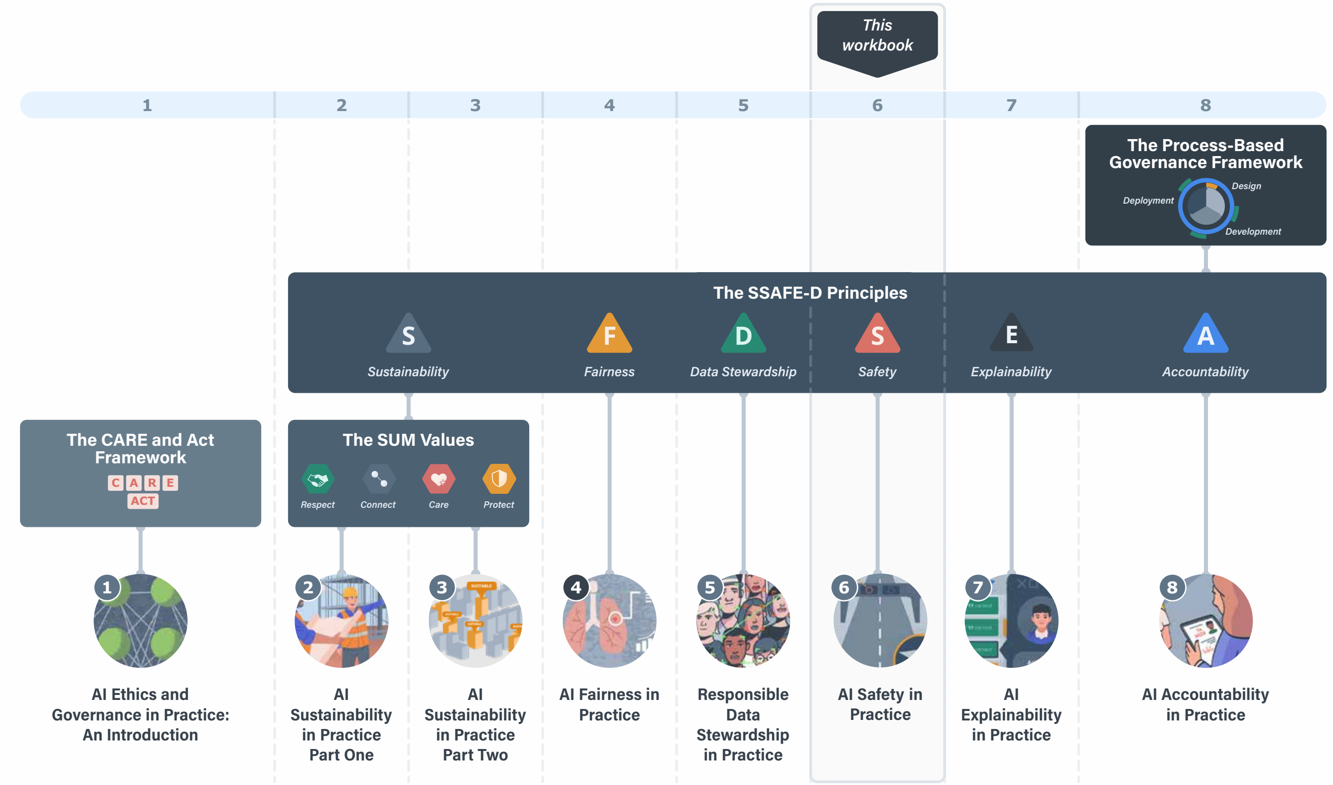
Explore the full curriculum and additional resources on the AI Ethics and Governance in Practice Platform at aiethics.turing.ac.uk.

Taken together, the workbooks are intended to provide public sector bodies with the skills required for putting AI ethics and governance principles into practice through the full implementation of the guidance. To this end, they contain activities with instructions for either facilitating or participating in capacity-building workshops.

Please note, these workbooks are living documents that will evolve and improve with input from users, affected stakeholders, and interested parties. We value your participation. Please share feedback with us at aiethics@turing.ac.uk.

Programme Roadmap

The graphic below visualises this workbook in context alongside key frameworks, values and principles discussed within this programme. For more information on how these elements build upon one another, refer to AI Ethics and Governance in Practice: An Introduction.


This workbook
1
2
3
4
5
6
7
8
The Process-Based Governance Framework
Design
Deployment
Development
The SSAFE-D Principles
S
Sustainability
F
Fairness
D
Data Stewardship
S
Safety
E
Explainability
A
Accountability
The CARE and Act Framework
C A R E
ACT
The SUM Values
Respect
Connect
Care
Protect
1
AI Ethics and Governance in Practice: An Introduction
2
AI Sustainability in Practice Part One
3
AI Sustainability in Practice Part Two
4
AI Fairness in Practice
5
Responsible Data Stewardship in Practice
6
AI Safety in Practice
7
AI Explainability in Practice
8
AI Accountability in Practice

Intended Audience

The workbooks are primarily aimed at civil servants engaging in the AI Ethics and Governance in Practice Programme — whether as AI Ethics Champions delivering the curriculum within their organisations by facilitating peer-learning workshops, or as participants completing the programmes by attending these workshops. Anyone interested in learning about AI ethics, however, can make use of the programme curriculum, the workbooks, and resources provided. These have been designed to serve as stand-alone, open access resources. Find out more at aiethics.turing.ac.uk.

There are two versions of each workbook:

- **Facilitator Workbooks** (such as this document) are annotated with additional guidance and resources for preparing and facilitating training workshops.
- **Participant Workbooks** are intended for workshop participants to engage with in preparation for, and during, workshops.

Introduction to This Workbook

Project teams frequently engage in tasks pertaining to the technical safety and sustainability of their AI projects. In doing so, they need to ensure that their resultant models are reproducible, robust, interpretable, reliable, performant, and secure. The issue of AI safety is of paramount importance, because possible failures have the potential to produce harmful outcomes and undermine public trust. This work of building safe AI outputs is an ongoing process requiring reflexivity and foresight. To aid teams in this, the workbook introduces the core components of AI Safety (reliability, performance, robustness, and security), and helps teams develop anticipatory and reflective skills which are needed to responsibly apply these in practice. The workbook is divided into two sections, Key Concepts and Activities.

Key Concepts Section

This section provides content for workshop participants and facilitators to engage with prior to attending each workshop. It covers the four safety objectives and provides case studies aimed to support a practical understanding of technical safety of AI systems. The section also provides best practices to put considerations of accuracy and performance, reliability, security, and robustness in operation at every stage of the AI project lifecycle. Topics discussed include:

Part One: Introduction to AI Safety

1. Accuracy and Performance Objective
2. Reliability Objective
3. Security Objective
4. Robustness Objective

Part Two: Putting AI Safety into Practice

1. Safety Self-Assessment and Risk Management

Activities Section

This section contains instructions for group-based activities (each corresponding to a section in the Key Concepts). These activities are intended to increase understanding of Key Concepts by using them.

Case studies within the AI Ethics and Governance in Practice workbook series are grounded in public sector use cases, but do not reference specific AI projects.

Conceptualising AI Safety

Build a common vocabulary and understanding of AI Safety objectives by reflecting on how participants would define these and discuss the definitions we share in this workshop.

Identifying AI Safety Risks

Enhance understanding of AI safety risks and mitigation strategies in the public sector.

Safety Self-Assessment

Recognise safety considerations at relevant stages of the project lifecycle model.

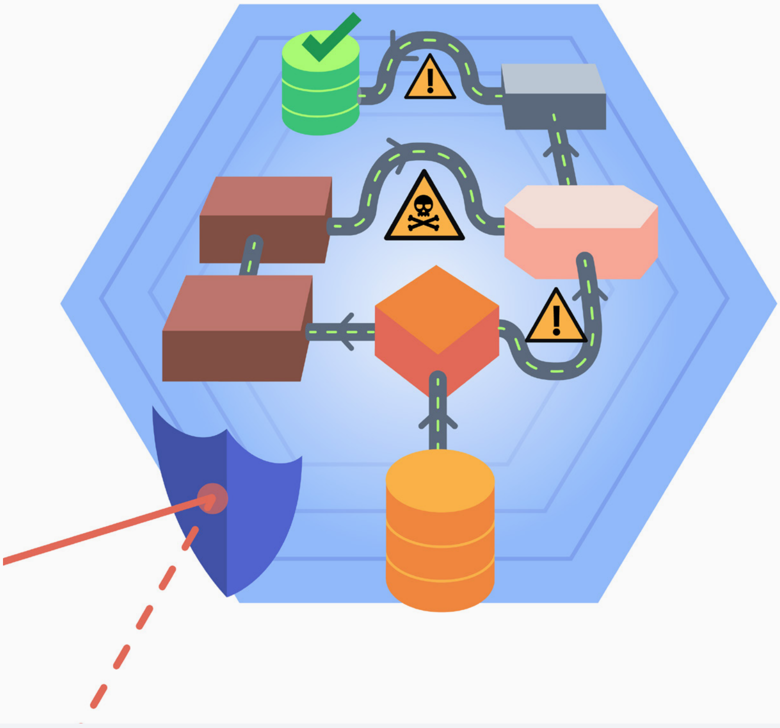
Note for Facilitators

Additionally, you will find facilitator instructions (and where appropriate, considerations) required for facilitating activities and delivering capacity-building workshops.

AI Safety in Practice

Key Concepts

Part One: Introduction to AI Safety



A technically safe and sustainable AI system is **accurate and performant, reliable, secure, and robust**. Ensuring these goals, however, is a difficult and unremitting task Project teams must consider two different but related aspects:

- AI safety of the model, which focuses on ensuring that the algorithms, architectures, and parameters within an AI system are both technically sustainable and safe.

- AI safety of the system, which focuses on ensuring that the safety goals are met considering the broader context in which the model operates (e.g. the interaction with its environment, users, and other systems).

Because AI systems operate in a world filled with uncertainty, volatility, and flux, the challenge of building technically safe and sustainable AI can be especially daunting. This job, however, must be met head-on. Only by making the goal of producing technically safe and sustainable AI technologies central to your project, will you be able to mitigate risks of your system failing at scale when faced with real-world unknowns and unforeseen events. The issue of **AI safety** is of paramount importance, because these potential failures may both produce harmful outcomes and undermine public trust.

A Note on AI Safety

More recently, the widespread deployment and adoption of pre-trained generative AI models have increased awareness of AI safety. However, the issue of what AI safety means, and what the scope of AI safety concerns should include, are increasingly contested matters.

For instance, some relate AI safety to one or more of the following processes:[1]

- ensuring that the deployment of an AI/ML system complies with its intended purpose;
- ensuring that future AI/ML systems align with human values and goals;
- ensuring limitations on the capabilities of AI/ML systems are put in place to prevent them from causing catastrophic or existential harms;
- ensuring that AI/ML systems embed a broader range of values and intentions of society as a whole; or
- ensuring the safe and reliable operation of AI/ML systems through the development of technical methods and tools.

Currently, the loose demarcation of the term is matched with efforts to force it into various broader contexts (i.e. to use 'AI safety' as a catch-all phrase that encompasses AI ethics and governance as such) without sufficient consideration for more nuanced discursive approaches. In the AI Ethics and Governance in Practice workbook series, we understand the concept of AI safety through a sociotechnical lens that focuses on the specific technical characteristics that need to be safeguarded in actual practices of designing, developing, and deploying technically safe AI systems. We therefore keep the concept of 'AI safety' distinct from other salient issues and concepts like AI sustainability, fairness, equity, transparency, and accountability, which have been central to AI ethics and responsible innovation debates and practices, while treating all of these concepts as interrelated areas of focus. In this way, AI ethics and governance can advance the mitigation of the broad range of potential AI risks.

In order to safeguard that your AI/ML model and system functions safely, you must prioritise the technical objectives of:

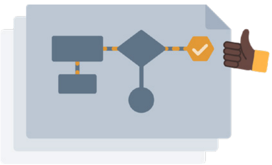
1. Performance

There are several performance metrics to evaluate model performance, such as accuracy, precision, and recall. Accuracy is a metric that refers to the degree to which it generates correct outputs (such as both true positives and true negatives). The accuracy metric is best applied when the dataset is balanced — that is all classes have relatively similar numbers of samples. Accuracy is also known for being misleading in the case of different class proportions (e.g. different classes of anomaly detection in images) since assigning samples to the prevalent class is one way of achieving high accuracy. Overall, the objective of accuracy is to give an overall indication of how good the performance of the system is in contributing to error-free and reliable predictions and decisions.

Example of Performance

It is important to note that there is not a one-size-fits-all approach to accuracy and performance. For instance, a facial recognition system may be marketed as having 85 to 95 per cent accuracy. However, this singular performance metric may obfuscate variation in accuracy across and between different racial groups, with studies noting that images of darker skinned women in particular have the highest error rates when comparing against subgroups.[2] [3] [4] Thus, while the AI system may have a high performance rate, the system does not perform with equal accuracy across all groups.

Considerations of performance relate closely with principles of fairness through a context-based and society-centred approach. Refer to the AI Fairness in Practice workbook.

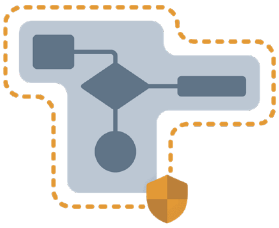


2. Reliability

Reliability of AI systems indicate the likelihood of adherence to intended functionality and to the specifications they were programmed to carry out. Additionally, a reliable system is capable of reporting uncertainty in decision making, adapting to changing data, and generalising robustly. As such, the objective of a reliable AI system is to behave exactly as its designers intended and anticipated while effectively adapting to dynamic environments. Whilst performance refers to the degree of successful operation and the proportion of correct outputs generated by a system, reliability indicates the level of adherence to design specifications and functionality while responding to changing environments.

Example of Reliability

In a medical setting, a reliable model needs to perform a stated function that will improve clinical decision-making and to reduce unknown or spurious results. For example, an AI system designed to assist doctors in identifying malignant nodules in CT image scans may be unreliable if it has not been trained on a diverse population or there is a lack of high-quality curated datasets, thus impeding detection.[5] This could result in stressful delays in follow-up diagnoses or, more worryingly, inaccurate assessments.[6]

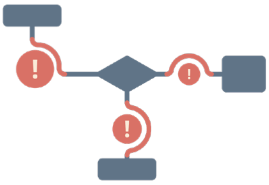


3. Security

A secure AI/ML system is capable of maintaining the integrity of the information that constitutes it, more specifically keeping confidential information safe. The goal of security encompasses the protection of several operational dimensions of an AI system when confronted with possible adversarial attacks. For instance, a system's security can be compromised by data poisoning wherein the attacker can manipulate the dataset to output false results.[7]

Example of Security

A healthcare platform employs AI to offer insights and recommend personalised treatment plans based on patient data, which includes sensitive information (e.g. medication history) found in medical records. However, because the AI project has in place inadequate security measures, the system is vulnerable to potential data breaches. If malicious actors gain unauthorised access to patients' personal health information, patients may be impacted due to infringement on their privacy rights and violations of data protection principles. The impacts would be even more acute for patients with historically stigmatised medical conditions (e.g. HIV/AIDS, mental health disorders).[8] For example, a breach exposing their private health details could exacerbate situations of discrimination, social stigma, and cause potential harms to their wellbeing.[9]



4. Robustness

The measure of robustness is the strength of a system's integrity and the soundness of its operation in response to unanticipated or abnormal conditions. The objective of robustness can be thought of as the goal that an AI system functions reliably and accurately under unforeseen conditions.

Example of Robustness

The robustness of AI/ML models may be weakened in cases where unexpected changes in the operating environment affect the reliability of the model's mapping function. For instance, a a pathology detection system, which has been trained on MRI scans made by a single scanner manufacturer, is deployed in a hospital that uses older MRI scanners made by a different manufacturer. This results in higher error rates and less reliability in the functioning of the model.[10]

This workbook focuses on safety considerations as it relates to narrow AI/ML projects (i.e. those models focused on specific tasks). Subsequent editions will be expanded to include generative AI examples and use cases.

Achieving these technical objectives requires that your technical team put careful forethought into how to construct a system **that accurately and dependably operates in accordance with its designers' expectations even when confronted with unexpected changes, anomalies, and perturbations**. Building an AI system that meets these safety goals also requires rigorous self-assessment, testing, validation, and re-assessment as well as the integration of adequate mechanisms of risk management, oversight, and control into its real-world operation.

Robust and secure AI systems are those that consistently maintain their expected performance levels, even in challenging and potentially hazardous situations. While security aims to safeguard the model or system from deliberate sabotage or forced failure, robustness aims to prevent model errors or unmodeled phenomena caused by uncertainties or complexities in the environment.[11]

Machine Learning Operations (MLOps)

The testability of AI systems is a growing concern as the non-deterministic nature of these systems means that traditional software testing methods widely used for debugging and ongoing maintenance and safety are inadequate in detecting model degradation.

Machine Learning Operations (MLOps) is a new field applying on-going maintenance efforts more relevant for AI systems, where issues such as model degradation over time mean that AI systems require continual efforts to remain accurate, reliable, secure and robust.[12] [13]

A Closer Look at AI Safety Objectives

It is important that you gain a strong working knowledge of each of the safety relevant operational objectives (**performance**, **reliability**, **security**, and **robustness**):

Objective 1

Performance

To measure the performance of a model, in machine learning, **accuracy** is the proportion of examples for which it generates a correct output. This performance measure is also sometimes characterised conversely as an **error rate**, that is the fraction of cases for which the model produces an incorrect output. Keep in mind that, in some instances, the choice of an acceptable error rate or accuracy level can be adjusted in accordance with the use case specified by the needs of the application. In other instances, it may be largely set by a domain established benchmark.

As a performance metric, accuracy should be a central component establishing and nuancing your team's approach to safe AI. That said, specifying a reasonable performance level for your system may also often require you to refine or exchange your measures of accuracy and performance. For instance, if certain errors are more significant or costly than others, a metric for total cost can be integrated into your model so that the cost of one class of errors can be weighed against that of another. For further consideration of performance metrics, also refer to AI Fairness in Practice.

One way to better understand the performance of the model as well as this possible trade-off of errors is through a table commonly referred to as a **confusion matrix.**[14] A confusion matrix provides values that can be used to calculate model metrics such as recall, precision, and AUC-ROC curves, which will be covered in greater depth in this section.

KEY CONCEPT

Accuracy

The proportion of examples for which the model generates a correct output.

KEY CONCEPT

Error Rate

The fraction of cases for which the model produces an incorrect output.

KEY CONCEPT

Confusion Matrix

A confusion matrix is a table that depicts the correctness values for the actual and predicted classes of the model's outputs.

A confusion matrix has four quadrants:

True Positives

Correct **positive** predictions

False Positives

Incorrect **positive** predictions

False Negatives

Incorrect **negative** predictions

True Negatives

Correct **negative** predictions

		Actual Class	
		P	*N*
Predicted Class	*P*	**True Positives** ***(TP)***	**False Positives** ***(FP)***
	N	**False Negatives** ***(FN)***	**True Negatives** ***(TN)***

To explain each of the quadrants we will use the example of a model used to determine whether or not a patient has a particular medical diagnosis:

True Positives model predicts that the patient **has** the disease and **they do**	**False Positives** model predicts the patient **has** the disease, but **they do not**
False Negatives model predicts that the patient does **not have** the disease, but **they do**	**True Negatives** model predicts that the patient does **not have** the disease, and **they do not**

False Positive and False Negative are also known as Type 1 and Type 2 errors. Differing project contexts and use cases place higher degrees of importance on minimising Type 1 errors over Type 2 errors and vice versa. Determining how to handle error trade-offs such as the Type 1-Type 2 trade-off can have serious implications for impacted stakeholders.

For example, in the case of medical diagnoses, a team might design the model to minimise Type 2 errors (False Negatives: model predicts that the patient does not have the disease, but they do) so that a diagnosis is not missed.

However, with this, comes a greater prevalence of Type 1 errors (False Positives: telling a patient that they have a disease when they do not). In this example, Type 1 errors come with costs such as psychological trauma or adverse physical effects if treatment must be undertaken — both resulting from receiving a medical diagnosis that is incorrect.

In addition to better understanding the components that make up the confusion matrix such as Type 1 and Type 2 errors, there are also various performance metrics that can be calculated using the values found in the confusion matrix.

Recall (also known as true positive rate or sensitivity) is a useful metric that explains the number of correct predictions the model made from all of the positive classes. This is calculated by dividing the number of True Positives by the sum of True Positives and False Negatives. To continue the above example, recall tells us how many patients we correctly identified as having the disease out of all of the patients who actually have the disease.

A Note on Metrics

An AI system's accuracy, precision, and recall are useful in gaining some quick insights into baseline trends within a dataset(s). Monitoring these trends over time can help teams gain a better understanding of the model’s overall performance and the factors that impact its reliability, security and robustness. Tracking model performance also helps to identify bigger safety factors such as model drift, as well as a dataset’s basic integrity and possible hints of bias or degradation within the model.

Together, accuracy, precision, and recall comprise three types (but not exhaustive types) in an expanding group of AI evaluation metrics and performance monitoring techniques. By applying existing techniques, there remains a better chance of identifying the nuances of a model /AI system performance an increased opportunity to become familiar with the data and possible inaccuracies, and an enhanced ability to improve and optimise the model.

Precision instead looks at all of the classes predicted as positive and determines how many of those predictions are positive in actuality. This is accomplished by taking the number of True Positives and dividing it by the sum of True Positives and False Positives. In this example, precision tells us the measure of patients that have been correctly identified as having the disease out of all the patients that actually have it.

A related term is that of specificity. Instead of looking at the likelihood that a given positive prediction is actually positive, specificity looks at all the classes predicted as negative and determines how many of them are actually negative.

Often, precision and recall are not sufficient when considering performance metrics. Additionally, there are trade-offs that can occur when comparing two models when precision may be low but recall is high or the opposite. This challenge is often referred to as the Precision-Recall trade-off.[15] Often when teams try to increase precision, this increase tends to come at a cost of lower recall and vice versa. Therefore, **in order to account for this and to measure recall and precision simultaneously, the F1-score can be employed**. F1-scores are particularly helpful when the classes of the model are imbalanced.

F1-score is the harmonic mean between recall and precision.

$$\text{F1-Score} = \frac{\textbf{2*Recall*Precision}}{\textbf{Recall + Precision}}$$

The F1-score (also known as F-score) is a measure of a model's accuracy on a dataset. It is used to evaluate binary classification systems which classify examples into 'positive' or 'negative'. F1-score is a method for combining the precision and recall of the model, defined as the 'harmonic mean' of that model's precision and recall. Harmonic mean is a type of average used for numbers representing a rate or ratio, such as precision and recall in information retrieval. Typically, therefore, the F1-score is used for evaluating information retrieval systems in search engines, and for different types of machine learning models, in particular those using natural language processing.

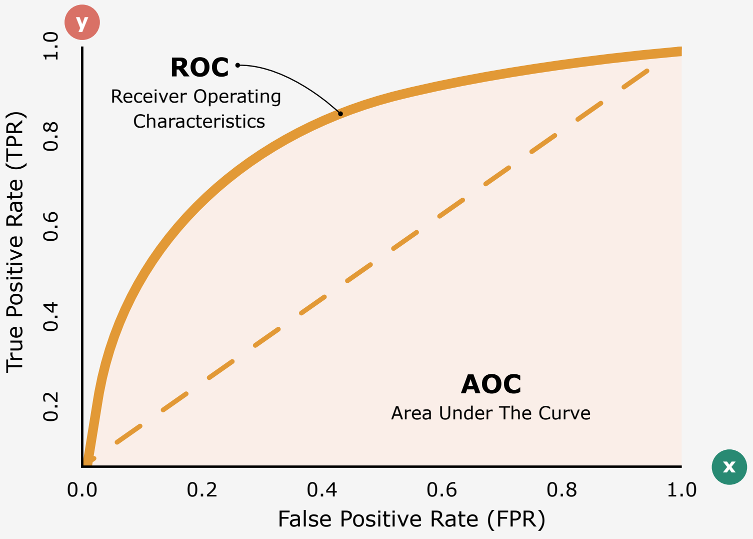
Area Under the Curve (AUC) and Receiver Operating Characteristics (ROC)

While the previous methods serve as useful metrics for better understanding specific aspects of model performance, a metric often used to assess the overall model performance is known as the AUC-ROC (Area Under the Curve-Receiver Operating Characteristics) curve. The AUC-ROC curve is particularly useful for binary classification and consists of two components: the ROC and the AUC.

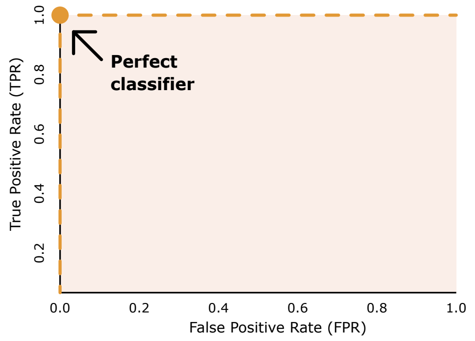
Receiver Operating Characteristics (ROC)

The True Positive Rate (also known as recall, discussed above) and False Positive Rate (also known as 1- Specificity) can be calculated from the confusion matrix. The receiver operating characteristics (ROC) is a probability curve that plots the True Positive Rate (TPR) of the model against the False Positive Rate (FPR) at multiple threshold levels, where the Y-axis value stands for TPR and X-axis value for FPR.

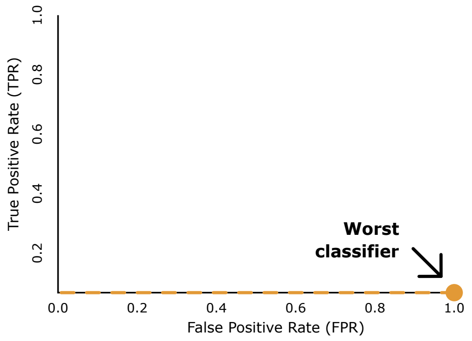
x	y	threshold
0.72	0.24	0.517
0.76	0.28	0.476
0.42	0.12	0.679
0.66	0.24	0.551
0.56	0.20	0.586
0.34	0.04	0.747



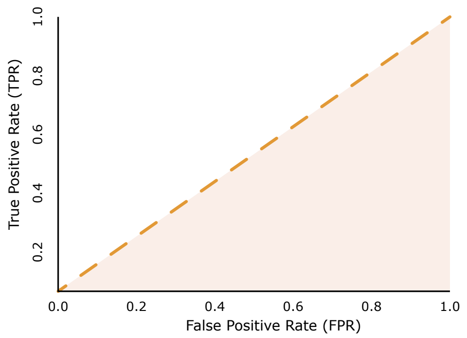
Area Under the Curve (AUC)



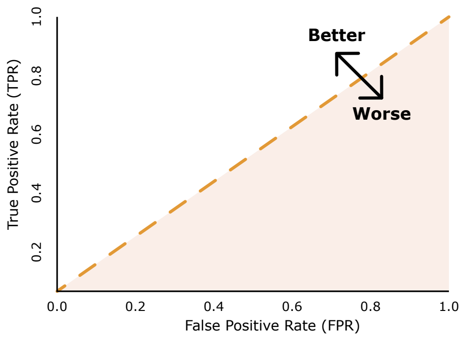
The area under the curve (AUC) explains how effective the model is at distinguishing between classes. A perfect classification would have an inverse L or a square shape and would have an AUC of 1.



A model capable of only making incorrect class predictions (or have 0% accuracy) would have an AUC of 0.



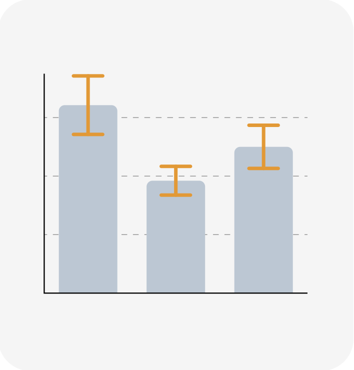
A classifier with 50% accuracy would have a triangle shape and an AUC of 0.5.



To take the previous example, the higher the AUC, the better the model is at correctly classifying patients as having the disease or not.

While there are various performance metrics for teams to choose from that each contribute to the team's understanding of the model performance, there are also two more tools that provide greater comprehension, namely, confidence intervals and error bars.

Confidence intervals account for possible error by calculating the mean of the data and subtracting and adding an error estimate on either side. Confidence intervals are depicted as ranges (e.g. 160.3cm, 168.7cm). This depicts the range of values that you and your team expect the estimate to fall between with a certain level of confidence, if you were to redo the test. These ranges can help the team to understand with varying degrees of confidence how large the range of values is when the error is calculated, and these are often graphically depicted as error bars on charts, as seen to the right.



Objective 2

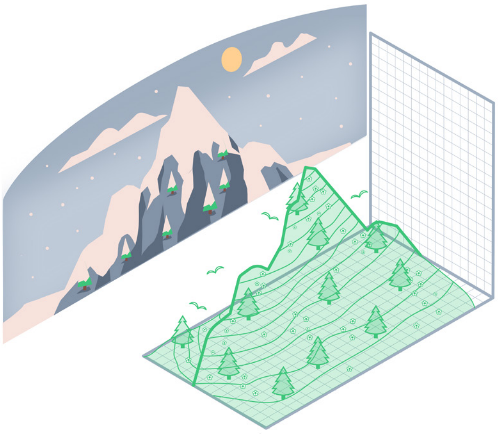
Reliability

The reliability of AI systems indicates the likelihood of adherence to the specification they were programmed to carry out in normal or known circumstances. Reliability is therefore a measure of **consistency** and can help to establish confidence in the safety of a system based upon the dependability with which it operationally conforms to its intended functionality.

Risks Posed to Performance and Reliability

Risk 1: Concept or Model Drift[16]

Once trained, most machine learning systems operate on static models of the world that have been built from historical data which have become fixed in the systems' parameters.[17] This freezing of the model before it is released 'into the wild' makes its performance and reliability especially vulnerable to changes in the underlying distribution of data. When the historical data that have crystallised into the trained model's architecture cease to reflect the population concerned, the model's mapping function will degrade and no longer be able to accurately and reliably transform its inputs into its target output values. These systems can quickly become prone to error in unexpected and harmful ways.



Consider a predictive model developed to optimise bus routes based on historical data. During a pandemic, the model may become less effective, as an increase in remote working alters typical commuter behaviour. Similarly, the implementation of new cycling lanes or changes in parking policies can contribute to concept drift. Large-scale events and human interventions can bring significant societal change. Some of these changes might be more subtle and take longer to understand.[18] This is because the transition of change varies. The target distribution may change abruptly (sudden drift) or progressively (gradual drift), the old concepts may reoccur after some time (recurring drift), or a new concept may replace the old one slowly in a continuous manner (incremental drift).[19]

There has been much valuable research done on methods of detecting and mitigating concept and distribution drift.[20] [21] You should consult with your technical team to ensure that its members have familiarised themselves with this research and have sufficient knowledge of the available ways to confront the issue. It is also important to adopt mechanisms that ensure the correct and appropriate use of the system, and that the system is not being repurposed for tasks for which it was not envisaged nor accounted for.[22] In all cases, you should remain vigilant to the potentially rapid concept drifts that may occur in the complex, dynamic, and evolving social or physical environments in which your AI project will intervene. This is especially relevant for sensitive domains and application contexts of application and use. Remaining aware of these transformations in the data is crucial for safe AI, and your team should actively formulate an action plan to anticipate and to mitigate their impacts on the performance of your system. This should include consulting experts with appropriate contextual and domain knowledge, who are well-equipped to identify underlying changes in relevant social environments that may lead to concept or model drift and decreased system performance and reliability.

Risk 2: Brittleness[23]

Another possible challenge to the accuracy, performance, and reliability of AI/ML systems arises from the inherent limitations of the systems themselves. Many high-performing and complex AI/ML models, such as deep neural nets (DNN), rely on massive amounts of data and brute force repetition of training examples to tune the thousands, millions, or even billions of parameters, which collectively generate their outputs.

However, when they are actually running in an unpredictable world, these systems may have difficulty processing unfamiliar events and scenarios or previously ‘unseen’ inputs that trigger unexpected inferences which generate inaccurate or nonsensical outputs. They may make unexpected and serious mistakes, because they have neither the capacity to contextualise the problems they are programmed to solve nor the common-sense ability to determine the relevance of new or unknown ‘unknowns’. Moreover, these mistakes may remain unexplainable given the high-dimensionality and computational complexity of their mathematical structures. This fragility or brittleness can have especially significant consequences in safety-critical applications like fully automated transportation and medical decision support systems where undetectable changes in inputs may lead to significant failures.[24] [25] Consider an advanced driver-assistance system using AI/ML that fails to detect and respond to unexpected events on the road caused by dynamic weather conditions, such as fallen tree branches or lane markings covered by snow.[26]

Gathering more data (e.g. more pictures of fallen tree branches in roads) is only a temporary fix. However, in dynamic scenarios where we can encounter an indefinite range of unexpected events, this approach may lead to high workload and not be impractical.[27]

While progress is being made in finding ways to make these models more robust, it is crucial to consider safety first when weighing up their viability.

Risk 3: Overfitting

A common, recurring challenge to the performance and reliability of AI/ML systems is overfitting. Overfitting occurs when the model's mapping function is matched too closely to the patterns arising in the training data.[28] [29] The result of this overfitting is that the trained system is unable to respond effectively to new, unseen data, thereby making it perform poorly or unreliably in real world scenarios. Overfitting occurs when models erroneously take the noise of the training dataset as the signals of the underlying data distribution. Therefore, when new data that does not have the same noise is applied to the same model, it is not able to pick out the actual, correct output-generating signal and accordingly fails.

Overfitting takes place for a variety of reasons including using too many features to predict the target variable in question (without enough data to support that quantity of variables–this is sometimes called the 'curse of dimensionality')[30] as well training the model too extensively on the same training dataset. In both scenarios the model is unable to generalise to new data which significantly reduces its predictive power.

In the context of transportation, consider an AI system designed to predict traffic congestion based on historical data from a specific city. If the model was extensively trained on a single dataset from that city, it will become specialised in predicting traffic patterns unique to that location. When confronted with new data from a different city, the AI system may struggle. Unique factors that could affect traffic in the new city, such as different commuting habits or the presence of major sporting venues, might not be accurately accounted for, leading to unreliable predictions and decreased performance.

There are various ways to prevent or mitigate overfitting. Among others, these include:[31] [32]

- Testing methods for overfitting, such as k-fold cross validation.
- Early stopping: pausing the training before the model learns the noise in the data.
- Feature selection (or pruning): identifying the most important features within the training dataset that impact the outputs.
- Dimensionality reduction (e.g. PCA, LDA, t-SNE): reducing the number of features without compromising the meaningful properties of the original data.

- Regularisation: grading features according to importance and eliminating those features that do not impact the prediction outcomes.[33]

It may also be helpful to make sure that the model has appropriate training-validation-testing splits, so that validation and testing sets can be effectively compared to the training dataset.[34] For example, if the error rates for the training dataset are low, but the test dataset produces high error rates, this signals overfitting may be taking place. Seeking out other means of external validation may support less overfitting as well. Overfitting is a very common issue, and you should consult with your technical team to ensure that its members have familiarised themselves with mitigation measures for this while simultaneously preventing the opposite issue of underfitting.

Risk 4: Unpredictable and Non-Deterministic Behaviour of Autonomous AI/ML Systems

Some autonomous AI/ML systems can use probabilistic mechanisms to learn from and adapt to changing run-time environments. Such self-updating systems, which evolve continuously when operating in the wild, may behave in unpredictable ways. Because they have to cope with uncertain and changeable surroundings, these systems typically have a non-deterministic and dynamic character, which resists commonly accepted methods of formal verification, testing, and validation. Moreover, in virtue of their complexity and non-linearity, these models often yield behaviours that cannot directly be interpreted or explained. This leads to a crucial two-pronged difficulty:

1. that the intended functionality of fully autonomous systems cannot be formalised at design-time into specific, suitable, and checkable requirements of the sort necessary to fully assure their trustworthiness; and,

2. that the logic underlying the behaviours of these systems cannot be readily accessed in a human understandable way so that their outcomes can be sufficiently demonstrated to reflect design intentions.

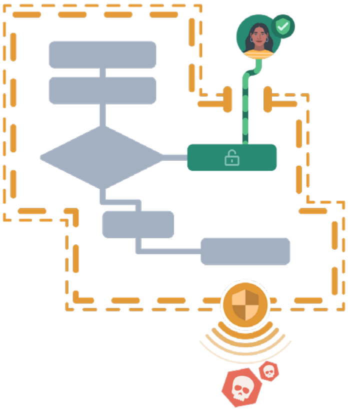
Non-deterministic models also create challenges for reproducibility, as the system can produce a different result each time it is run even when the same inputs are used.[35] This differs from deterministic algorithms which produce a single output for the same input, all things held equal. It is important for teams to consider whether a non-deterministic model is appropriate for the specific context. Teams must consider the importance of interpretability and reproducibility in safety-critical sectors and contexts that produce significant impacts on individual lives.

For more information about interpretability, refer to AI Explainability in Practice.

Objective 3

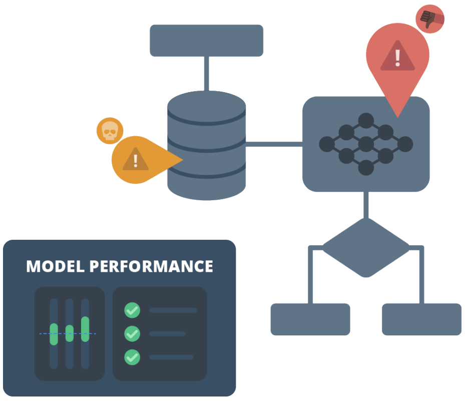
Security

The goal of security encompasses the protection of several operational dimensions of an AI system when confronted by adversarial attacks. A secure system is capable of maintaining the **integrity** of the information that constitutes it. This includes protecting its architecture from the unauthorised modification or damage of any of its component parts. A secure system also remains continuously **functional** and **accessible** to its authorised users and keeps **confidential** and **private information** secure even under hostile or adversarial conditions.



Objective 4

Robustness



The objective of robustness can be thought of as the goal that an AI system functions reliably and accurately under harsh conditions. These conditions may include adversarial intervention, implementer error, or skewed goal-execution by an automated learner (in reinforcement learning applications). The measure of robustness is therefore the strength of a system's integrity and the soundness of its operation in response to difficult conditions, adversarial attacks, perturbations, data poisoning, and undesirable reinforcement learning behaviour.

⚠ Risks Posed to Security and Robustness

Risk 1: Adversarial Attack[36]

Adversarial attacks on machine learning models maliciously modify input data — often in imperceptible ways — to induce them into misclassification or incorrect prediction. For instance, by undetectably altering a few pixels on a picture, an adversarial attacker can mislead a model into generating an incorrect output (like identifying a panda as a gibbon or a 'stop' sign as a 'speed limit' sign) with extremely high confidence. While a good amount of attention has been paid to the risks that adversarial attacks pose in deep learning applications like computer vision, these kinds of perturbations are also effective across a vast range of machine learning techniques and uses such as spam filtering and malware detection.

These vulnerabilities of AI systems to adversarial attacks have serious consequences for AI safety. The existence of cases where subtle but targeted perturbations cause models to be misled into gross miscalculation and incorrect decisions have potentially serious safety implications for the adoption of critical systems like applications in autonomous transportation, medical imaging, and security and surveillance.

A subset of adversarial attacks include model inversion (MI) attacks wherein malicious actors attempt to reconstruct training data or access sensitive information from model parameters (white-box attacks) or from their outputs (black-box attacks). For instance, an MI attack is used reconstruct individuals' identity, including images and biometric details, through class labels (say, their name) or target models.[37]

In response to concerns about the threats posed to a safe and trusted environment for AI technologies by adversarial attacks a field called **adversarial machine learning** has emerged over the past several years.[38] [39] Work in this area focuses on securing systems from disruptive perturbations at all points of vulnerability across the AI pipeline.

One of the major safety strategies that has arisen from this research is an approach called **model hardening**, which has advanced techniques that combat adversarial attacks by strengthening the architectural components of the systems.[40] Hardening is the process of enhancing the security of an AI model that includes identifying inputs that cause the model to produce incorrect outputs, such as false positives or false negatives. To harden effectively, it is important to consider how the AI system collects, processes and stores data, and how this may impact an organisation's privacy and data protection obligations. For example, an AI system hosted on the cloud may send data between different regions. If using a third-party AI system, you will need to understand how inputs from the organisation will be used to retrain the AI system's model.[41]

In addition, teams can protect their data and models through other security techniques including differential privacy (i.e. by adding noise to obfuscate personal data) or federated learning (i.e., training a model with data on decentralised servers).

You should consult with members of your technical team to ensure that the risks of adversarial attack have been taken into account and mitigated throughout the AI lifecycle. A valuable collection of resources to combat adversarial attack can be found at https://github.com/IBM/adversarial-robustness-toolbox. Other resources are available in the Endnotes.[42] [43] [44]

Risk 2: Data Poisoning[45]

A different but related type of adversarial attack is called data poisoning.[46] [47] This threat to safe and reliable AI involves a malicious compromise of data sources at the point of collection and pre-processing. Data poisoning occurs when an adversary modifies or manipulates part of the dataset upon which a model will be trained, validated, and tested. In the case of extremely large models, access to a small subset of the training data is enough to carry out data poisoning.[48] By altering a selected subset of training inputs, a poisoning attack can induce a trained AI system into curated misclassification, systemic malfunction, and poor performance.

In order to combat data poisoning, your technical team should become familiar with the state of the art in filtering and detecting poisoned data.[49] However, such technical solutions are not enough. Data poisoning is possible because data collection and procurement often involves potentially unreliable or questionable sources. When data originates in uncontrollable environments like the internet, social media, or the Internet of Things, many opportunities present themselves to ill-intentioned attackers, who aim to manipulate training examples. Likewise, in third-party data curation processes (such as ‘crowdsourced’ labelling, annotation, and content identification), attackers may simply handcraft malicious inputs. Your project team should focus on the best practices of responsible data management, so that they are able to tend to data quality as an end-to-end priority. This may include tracking the data provenance and lineage. Safety considerations by the project team should further include data security.

For more information on responsible data management and data security, refer to Responsible Data Stewardship in Practice.

Risk 3: Transfer Learning Attacks

Transfer learning attacks are another form of adversarial attack that target large pre-trained models. Many ML systems rely on pre-trained base models that are then tuned to serve a specific purpose. These pre-trained models are often publicly available on open-source platforms, thereby increasing their susceptibility to malicious adversarial modifications which target the pre-trained model with the ultimate aim of damaging the task-specific models that draw on them.

Another form of transfer learning attack is referred to as a backdoor attack which occurs when weights of the pre-trained model are poisoned-injected with vulnerabilities that illuminate “backdoors”. These vulnerabilities allow an attacker to manipulate the fine-tuned model through inputting an arbitrary keyword. Thus, the poisoned pre-trained model is used to fine-tune a task-specific model resulting in the poisoning of that model. Backdoor attacks may go undetected during normal testing with the absence of a trigger. They differ from other evasive adversarial attacks because they rely on trigger embeddings that are ‘input- and model-agnostic’, meaning the trigger will always cause an incorrect prediction on any poisoned model or input.[50]

For instance, an autonomous vehicle navigation system uses pre-trained deep learning model for object detection. The project team found a model on an open-source platform and is aware that it was trained on dataset of diverse traffic scenarios. At attacker poisons the weights of the pre-trained model with an “invisible trigger”: a “Detour” sign with slight modifications. When the autonomous vehicle encounters a road sign resembling the poisoned “Detour” sign, the trigger activates. Because of the injected backdoor, the model incorrectly interprets this sign as a speed limit increase change, causing the vehicle to accelerate and potentially leading to an accident.

There are several ways to mitigate transfer learning attacks. One way to defend against these risks is for developers who created the pre-trained models to clearly describe what their pre-trained models do and how to mitigate risks. Additionally, developers using pre-trained models should ensure they retrieved the model from a secure source. In the case of image classification, several examples of robust defense mechanisms include randomising the input via dropout, modifying the weights of the task-specific model, and utilising ensemble methods. When considering backdoor attacks via weight poisoning, researchers at Carnegie Mellon University have developed a method entitled Label Flip Rate. This method is applicable to natural language processing algorithms and attempts to identify the triggers that have been embedded in the pre-trained model by calculating the percentage of instances that were not initially in the target class but were then classified as the target as a result of the attack.[51]

Risk 4: Misdirected Reinforcement Learning Behaviour[52]

A different set of safety risks emerges from a form of machine learning called reinforcement learning (RL). In the more widely applied methods of supervised learning that have largely been the focus of this guide, a model transforms inputs into outputs according to a fixed mapping function that has resulted from its passively received training. In RL, by contrast, the learner system actively solves problems by engaging with its environment through trial and error. This exploration and ‘problem-solving’ behaviour is determined by the objective of maximising a reward function that is defined by its designers.

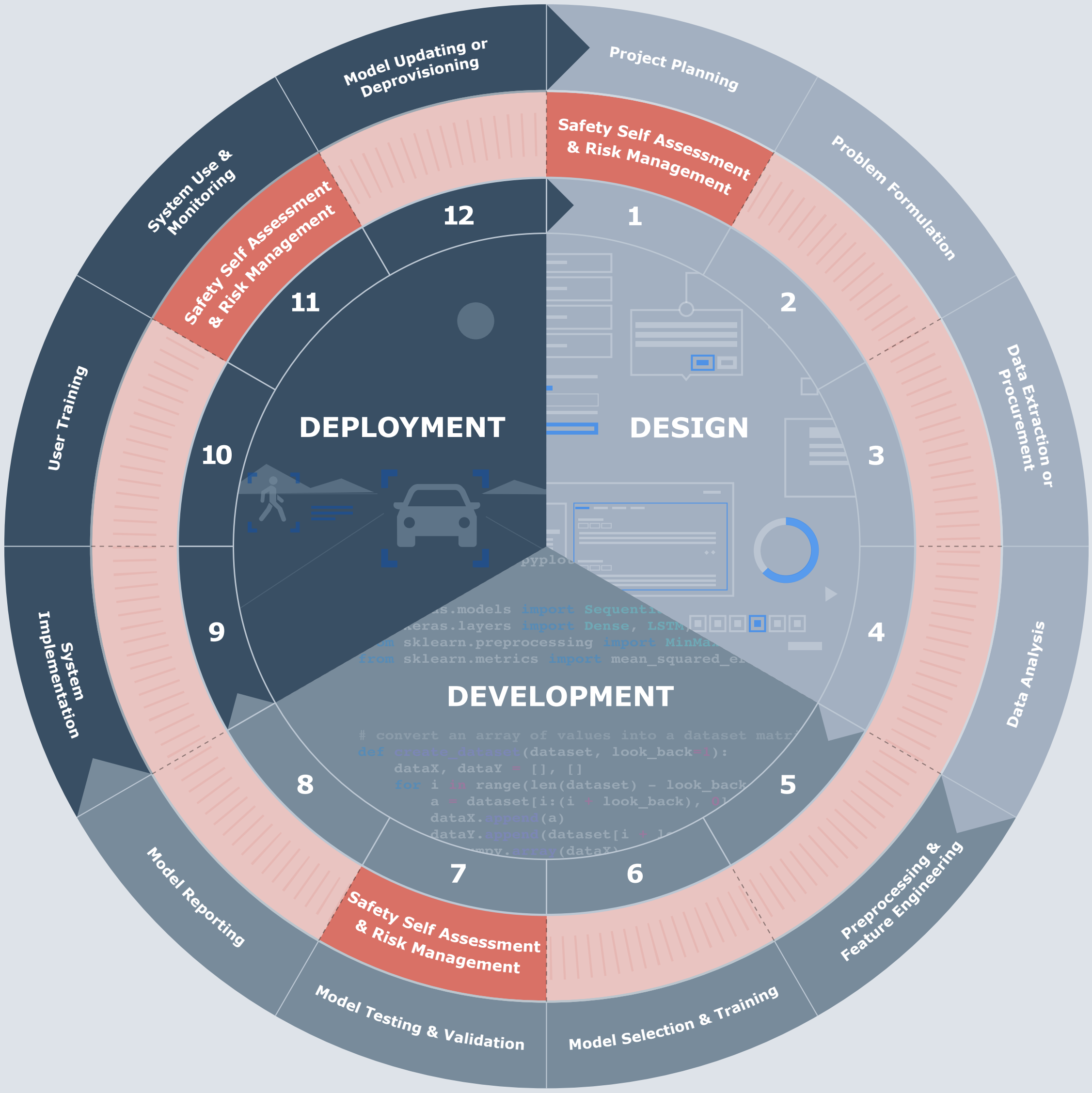
This flexibility in the model, however, comes at the price of potential safety risks. An RL system, which is operating in the real-world without sufficient controls, may determine a reward-optimising course of action that is optimal for achieving its desired objective but harmful to people. Because these models lack context-awareness, common sense, empathy, and understanding, they are unable to identify, on their own, scenarios that may have damaging consequences but that were not anticipated and constrained by their programmers. This is a difficult problem, because the unbounded complexity of the world makes anticipating all of its pitfalls and detrimental variables veritably impossible.

Existing strategies to mitigate such risks of misdirected reinforcement learning behaviour include:

- Running extensive simulations during the testing stage, so that appropriate measures of constraint can be programmed into the system.
- Continuous inspection and monitoring of the system, so that its behaviour can be better predicted and understood.
- Finding ways to make the system more interpretable so that its decisions can be better assessed.
- Hard-wiring mechanisms into the system that enable human override and system shut-down.

Part Two: Putting AI Safety into Practice

The safety risks you face in your AI project will depend, among other factors, on the sort of AI models you are using, the type of applications in which those techniques are going to be deployed, the provenance of your data, the way you are specifying your objective, and the problem domain in which that specification applies. As a best practice, regardless of this variability of techniques and circumstances, **safety considerations of performance, reliability, security, and robustness should be in operation at every stage of your AI project lifecycle.**


DESIGN
DEVELOPMENT
DEPLOYMENT
1
2
3
4
5
6
7
8
9
10
11
12
Project Planning
Problem Formulation
Data Extraction or Procurement
Data Analysis
Preprocessing & Feature Engineering
Model Selection & Training
Model Testing & Validation
Model Reporting
System Implementation
User Training
System Use & Monitoring
Model Updating or Deprovisioning
Safety Self Assessment & Risk Management
Safety Self Assessment & Risk Management
Safety Self Assessment & Risk Management

Safety Self-Assessment and Risk Management

Putting this principle into practice involves the completion of AI safety self-assessments by relevant members of your team at each stage of the workflow. The AI Safety protocols of testing, validating, verifying, and monitoring the AI/ML system are:

- conducting a **Safety Self-Assessment** (that works from safety properties – performance, reliability, security, and robustness);
- identifying risks (that respond to the specific context of the use case and design process);
- taking action to manage, eliminate, or mitigate risks; and
- documenting the actions taken (who, when, how) in a **Risk Management Plan**.

The Safety Self-Assessment and Risk Management Plan should evaluate how your team's design and implementation practices line up with the AI safety objectives. Your AI safety self-assessments should be logged across the workflow on a single document in a running fashion that allows review and re-assessment. The plan consists of three steps:

Step 1

Familiarise yourself with the **Safety Assurance Activities** (on page 32) that are relevant to each project stage by reviewing the following actions intended to assure the achievement of each safety objective.

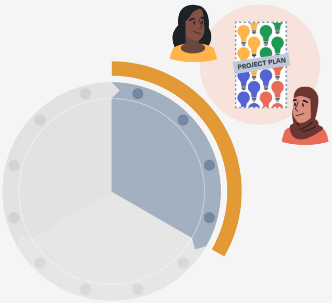
Step 2

Reflect on how your particular AI project might be vulnerable to safety risks at each stage and what challenges exist to achieving each safety objective, within the context of the specific context and use case of your project.

Step 3

Determine and document safety assurance activities you will conduct to manage, eliminate, or mitigate risks to assure that each safety objective is met.

The Safety Self-Assessment and Risk Management template will help you and your team go through steps 2 and 3. Having familiarised yourself with safety assurance activities across the AI project workflow, this template will allow you to identify and document potential risks identified for your project, as well as assurance actions you will implement to address them.



The Stakeholder Engagement Process and Stakeholder Impact Assessment play a crucial role in informing safety self-assessments within AI projects. A participatory approach ensures a comprehensive understanding of the context within which the model will operate and the potential impacts of the AI system. This helps in identifying safety risks that may not be immediately apparent, fostering a more thorough Safety Self-Assessment.

For more information about the Stakeholder Engagement Process and Stakeholder Impact Assessment, refer to AI Sustainability in Practice Part One and AI Sustainability in Practice Part Two, respectively.

Safety Assurance Activities

DESIGN

1 Project Planning

Activities for Performance and Reliability

a. Wherever the prospective AI technology is replacing an existing algorithmic system or a human, or is part of another technical platform or hybrid system that serves the same or similar function, weigh potential improvements in performance and reliability against any disadvantage that could arise because of the replacement.

b. Establish how you will assess the quality and fitness for purpose of procured pre-trained models during Model Selection.

c. Establish who will be supporting the model and AI safety objectives, e.g. in testing, maintenance, monitoring and evaluation over its lifecycle.

d. Drawing on your Data Factsheet, establish proportional safety protocols to ensure data quality is retained during Data Collection and Procurement, and Preprocessing and Feature Engineering.

Activities for Security and Robustness

a. Drawing on your Data Factsheet, establish proportional safety protocols to ensure the reliability of data sources.

b. Establish limitations on who is able to access the system, when, and how.

c. Establish governance over the security of confidential and private information that is processed in the project. In particular, consider how the AI system collects, processes and stores data, and how this may impact your organisation's privacy and data protection obligations.

d. Document what security measures were chosen and the reasons for choosing those measures. Statistically verify the achievement of these safety objectives, in terms of:

   - the system's ability to protect its architecture from unauthorised modification or damage of any of its component parts;

   - the system's ability to remain continuously functional and accessible to its authorised users; and

   - the system's ability to keep confidential and private information secure under adversarial conditions.

e. Put in place procedures and controls to ensure that the system functions reliably and accurately under harsh conditions (which may include adversarial intervention, implementation error, perturbations, data poisoning, or undesirable reinforcement learning behaviour).

f. Put in place measures to minimise risks of physical, psychological, or moral harm from the AI system operations (e.g. ongoing monitoring from a human-in-the-loop).

DESIGN

2 Problem Formulation

Activities for Performance and Reliability

a. Assess the risks of explainability, accuracy, and reliability in using the given model over other alternatives.

b. Consider whether the system can meet acceptable thresholds of accuracy, performance, and reliability given its use context and level of impact (e.g. a high impact system operating in a safety critical sector like transport, social care, or healthcare will likely have high performance thresholds).

Activities for Security and Robustness

a. Assess the computational methods and sociotechnical considerations, including the use and relevance of proxy data and where the algorithmic models open up to cybersecurity risks.

b. Statistically verify the system robustness in terms of the system's immunity to adversarial attacks and adverse conditions. Document the reasons for choosing those measures and how acceptance criteria for security will be encoded into system specifications.

DESIGN

3 Data Extraction or Procurement

Activities for Performance and Reliability

- Sufficiently vet and verify data sources for their measurement accuracy and reliability.

Activities for Security and Robustness

- Assess reliability of data sources to avoid procuring compromised data and to protect against data poisoning.

DESIGN

4 Data Analysis

Activities for Performance and Reliability

a. Assess and validate data quality. This includes considering and evaluating whether the data is of appropriate relevance, timely, complete and representative, and in sufficient quantity to meet the accuracy, performance, and reliability needs of your project given its use context.

b. Assess and validate extent to which procured datasets are fit for purpose in the context of your project.

DEVELOPMENT

5 Preprocessing & Feature Engineering

Activities for Performance and Reliability

a. Augment datasets using synthetic data, in collaboration with domain and subject matter experts, to address known gaps and ensure the training data are representative and complete.

b. Conduct early stopping, feature selection, regularisation or resampling techniques, such as k-fold cross-validation to avoid overfitting.

6 Model Selection & Training

Activities for Performance and Reliability

a. Ensure the prioritisation of performance metrics for the system beyond accuracy (e.g. sensitivity, precision, specificity) and the metrics' responsiveness to the given context.

b. Consider the shortcomings of the various established accuracy metrics and how the achievement of the chosen accuracy objective could be statistically verified. Consider and assess whether other related performance measures will be measured and document the reason for choosing those measures. Undertake external validation to test and confirm your model's 'ground truth'. Assess how acceptance criteria for accuracy will be encoded into system specifications.

c. Ensure appropriate understanding and control. Train the model towards performance in a runtime environment.

d. Put in place actions such as feature selection, regularisation, and k-fold cross validation to avoid overfitting.

e. Statistically verify the achievement of reliability and document the reasons for choosing those measures. Assess how acceptance criteria for reliability will be encoded into system specifications.

f. Assess the reliability of any procured pre-trained models and their sources to protect against transfer learning attacks (e.g. reviewing documentation addressing model functionality and risk mitigation).

g. Base the system's design on well-understood techniques that have previously been in operation and externally validated for similar purposes and in the same sector. If not, ensure that diligent processes of testing, verifying, and externally validating the performance of the system occurs.

h. Establish system monitoring and performance evaluation protocols that are proportionate to the system's technological maturity.

i. Establish a test and validation set to compare to the training dataset and avoid overfitting.

j. Assess the runtime environment to ensure appropriate understanding and control. Train the model towards performance in this environment.

k. Monitor and/or test the system to validate for the achievement of acceptance criteria for these objectives during the deployment phase.

l. Set in place standards triggering either the update or deprovisioning of the model.

Activities for Security and Robustness

a. Statistically verify security in terms of:

- the system's ability to protect its architecture from unauthorised modification or damage of any of its component parts;
- the system's ability to remain continuously functional and accessible to its authorised users; and
- the systems ability to keep confidential and private information secure under adversarial conditions.

b. Document the reasons for choosing those measures. Assess how acceptance criteria for security will be encoded into system specifications.

c. Statistically verify the achievement of robustness and document the reasons for choosing those measures. Assess how acceptance criteria for robustness will be encoded into system specifications.

d. Monitor and/or test the system to validate for the achievement of acceptance criteria for these objectives during the deployment phase.

e. Assess the reliability of any procured pre-trained models and their sources to protect against transfer learning attacks, this can include reviewing documentation addressing model functionality and how to mitigate risks.

f. Assess the extent to which pre-trained models developed for different purposes are fit for purpose within your project context.

g. Incorporate defence mechanisms against backdoor or adversarial attacks during Model Selection & Training, and Model Testing & Validation. Model Selection and Training, and Model Testing and Validation.

h. Implement model hardening techniques against inversion attacks during Model Selection & Training, and Model Testing & Validation.

i. Internally and externally validate the model across a wide range of environments.

j. Assess the quality and fitness-for purpose of procured pre-trained models.

k. Set in place standards triggering either the update or deprovisioning of the model in this phase of the project.

7 Model Testing & Validation

Activities for Performance and Reliability

a. Involve diverse stakeholder groups in the model's development and testing to ensure it performs reliably within and between each group.

b. Evaluate and optimise the model using sensitivity analyses and perturbations to training data to minimise the risk of encountering novel data in the runtime environment.

c. Put in place measures to adjust the model testing and validating methods to ensure its accuracy, performance, and reliability.

d. Conduct forms of monitoring, testing and/or validation to verify the system's achievement of benchmarks for performance/reliability and document the results of these verification and whether these benchmarks have been met. If needed, reassess the acceptance criteria for performance/reliability and establish how this criteria be adjusted and why.

e. Monitor and/or test the system to validate for the achievement of acceptance criteria for these objectives during the deployment phase.

Activities for Security and Robustness

a. Evaluate and develop the model to minimise its vulnerability to inversion attacks through model hardening techniques such as adversarial training, architectural modification, regularisation, and data pre-processing manipulation.

b. Evaluate and develop a model that incorporates appropriate defence mechanisms for your model type (such as randomising the input via dropout, modifying the weights of the task-specific model, and utilising ensemble methods) to protect against backdoor or adversarial attacks.

c. Stress-test the system to understand how it responds to adversarial intervention, implementation error, or skewed goal-execution by an automated learner (in reinforcement learning applications).

d. Conduct forms of monitoring, testing and/or validation to verify the system's achievement of benchmarks for security/robustness and document the results of these verification and whether the acceptance criteria for security/robustness has been met. If there is a need to re-assess the acceptance criteria for security/robustness, establish how the criteria will be adjusted and why.

DEVELOPMENT

8 Model Reporting

Activities for Performance and Reliability

- Provide a comprehensive report documenting the performance measures used for evaluating the model (e.g. decision thresholds for classifiers, accuracy metrics).

Activities for all objectives

- Document the intended use of the model, features implemented, training-testing distributions, and other decisions taken during the development phase.

DEPLOYMENT

9 System Implementation

Activities for Performance and Reliability

- Evaluate and optimise the model using sensitivity analyses and perturbations to the training data during System Implementation to minimise risk of encountering novel data in the runtime environment.

Activities for Security and Robustness

a. Internally and externally validate the model across a wide range of environments to ensure robustness.

b. Undertake penetration testing during System Implementation and System Use and Monitoring to mitigate any risks associated with revealing sensitive data to non-trusted third parties.

c. Incorporate run-time detection during System Implementation to identify and trace adversarial examples in real time.

d. Set in place measures, such as hard-wiring mechanisms that enable human override.

DEPLOYMENT

10 User Training

Activities for Performance and Reliability

a. Explain to users the system's functionalities and limitations, familiarise them with the system's expected behaviours and train them on how to recognise and respond to errors or unexpected behaviours (including how to evaluate and feedback on the system's performance).

b. Train users on the ways in which human factors may affect the system's performance in real-world settings (e.g. cognitive biases, the social and environmental context within which the system is embedded).

c. Incorporate sufficient processes to ensure that the deployment of the system does not harm the physical, psychological, or moral integrity of implementers and users.

Activities for Security and Robustness

- Put in place ongoing monitoring to ensure there is a human-in-the-loop, where the system could cause physical, psychological, or moral harm from its operations.

DEPLOYMENT

11 System Use & Monitoring

Activities for Performance and Reliability

a. Monitor and/or test the system to minimise potential risks when considering the achievement of acceptance criteria for accuracy, performance, and reliability.

b. Regularly reevaluate the performance of the system to ensure it is able to keep pace with real world changes that may cause concept drifts and shifts in underlying data distributions.

Activities for Security and Robustness

a. Carry out extensive penetration testing to ensure that sensitive data will not be revealed to non-trusted parties.

b. Incorporate run-time detection to identify and trace in real-time the existence of adversarial examples.

c. Undertake penetration testing during System Use and Monitoring to ensure that sensitive data is not revealed to third parties.

d. Incorporate run-time detection during System Use and Monitoring to identify and trace adversarial examples in real time.

Activities for all objectives

a. Conduct forms of monitoring, testing and/or validation to verify the system's achievement of benchmarks for performance/reliability/security/robustness and document the results of these verification and whether the acceptance criteria for the objective has been met. If there is a need to reassess the acceptance criteria for reliability, establish how the criteria will be adjusted and why.

b. Set in place standards triggering either the update or deprovisioning of the model in this phase of the project.

DEPLOYMENT

12 Model Updating or Deprovisioning

Activities for Performance and Reliability

a. Put in place standards to trigger either the update or deprovisioning of the model.

b. Put in place measures to ensure the necessary updates are made (e.g. parameter re-tuning).

c. Ensure deprovisioning is carried out as required.

Activities for Security and Robustness

d. Document the results of verifications for security and robustness acceptance criteria being achieved.

e. If there is a need to reassess the acceptance criteria for this objective, document how this criteria will be adjusted and why.

Activities for all objectives

- Determine model updating or deprovisioning needs based on standards determined in the Safety Self-Assessment and Risk Management (on page 43) template. Conduct updates (i.e. parameter re-tuning) or deprovisioning as required.

Download this template here on the AI Ethics and Governance in practice Platform.

SSA & RM

Safety Self-Assessment and Risk Management Template for Project Name

Date completed: ....

Team members involved: ....

DESIGN

1 Project Planning

Step 1

Review the **Safety Assurance Activities** (on page 32) for this lifefcycle stage.

Step 2

Consult previous safety cases for similar technologies and work with subject matter experts to scope possible risks to AI safety objectives.

1. Additional considerations for Security and Robustness

a. Where the prospective AI technology is replacing an existing algorithmic system or a human, or is part of another technical platform or hybrid system that serves the same or similar function, are there any disadvantages that could arise because of the replacement (e.g. less explainability)?

....

- What, if any, risks could arise because of any such trade-offs related to the replacement?

....

b. To what extent has your team considered the risks of procured pre-trained models of poor quality?

....

c. To what extent has your team considered the risks of not having relevant expertise supporting model safety over its lifecycle?

....

d. To what extent could the system cause physical, psychological, or moral harm from its operations?

....

2. Safety risks identified for:

- **Performance**

....

- **Reliability**

....

- **Security**

....

- **Robustness**

....

Step 3

Establish proportional safety protocols for this phase of the project lifecycle to mitigate against risks and support the achievement of safety objectives.

1. Action taken to address, mitigate, or manage identified risks:

....

2. Documentation of actions taken:

....

DESIGN

2 Problem Formulation

Step 1

Review the **Safety Assurance Activities** (on page 32) for this lifefcycle stage.

Step 2

Consult previous safety cases for similar technologies and work with subject matter experts to scope possible risks to AI safety objectives.

1. Additional considerations for Performance and Reliability

a. What, if any, risks could arise if the system fails to meet acceptable thresholds of accuracy, performance, and reliability?

2. Additional considerations for Security and Robustness

a. What, if any, risks could arise if the system fails to meet acceptable thresholds of performance and reliability?

b. To what extent are there risks of the acceptance criteria for security not being fully embedded in the system's specifications?

3. **Safety risks identified for:**

- **Performance**
- **Security**
- **Reliability**
- **Robustness**

Step 3

Establish proportional safety protocols for this phase of the project lifecycle to mitigate against risks and support the achievement of safety objectives.

1. **Action taken to address, mitigate, or manage identified risks:**

2. **Documentation of actions taken:**

DESIGN

3 Data Extraction or Procurement

Step 1

Review the **Safety Assurance Activities** (on page 32) for this lifefcycle stage.

Step 2

Consult previous safety cases for similar technologies and work with subject matter experts to scope possible risks to AI safety objectives.

1. **Additional considerations for Performance and Reliability**

   a. To what extent do the quality and integrity of the data that are being used to train, test, and validate the prospective system pose risks to its accuracy, performance, and reliability?

2. **Safety risks identified for:**

- **Performance**
- **Security**
- **Reliability**
- **Robustness**

Step 3

Establish proportional safety protocols for this phase of the project lifecycle to mitigate against risks and support the achievement of safety objectives.

1. **Action taken to address, mitigate, or manage identified risks:**
2. **Documentation of actions taken:**

DESIGN

4 Data Analysis

Step 1

Review the **Safety Assurance Activities** (on page 32) for this lifefcycle stage.

Step 2

Consult previous safety cases for similar technologies and work with subject matter experts to scope possible risks to AI safety objectives.

1. **Additional considerations for Performance and Reliability**

a. What, if any, risks could arise if the data are not fit-for-purpose and not of appropriate relevance, timely, complete or representative and of sufficient quantity to meet your accuracy, performance, and reliability needs?

2. **Safety risks identified for:**

- **Performance**
- **Security**
- **Reliability**
- **Robustness**

Step 3

Establish proportional safety protocols for this phase of the project lifecycle to mitigate against risks and support the achievement of safety objectives.

1. **Action taken to address, mitigate, or manage identified risks:**
2. **Documentation of actions taken:**

DEVELOPMENT

5 Preprocessing & Feature Engineering

Step 1

Review the **Safety Assurance Activities** (on page 32) for this lifefcycle stage.

Step 2

Consult previous safety cases for similar technologies and work with subject matter experts to scope possible risks to AI safety objectives.

1. Additional considerations for Performance and Reliability

a. To what extent are there risks to assuring data quality during data pre-processing and feature engineering?

2. Additional considerations for Security and Robustness

a. To what extent is your system at risk to adversarial and backdoor attacks?

b. To what extent is your system at risk for inversion attacks (e.g. adversarial training, architectural modification, regularisation, and data pre-processing manipulation)?

3. **Safety risks identified for:**

- **Performance**

  ....

- **Security**

  ....

- **Reliability**

  ....

- **Robustness**

  ....

Step 3

Re-assess the results of previous SSA risk assessment, define reasonable measurements for each objective, and establish proportional safety protocols for this phase of the lifecycle to mitigate against these risks and support the achievement of safety objectives.

1. **Action taken to address, mitigate, or manage identified risks:**

   ....

2. **Documentation of actions taken:**

   ....

DEVELOPMENT

6 Model Selection & Training

Step 1

Review the **Safety Assurance Activities** (on page 32) for this lifefcycle stage.

Step 2

Consult previous safety cases for similar technologies and work with subject matter experts to scope possible risks to AI safety objectives.

1. Additional considerations for Performance and Reliability

a. To what extent do we risk presenting performance metrics that are not informed by the specific context of the use case and its performance needs (e.g. a system whose effective identification of rare events is more critical than its overall accuracy rate)?

....

b. If using a pre-trained model, to what extent is your system at risk for transfer learning attacks?

....

c. To what extent are there risks to the system's accuracy, performance, and reliability if the algorithmic model(s) or technique(s) intended to be used by the AI system (e.g. which have a non-deterministic, probabilistic, evolving, or dynamic character) prevent or hinder the system's intended functionality?

....

d. To what extent are there risks to the system's accuracy, performance, and reliability if the algorithmic model(s) or technique(s) intended to be used by the AI system are not formalised into specific and checkable design-time requirements (and thus impairs commonly accepted methods of formal verification and validation)?

....

2. **Additional considerations for Security and Robustness**

a. To what extent is your system at risk of backdoor or adversarial attacks?

b. To what extent is your system at risk for inversion attacks (e.g. adversarial training, architectural modification, regularisation, and data pre-processing manipulation)?

c. To what extent is the system at risk when considering robustness?

3. **Safety risks identified for:**

- **Performance**
- **Reliability**
- **Security**
- **Robustness**

Step 3

Assess the success measures of the system's safety, reassessment results of the previous (i.e. in Development) SSA risk assessment, and establish proportional safety protocols for the deployment phase of the lifecycle to mitigate against risks and support the achievement of safety objectives.

1. **Action taken to address, mitigate, or manage identified risks:**

2. **Documentation of actions taken:**

DEVELOPMENT

7 Model Testing & Validation

Step 1

Review the **Safety Assurance Activities** (on page 32) for this lifefcycle stage.

Step 2

Consult previous safety cases for similar technologies and work with subject matter experts to scope possible risks to AI safety objectives.

1. Additional considerations for Performance and Reliability

a. To what extent is your model at risk of overfitting?

....

b. To what extent is there a risk of the runtime environment operating differently than intended?

....

c. To what extent are there risks of the model performing unreliably within and between each group?

....

d. To what extent is our runtime environment at risk of being less controllable and more unpredictable?

....

e. To what extent are there risks of the performance metrics not being informed by the specific context of the use case and by the potential effects of differential error rates on affected sub-populations (in particular, on vulnerable or protected groups)?

....

f. To what extent is there a risk of the performance metrics not being presented in a clear and accessible manner in plain, non-technical language?

....

g. Have any standards set to trigger the system's update or deprovisioning been breached?

....

2. Additional considerations for Security and Robustness

a. To what extent are there risks in not having appropriate or adequate or timely hardening techniques, such as adversarial training, to minimise vulnerability and inversion attacks on the model?

b. To what extent are there risks in not having appropriate or adequate or timely monitoring and testing of the model for validation across a wide range of environments to ensure robustness?

c. To what extent are there risks in not meeting benchmarks for security/robustness?

d. Have any standards which are set to trigger the system's update or deprovisioning been breached?

3. Safety risks identified for:

- **Performance**
- **Reliability**
- **Security**
- **Robustness**

Step 3

Assess the success measures of the system's safety, reassessment results of the previous (i.e. in Development) SSA risk assessment, and establish proportional safety protocols for the deployment phase of the lifecycle to mitigate against risks and support the achievement of safety objectives.

1. Action taken to address, mitigate, or manage identified risks:

2. Documentation of actions taken:

DEVELOPMENT

8 Model Reporting

Step 1

Review the **Safety Assurance Activities** (on page 32) for this lifefcycle stage.

Step 2

Consult previous safety cases for similar technologies and work with subject matter experts to scope possible risks to AI safety objectives.

1. Additional considerations for Security and Robustness

a. To what extent is the system at risk for revealing sensitive data to non-trusted third parties?

b. To what extent is the system at risk of novel adversarial attacks?

c. To what extent could the system operate autonomously without the need for intervention to curtail instances of misdirected reinforcement learning?

2. Safety risks identified for:

- **Performance**
- **Reliability**
- **Security**
- **Robustness**

Step 3

Establish proportional safety protocols for this phase of the project lifecycle to mitigate against risks and support the achievement of safety objectives.

1. **Action taken to address, mitigate, or manage identified risks:**

   ....

2. **Documentation of actions taken:**

   ....

DEPLOYMENT

9 System Implementation

Step 1

Review the **Safety Assurance Activities** (on page 32) for this lifefcycle stage.

Step 2

Consult previous safety cases for similar technologies and work with subject matter experts to scope possible risks to AI safety objectives.

1. **Additional considerations for Performance and Reliability**

   a. To what extent is the system at risk of encountering novel data in the runtime environment?

   ....

   b. To what extent are there possible risks resulting from model brittleness (e.g. inability to respond to ‘unseen’ inputs not contextualise problems)?

   ....

2. Additional considerations for Security and Robustness

a. To what extent is the system at risk for revealing sensitive data to non-trusted third parties?

b. To what extent is the system at risk to novel adversarial attacks?

c. To what extent could the system operate autonomously without need for intervention to curtail instances of misdirected reinforcement learning?

3. Safety risks identified for:

- **Performance**
- **Reliability**
- **Security**
- **Robustness**

Step 3

Assess the success measures of the system's safety, reassessment results of the previous (i.e. in Development) SSA risk assessment, and establish proportional safety protocols for the deployment phase of the lifecycle to mitigate against risks and support the achievement of safety objectives.

1. **Action taken to address, mitigate, or manage identified risks:**

2. **Documentation of actions taken:**

DEPLOYMENT

10 User Training

Step 1

Review the **Safety Assurance Activities** (on page 32) for this lifefcycle stage.

Step 2

Consult previous safety cases for similar technologies and work with subject matter experts to scope possible risks to AI safety objectives.

1. Additional considerations for Performance and Reliability

a. To what extent are the implementers and users of the system at risk of harms that adversely impact their dignity, autonomy, and ability to make free, independent, and well-informed judgements?

b. To what extent could the system cause physical, psychological, or moral harm from its operations?

2. Safety risks identified for:

- Performance
- Reliability
- Security
- Robustness

Step 3

Assess the success measures of the system’s safety, reassessment results of the previous (i.e. in Development) SSA risk assessment, and establish proportional safety protocols for the deployment phase of the lifecycle to mitigate against risks and support the achievement of safety objectives.

1. **Action taken to address, mitigate, or manage identified risks:**

   ....

2. **Documentation of actions taken:**

   ....

DEPLOYMENT

11 System Use & Monitoring

Step 1

Review the **Safety Assurance Activities** (on page page 32) for this lifefcycle stage.

Step 2

Consult previous safety cases for similar technologies and work with subject matter experts to scope possible risks to AI safety objectives.

1. **Additional considerations for Performance and Reliability**

   a. To what extent is the system at risk if achieving acceptance criteria for accuracy, performance, and reliability are not met?

   ....

   b. To what extent is the system at risk of concept drift and shifts in underlying data distributions?

   ....

2. Additional considerations for Security and Robustness

a. To what extent is the system at risk if achieving the acceptance criteria for security and robustness is not met?

b. To what extent are there risks of sensitive data being revealed over time in the runtime environment?

c. Have any standards which are set to trigger the system's update or deprovisioning been breached?

3. Safety risks identified for:

- Performance
- Reliability
- Security
- Robustness

Step 3

Assess the success measures of the system's safety, reassessment results of the previous (i.e. in Development) SSA risk assessment, and establish proportional safety protocols for the deployment phase of the lifecycle to mitigate against risks and support the achievement of safety objectives.

1. Action taken to address, mitigate, or manage identified risks:

2. Documentation of actions taken:

DEPLOYMENT

12 Model Updating or Deprovisioning

Step 1

Review the **Safety Assurance Activities** (on page 32) for this lifefcycle stage.

Step 2

Consult previous safety cases for similar technologies and work with subject matter experts to scope possible risks to AI safety objectives.

1. Additional considerations for Performance and Reliability

a. To what extent is the system at risk for missing triggers to update or de-provision?

b. To what extent are there risks of not undertaking appropriate or adequate updating or deprovisioning of the system?

2. Additional considerations for Security and Robustness

a. To what extent is the system at risk if achieving the acceptance criteria for security and robustness are not met?

b. Have any standards which are set to trigger the system's update or deprovisioning been breached?

3. **Safety risks identified for:**

- **Performance**

  ………………………………………………

- **Security**

  ………………………………………………

- **Reliability**

  ………………………………………………

- **Robustness**

  ………………………………………………

Step 3

Assess the success measures of the system's safety, reassessment results of the previous (i.e. in Development) SSA risk assessment, and establish proportional safety protocols for the deployment phase of the lifecycle to mitigate against risks and support the achievement of safety objectives.

1. **Action taken to address, mitigate, or manage identified risks:**

   ………………………………………………

2. **Documentation of actions taken:**

   ………………………………………………

AI Safety in Practice

Activities

Activities Overview

In the previous sections of this workbook, we have presented the various elements of technical sustainability or safety. In this section we provide concrete tools for applying these concepts in practice. Activities will help participants to build a common vocabulary around the objectives of AI safety, enhance their understanding of AI safety risks, and take actions throughout the AI/ML project lifecycle to ensure that AI safety objectives are achieved.

We offer a collaborative workshop format for team learning and discussion about the concepts and activities presented in the workbook. To run this workshop with your team, you will need to access the resources provided in the link below. This includes a digital board and printable PDFs with case studies and activities to work through.

Workshop resources for AI Safety in Practice

A Note on Activity Case Studies

Case studies within the Activities sections of the AI Ethics and Governance in Practice workbook series offer only basic information to guide reflective and deliberative activities. If activity participants find that they do not have sufficient information to address an issue that arises during deliberation, they should try to come up with something reasonable that fits the context of their case study.

Note for Facilitators

In this section, you will find the participant and facilitator instructions required for delivering activities corresponding to this workbook. Where appropriate, we have included considerations to help you navigate some of the more challenging activities.

Activities presented in this workbook can be combined to put together a capacity-building workshop or serve as stand-alone resources. Each activity corresponds to a section within the Key Concepts in this workbook. Some activities have prerequisites, which are detailed on the following page.

We sometimes provide ideas of how a **co-facilitator** can help manage large groups.

Conceptualising AI Safety

Build a common vocabulary and understanding of AI Safety objectives by reflecting on how participants would define these and discuss the definitions we share in this workshop.

Corresponding Sections

→ A Closer Look at AI Safety Objectives (page 15)

Idenitifying AI Safety Risks

Enhance understanding of AI safety risks and mitigation strategies in the public sector.

Corresponding Sections

→ A Closer Look at AI Safety Objectives (page 15)

→ Risks Posed to Performance and Reliability (page 21)

→ Risks Posed to Security and Robustness (page 26)

Safety Self-Assessment

Recognise safety considerations at relevant stages of the project lifecycle.

Corresponding Sections

→ Part Two: Putting AI Safety into Practice (page 30)

45 mins | Participant Instructions

Conceptualising AI Safety

Objective

Build a common vocabulary and understanding of AI Safety objectives by reflecting on how participants would define these and discuss the definitions we share in this workshop.

Part One: Building a Common Vocabulary

1. Individually, take a moment to read over the activity instructions.
2. Depending on the team size, you will split into groups.
3. Next, take a few minutes to think about what your understanding of each of the four safety objectives is and then write/ add post it notes in your groups section that answers to the following questions:
   - What is your understanding of reliability, performance, security, and robustness?
   - How do you interpret these safety objectives?
4. When the assigned time for individual reflection is up, your facilitator will introduce the definitions shared in this workbook.
5. After your facilitator introduces the definitions shared in this workbook, have a group discussion about how your initial understanding matched the definitions of the AI safety objectives in this workbook.

Part Two: Exemplifying Safety Objectives

1. Your team will be split into groups.
2. Each group will have a discussion about how the four safety objectives are pertinent to your work.
3. Have a volunteer write notes about your group discussion in your group's section. Consider the following questions:
   - Are you applying or will you apply any of these objectives? Are there any objectives you might prioritise over others?
   - Can you think of an example from a project that demonstrated any of these objectives?
4. Having had a discussion, reconvene as a team, having volunteers share each group's insights.

Activity 2 | 45 mins | Facilitator Guidance

Contextualising AI Safety

Part Two: Building a Common Vocabulary

1. Give participants a moment to read over the activity instructions, answering any questions.

2. Depending on the team size, split the team into groups.

3. Let the team know that they will have a couple of minutes for individual reflection about their understanding of each of the four safety objectives.

4. When time is up, ask the team to reconvene and introduce the definitions shared in this workbook.

5. After you introduce the definitions of the safety objectives shared in this workbook, facilitate a group discussion about how participants' initial understanding matched the definitions of the AI safety objectives in this workbook.

Part Two: Exemplifying Safety Objectives

1. Split the team into groups. Each group will have a discussion about how the four safety objectives are pertinent to their work.
    - **Facilitators** and **co-facilitators** should join groups and support their discussions.

2. Let the groups know that they will have XX minutes to discuss, considering the following questions:
    - Are you applying or will you apply any of these objectives? Are there any objectives you might prioritise over others?
    - Can you think of an example from a project that demonstrated any of these objectives?

3. When time is up, ask the team to reconvene.

4. Give each volunteer note-taker a few minutes to share their group's discussion.

5. Next, facilitate a team discussion about the AI Safety Objectives. Consider the following questions to prompt discussion:
    - What objectives did groups find most important in their examples? Why?
    - What are the similarities and differences between objectives?
    - What are the benefits of achieving all of the safety objectives within AI projects?

6. **Co-facilitators** may take notes about this discussion in sticky notes, placing them in the Team Discussion section.

60 mins | Participant Instructions

Identifying AI Safety Risks

Objective

Enhance understanding of AI safety risks and mitigation strategies in the public sector. Participants will analyse case studies of AI in the public sector, brainstorm potential risks, and propose measures to ensure safety objectives are met.

Part One: Mapping Risks

1. Individually, take a moment to read over the activity instructions.
2. Your team will be divided into groups, each assigned an example use of AI in the public sector. Take a few minutes to review your assigned case study.
3. In your groups, engage in a discussion about your assigned use case. Consider each safety objective and discuss potential safety risks that would prevent your team from meeting these objectives.
   - Have a volunteer within your group take notes of your discussion within your group's section of the board.
4. Reconvene as a team to share insights. Each group's note-taker will present their assigned use case and the group's discussion points.
5. Next, have a team-wide discussion on AI safety risks, drawing from the shared insights.

Part Two: Responding to Changing Scenarios

1. Your team will once again divide into the same groups to further analyse their assigned case study.
2. Each group will receive a scenario related to their case study for analysis. Take a few minutes to read over the provided scenario.
3. Within your groups, discuss the vulnerabilities of the AI project to safety risks outlined in the scenario. Consider the following questions:
   - How is this particular AI project vulnerable to safety risks associated with the changes described? To what extent is the system at risk?
   - If there are identified risks, what measures could be put in place to ensure that the safety objectives are met? Refer to potential actions to take and document the practices.
4. Have a volunteer within your group take notes pertaining to your discussion of the objectives within your group's section of the board.
5. Once time is up, regroup as a team to share findings. Each group's note-taker will present the group's discussions.

60 mins | Facilitator Instructions

Identifying AI Safety Risks

Part One: Mapping Risks

1. Give participants a moment to read over the activity instructions, answering any questions.

2. Next, split the team into groups, each assigned an example of AI in the public sector.

   Group 1: An AI system uses real-time video data collected from traffic light cameras to predict traffic flows and dynamically control traffic signals timings to optimise the flow of vehicles. By detecting the movements of vehicles, identifying pedestrians, and recognising traffic patterns, the AI system predicts the most effective timing for traffic signal changes at intersections

   Group 2: A local foundation trust is planning to use a combination of electronic health record data and patient symptoms to predict a patients' risk level in emergency rooms to inform individual wait times.

3. Let the group know that they will have 20 minutes to consider their use case and discuss potential safety risks that would prevent their team from meeting the safety objectives.
   - **Facilitators** and **co-facilitators** should join groups and support their discussions.

4. Let the team know when they have 5 minutes left to discuss.

5. When the 20 minutes have passed, ask the team to reconvene.

6. Give each volunteer note-taker a few minutes to share their group's use case and discussion.

7. Next, facilitate a team discussion about AI safety risks. Consider the following questions to prompt discussion:
   - What are some common themes or patterns in the safety risks identified across the case studies?
   - What are the ethical considerations in deploying AI systems in these scenarios, and how do they relate to safety?
   - Can you think of any real-world examples where similar AI systems faced safety challenges?

8. Co-facilitators are to take notes about this discussion in sticky notes, placing them in the **Team Discussion** section of the board.

Part Two: Responding to Changing Scenarios

1. Split the team into the same groups and provide each group with a scenario related to their case study for analysis.

   Group 1: After the AI system was deployed, the Highway Code (2022) was updated. It now states clearly that, at a junction, you should give way to pedestrians crossing or waiting to cross a road that you're turning into. Previously, vehicles had priority at a junction. The updated code clarifies that:

   - when people are crossing or waiting to cross at a junction, other traffic should give way;
   - if people have started crossing and traffic wants to turn into the road, the people crossing have priority and the traffic should give way; and
   - people driving, riding a motorcycle or cycling must give way to people on a zebra crossing and people walking and cycling on a parallel crossing.

   A parallel crossing is similar to a zebra crossing, but includes a cycle route alongside the black and white stripes.

   Group 2: After the AI system to predict patients' risk levels in emergency rooms was deployed, it was discovered that an unknown actor had infiltrated the system and intentionally injected false patient symptom data into the training dataset used by the AI. As a result, the AI system began to make inaccurate risk predictions, leading to potentially harmful consequences for patients.

   - **Facilitators** and **co-facilitators** should join groups and support their discussions.

2. Let the groups know that they will have 15 minutes to discuss, considering the following questions:

   - How is this particular AI project vulnerable to safety risks associated with the changes described? To what extent is the system at risk?
   - How do the risk(s) identified compromise the safety objectives of the AI system?
   - What measures could be put in place to ensure that the safety objectives are met? Refer to potential actions to take and documentation practices.

3. When time is up, ask the team to reconvene.

4. Give each volunteer note-taker a few minutes to share their group's discussion.

5. **Co-facilitators** may take notes about this discussion in sticky notes, placing them in the **Team Discussion** section.

30 mins | Participant Instructions

Safety Self-Assessment

Objective

Recognise safety considerations at relevant stages of the AI/ML project lifecycle.

Team Instructions

1. Individually, take a moment to thoroughly read over the activity instructions.
2. Your team will be divided into groups. If you had carried out the activity Identifying AI Safety Risks, you will be divided into the same groups.
3. In your groups, have a discussion about the safety considerations associated with your case study and place each safety consideration marker on the stage(s) of the AI project lifecycle where it is most relevant.
    - Have a volunteer within your group take notes of your discussion within your group’s section of the board.
4. As you talk through each safety consideration, discuss:
    - Where is the chosen consideration likely to have the most impact on project activities?
    - Where could actions be taken to most effectively mitigate this impact?
5. Once a consideration is placed on the project lifecycle, add a second sticky note with information on:
    - Any possible impacts of the safety consideration on your project.
    - Actions that could be taken to mitigate the impact of this consideration.
6. Reconvene as a team to share insights. Each group’s note-taker will present the group’s discussion points.
7. Next, have a team-wide discussion on the Safety Self-Assessment, drawing from the shared insights.

Notes

- Each group should engage in discussion to determine the placement and details for each safety consideration.
- ”Relevant” means a stage where the consideration is readily identifiable or can be effectively mitigated. Effects of a consideration can cascade through downstream stages.
- There isn’t a single right answer as many considerations may impact various stages across the AI/ML project lifecycle.

Activity 3 | 30 mins | Facilitator Guidance

Safety Self-Assessment

1. Give participants a moment to read over the activity instructions, answering any questions.

2. Next, split the team into the same groups from the Identifying AI Safety Risks activity. If the activity is done as a stand-alone activity, split the team into groups, each assigned an example of AI in the public sector.

   Group 1: Traffic flow analysis. An AI system uses real-time video data collected from traffic light cameras to predict traffic flows and dynamically control traffic signals timings to optimise the flow of vehicles. By detecting the movements of vehicles, identifying pedestrians, and recognising traffic patterns, the AI system predicts the most effective timing for traffic signal changes at intersections

   Group 2: A local foundation trust is planning to use a combination of electronic health record data and patient symptoms to predict a patients' risk level in emergency rooms to inform individual wait times.

3. Let the group know that they will have 15 minutes to consider their use case and to discuss safety considerations associated with their case study, along with possible impacts of the safety consideration on the project, and actions that could be taken to mitigate this impact.

4. Let the team know when they have 5 minutes left to discuss.

5. When the 15 minutes have passed, ask the team to reconvene.

6. Give each volunteer note-taker a few minutes to share their group's discussion.

7. Facilitate a team-wide discussion on the Safety Self-Assessment. Encourage participants to draw from the shared insights and group discussions.

   - **Co-facilitators** may take notes about this discussion in sticky notes, placing them in the Team Discussion section.

Endnotes

1 Lazar, S., & Nelson, A. (2023). AI safety on whose terms?. *Science, 381*(6654), 138-138.

2 Adamson, A. S., & Smith, A. (2018). Machine learning and healthcare disparities in dermatology. *JAMA Dermatology, 154*(11), 1247. https://doi.org/10.1001/jamadermatol.2018.2348

3 Goyal, M., Knackstedt, T., Yan, S., & Hassanpour, S. (2020). Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. *Computers in Biology and Medicine, 127,* 104065. https://doi.org/10.1016/j.compbiomed.2020.104065

4 Kamulegeya, L. H., Okello, M., Bwanika, J. M., Musinguzi, D., Lubega, W., Rusoke, D., Nassiwa, F., & Börve, A. (2019). Using artificial intelligence on dermatology conditions in Uganda: A case for diversity in training data sets for machine learning [Preprint]. *Bioinformatics*, 1–30. https://doi.org/10.1101/826057

5 see Lee, E. W., & Viswanath, K. (2020). Big data in context: addressing the twin perils of data absenteeism and chauvinism in the context of health disparities research. *Journal of medical Internet research, 22*(1), e16377. https://doi.org/10.2196/16377

6 Balagurunathan, Y., Mitchell, R., & El Naqa, I. (2021). Requirements and reliability of AI in the medical context. *Physica medica* (AIFB), 83, 72–78. https://doi.org/10.1016/j.ejmp.2021.02.024

7 Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines. *arXiv preprint* arXiv:1206.6389. https://doi.org/10.48550/arXiv.1206.6389

8 Calo, R. (2016). Privacy, vulnerability, and affordance. *DePaul Law Review,* 66, 592–593. https://via.library.depaul.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=4023&context=law-review

9 Malgieri, G., & Niklas, J. (2020). Vulnerable data subjects. *Computer Law & Security Review,* 37, 105415. https://doi.org/10.1016/j.clsr.2020.105415

10 Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Lu, M. Y., Sahai, S., & Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering, 7*(6), 719–742. https://doi.org/10.1038/s41551-023-01056-8

11 Barmer, H., Dzombak, R., Gaston, M., Heim, E., Palat, V., Redner, F., ... & VanHoudnos, N. (2021). Robust and Secure AI. https://insights.sei.cmu.edu/documents/609/2021_019_001_735346.pdf

12 John, M.M., Olsson, H.H., and Bosch J. 2021. Towards MLOps: A Framework and Maturity Model. 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), Palermo, Italy, 2021, pp. 1-8, https://doi.org/10.1109/SEAA53835.2021.00050

13 Vela, D., Sharp, A., Zhang, R. et al. 2022. Temporal quality degradation in AI models. Sci Rep 12, 11654. https://doi.org/10.1038/s41598-022-15245-z

14 Story, M., & Congalton, R. G. (1986). Accuracy assessment: a user's perspective. *Photogrammetric Engineering and remote sensing, 52*(3), 397-399. https://www.asprs.org/wp-content/uploads/pers/1986journal/mar/1986_mar_397-399.pdf

15 Buckland, M., & Gey, F. (1994). The relationship between recall and precision. *Journal of the American society for information science, 45*(1), 12-19. https://doi.org/10.1002/(SICI)1097-4571(199401)45:1%3C12::AID-ASI2%3E3.0.CO;2-L

16 Leslie, D. (2019). Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector. Available at SSRN 3403301.

17 Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L., & Petitjean, F. (2016). Characterizing concept drift. *Data Mining and Knowledge Discovery, 30*(4), 964-994.

18 Delacroix, S. (2022). Diachronic Interpretability & Machine Learning Systems. *Journal of Cross-disciplinary Research in Computational Law*.

19 Bayram, F., Ahmed, B. S., & Kassler, A. (2022). From concept drift to model degradation: An overview on performance-aware drift detectors. *Knowledge-Based Systems*, 245, 108632.

20 Gama, J., Medas, P., Castillo, G., & Rodrigues, P. (2004). Learning with drift detection. In Advances in Artificial Intelligence–SBIA 2004: 17th Brazilian Symposium on Artificial Intelligence, Sao Luis, Maranhao, Brazil, September 29-Ocotber 1, 2004. Proceedings 17 (pp. 286-295). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-28645-5_29

21 Fields, T., Hsieh, G., & Chenou, J. (2019, December). Mitigating drift in time series data with noise augmentation. In 2019 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 227-230). IEEE. https://doi.org/10.1109/CSCI49370.2019.00046

22 Leslie, D., Katell, M., Aitken, M., Singh, J., Briggs, M., Powell, R., ... & Burr, C. (2022). Data Justice in Practice: A Guide for Developers. Available at SSRN 4080058.

23 Symbal, A. (2004). The problem of concept drift: definitions and related work. Computer Science Department, Trinity College Dublin, 106(2), 58. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=

24 Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine,* 17, 1-9. https://doi.org/10.1186/s12916-019-1426-2

25 Cummings, M. (2021). Rethinking the maturity of artificial intelligence in safety-critical settings. AI Magazine, 42(1), 6-15. https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/7394

26 Cummings, M. (2021). Rethinking the maturity of artificial intelligence in safety-critical settings. *AI Magazine,* 42(1), 6-15.

27 Cummings, M. (2021). Rethinking the maturity of artificial intelligence in safety-critical settings. *AI Magazine,* 42(1), 6-15.

28 see Rice, L., Wong, E., & Kolter, Z. (2020, November). Overfitting in adversarially robust deep learning. In International conference on machine learning (pp. 8093-8104). PMLR. http://proceedings.mlr.press/v119/rice20a

29 Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM computing surveys* (CSUR), 27(3), 326-327. https://dl.acm.org/doi/pdf/10.1145/212094.212114

30 Bellmann, R. (1961). *Adaptive Control Processes: A Guided Tour.* Princeton University Press.

31 AWS (n.d.). What is overfitting? https://aws.amazon.com/what-is/overfitting/

32 IBM (2024). What is dimensionality reduction? https://www.ibm.com/topics/dimensionality-reduction

33 Ying, X. (2019). An Overview of Overfitting and its Solutions. In Journal of Physics: Conference Series, 1168(2). https://doi.org/10.1088/1742-6596/1168/2/022022

34 Dobbin, K. K., & Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC medical genomics,* 4, 1-8. https://doi.org/10.1186/1755-8794-4-31

35 Piantadosi, G., Marrone, S., Sansone, C. (2019). On Reproducibility of Deep Convolutional Neural Networks Approaches. In: Kerautret, B., Colom, M., Lopresti, D., Monasse, P., Talbot, H. (eds) *Reproducible Research in Pattern Recognition.* Lecture Notes in Computer Science, vol 11455. Springer, Cham. https://doi.org/10.1007/978-3-030-23987-9_10

36 Leslie, D. (2019). Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector. Available at SSRN 3403301.

37 Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security (pp. 1322-1333).

38 Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I., & Tygar, J. D. (2011, October). Adversarial machine learning. In Proceedings of the 4th ACM workshop on Security and artificial intelligence (pp. 43-58). https://doi.org/10.1145/2046684.2046692

39 Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint* arXiv:1706.06083. https://doi.org/10.48550/arXiv.1706.06083

40 Athalye, A., Carlini, N., & Wagner, D. (2018, July). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In International conference on machine learning (pp. 274-283). PMLR.

41 National Cyber Security Centre. (2023). *Guidelines for secure AI system development.* https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development/guidelines/secure-design

42 The National Cyber Security Centre has developed guidelines for secure AI development: These guidelines are for providers of any systems that use AI, whether those systems have been created from scratch or built on top of tools and services provided by others. The guidelines are endorsed by Australia, Canada, New Zealand, United Kingdom, and the United States. See: https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development

43 MITRE Atlas - https://atlas.mitre.org. MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems) is a globally accessible and living knowledge resource of adversary tactics and techniques based on real-world attack observations and realistic demonstrations provided by AI red teams and security groups.

44 NIST Adversarial Machine Learning (Report - January 2024). This NIST report on AI develops a taxonomy of attacks and mitigations and defines terminology in the field of adversarial machine learning. See: https://csrc.nist.gov/pubs/ai/100/2/e2023/final

45 Leslie, D. (2019). Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector. Available at SSRN 3403301.

46 Li, B., Wang, Y., Singh, A., & Vorobeychik, Y. (2016). Data poisoning attacks on factorization-based collaborative filtering. *Advances in neural information processing systems,* 29. https://doi.org/10.48550/arXiv.1608.08182

47 Rubinstein, B. I., Nelson, B., Huang, L., Joseph, A. D., Lau, S. H., Rao, S., Taft, N., & Tygar, J. D. (2009). Antidote: understanding and defending against poisoning of anomaly detectors. In Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement (pp. 1-14). https://doi.org/10.1145/1644893.1644895

48 Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., ... & Tramèr, F. (2024, February). Poisoning Web-Scale Training Datasets is Practical. In *2024 IEEE Symposium on Security and Privacy (SP)* (pp. 176-176). IEEE Computer Society.

49 Baracaldo, N., Chen, B., Ludwig, H., & Safavi, J. A. (2017). Mitigating poisoning attacks on machine learning models: A data provenance based approach. In Proceedings of the 10th ACM workshop on artificial intelligence and security (pp. 103-110). https://doi.org/10.1145/3128572.3140450

50 Goldblum, M., Tsipras, D., Xie, C., Chen, X., Schwarzschild, A., Song, D., Madry, A., Li, B., & Goldstein, T. (2020). Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses. ArXiv. https://arxiv.org/pdf/2012.10544

51 Kurita, K., Michel, P., & Neubig, G. (2020). Weight Poisoning Attacks on Pre-trained Models. Annual Conference of the Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2004.06660

52 Leslie, D. (2019). Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector. Available at SSRN 3403301.

To find out more about the AI Ethics and Governance in Practice Programme please visit:

aiethics.turing.ac.uk

Version 1.2

The Alan Turing Institute