

The Alan Turing Institute

Process Based Governance in Action

A guide to ethical and responsible design,
development, and deployment
of Artificial Intelligence

Prepared for the Department for Business & Trade by The Alan Turing Institute
October 2023



Table of Contents

FOREWORD BY JASON KITCAT, DEPARTMENT FOR BUSINESS AND TRADE	I
PREFACE	II
BRIEF BACKGROUND	III
SECTION 1.1: PROCESS-BASED GOVERNANCE	1
AI ETHICS	1
THE SSAFE-D PRINCIPLES	3
THE AI PROJECT LIFECYCLE	6
SECTION 1.2 - THE PBG FRAMEWORK	10
SCOPING AND ANTICIPATORY REFLECTION.....	13
THE PROCESS LOG	16
PROJECT SUMMARY REPORT (PS REPORT).....	18
SECTION 2 - DELIBERATION AND ENGAGEMENT	28
FROM PRINCIPLES TO CORE ATTRIBUTES	28
STAKEHOLDER ENGAGEMENT PROCESS (SEP).....	33
STAKEHOLDER IMPACT ASSESSMENTS.....	45
READINESS SELF-ASSESSMENT	47
SECTION 3: DATA PROTECTION AND INTELLECTUAL PROPERTY CONSIDERATIONS	48
DATA PROTECTION CONSIDERATIONS.....	48
INTELLECTUAL PROPERTY CONSIDERATIONS.....	49
RELATED CONSIDERATIONS	50
DATA PROTECTION AND INTELLECTUAL PROPERTY GOVERNANCE ACTIONS	51
SECTION 4. ACTION AND DECISION-MAKING.....	52
USING THE PROJECT LIFECYCLE MODEL	52
MAP GOVERNANCE WORKFLOW	63
SECTION 5: ONGOING GOVERNANCE	65
MECHANISMS FOR MONITORING, EVALUATION, AND COMMUNICATION	65
SECTION 6: WORKED EXAMPLES	69
APPENDIXES	77
APPENDIX A: GLOSSARY	78
APPENDIX B: PROCESS-BASED GOVERNANCE LOG TEMPLATE.....	87
APPENDIX C: PROJECT SUMMARY REPORT TEMPLATE	88
APPENDIX D: DATA FACTSHEET	93
APPENDIX E: PROCUREMENT GUIDANCE TOOL.....	96
APPENDIX F: CONTEXT-BASED RISK ASSESSMENT (COBRA) WORKSHEET	103
APPENDIX H: READINESS SELF-ASSESSMENT	110
APPENDIX I: SSAFE-D PRINCIPLES CORE ATTRIBUTES	121
APPENDIX J: BIAS SELF-ASSESSMENT	127
SELECTED BIBLIOGRAPHY	164
END NOTES.....	167



Foreword from Jason Kitcat, Director of Digital, Data and Technology, Department for Business and Trade

Public services have long looked to technology and innovation as ways to improve our efficiency and efficacy in delivering for the public. To that end, machine learning, natural language processing, large language models and other approaches generally labelled as “Artificial Intelligence”, are rightly being explored. However, as they become more complex, and potentially unpredictable, we as public servants have a duty to ensure we consider the risks as well as benefits in holistic, meaningful ways. Accuracy, fairness and openness really matter in public services. Government’s processes, decision-making and service delivery should be rigorous, free from bias and open to challenge to avoid maladministration.

Moving from fairly predictable techniques to the much harder to fathom workings of systems like large language models forces us to consider how we hold ourselves to the high ethical standards that the public rightly demand of us. To that end we’ve been delighted to work with the Alan Turing Institute to formulate this framework for assessing how we decide the risks and benefits of adopting AI technologies. This is a challenging, emerging area of work filled with huge amounts of hype and misunderstanding. I think the approach set out here is thoughtful, rigorous and rooted in public service values. I hope you find it useful in formulating your own approaches to governing the use of AI.

Preface



Welcome to *Process Based Governance in Action*, a practical guide to ethical and responsible design, development, and deployment of artificial intelligence. This governance framework was produced by The Alan Turing Institute for the Department for Business & Trade to support departmental management and staff in the ethical adoption and use of AI and related data-driven technologies. This document is intended to support your organisation with an approach to AI ethics called the **Process-Based-Governance Framework**, which is an established method for applying principles of AI ethics and safety to the design, development, and deployment of algorithmic systems.¹ The guidance outlines how AI project teams can put ethical values and practical principles into practice across the AI project lifecycle—beginning with the decision to use a data-driven system through its adoption and ultimate retirement—ensuring that AI is used ethically, safely, and responsibly.

Brief Background

The National AI Strategy encourages ministries and departments to responsibly incorporate AI as part of an overall effort to improve public services. Following the advice of the Government Internal Audit Agency and in light of recent developments in the field of large language models (LLMs) and generative AI (GenAI), DBT leadership is motivated to improve and clarify its governance over AI technologies. *Process Based Governance in Action* is an ethical framework designed to support responsible innovation by DBT personnel who design, develop, or deploy AI in their daily work. This includes AI system developers and procurers, users, and anyone tasked with making strategic decisions about AI at DBT.

Who We Are

The Public Policy Programme at The Alan Turing Institute develops research, tools, and techniques that help governments innovate with data-intensive technologies and improve the quality of people's lives. We work alongside policy makers 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.

Additional Resources

Process Based Governance in Action was designed for the Department for Business and Trade is complemented by the [AI Ethics and Governance in Practice Programme](#), developed by the Turing Public Programme to equip the public sector with tools, training, and support for adopting the PBG framework and carrying out projects in line with state-of-the-art practices in responsible and trustworthy AI innovation. The Programme features a set of eight workbooks that provides guidance and activities for implementing the components of the PBG Framework.²

Another resource we recommend is the [Turing Commons](#), a home for resources and tools to help you reflect, discuss, and take responsibility for the design, development, and use of data-driven technologies. The Turing Commons includes guidebooks, activities, case studies, blog posts, and more.

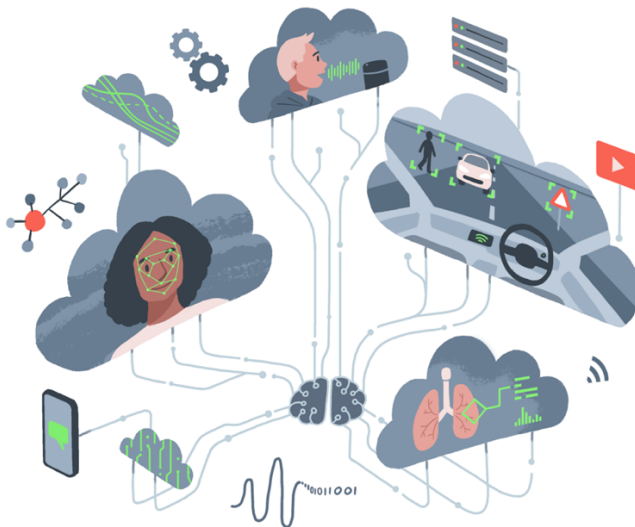
Acknowledgements

This project was supervised by Michael Katell and received tremendous support from members of the Turing's Ethics and Responsible Innovation Team, including Ann Borda, Semeli Hadjiloizou, Smera Jayadeva, Sabeedah Mahomed, and Anto Perini. Illustrations in this guide were designed for The Alan Turing Institute by Conor Rigby.

Section 1.1: Process-Based Governance

In this chapter, we introduce the Process-Based Governance (PBG) framework, which is a framework for developing strategies and producing documentation to demonstrate the work of a project team's ethical reflection and deliberation about an AI solution. We begin with some fundamental concepts before moving on to the details of the framework itself.

AI Ethics



Artificial intelligence (AI) is a powerful suite of tools and techniques for automating tasks and analysing complex data in almost any format.³ The remarkable capabilities and desired efficiencies of AI are amongst the reasons the National AI Strategy encourages the use of AI by government to support decision-making and improve service delivery. However, all technology use can have downstream social effects that must be considered for responsible development and use. The capabilities of AI that enable it to be a participant in important decisions coupled to the challenges of understanding how it functions amplify these effects, leading to concerns about the potential for AI use to contribute to various forms of harm, including bias and discrimination. The errors and inaccuracies of AI can be harder to immediately detect or

AI ethics is a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies.

solve, placing users at increased risk of liability and reputational harm. Securing and maintaining public trust is essential for the operation of government in a democratic society and a solemn duty for every public sector employee.

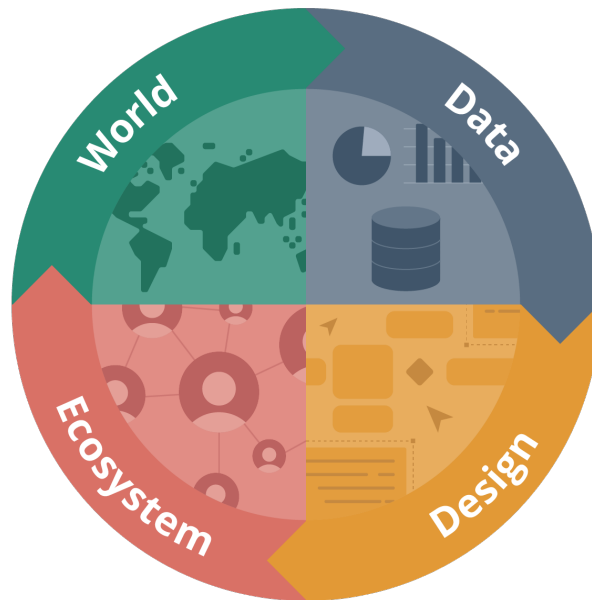
As AI is both a rapidly evolving and powerful technology, it is unavoidable that mistakes and miscalculations will be made and that both unanticipated and harmful impacts will inevitably occur. AI is an exciting yet complex set of technologies that can catch people off guard and result in error and harm. A key strategy for managing these impacts responsibly and to directing the development of AI systems toward optimal public benefit is to implement governance strategies guided by ethical principles.

The elements of this framework are rooted in core concepts from the field of AI ethics, which is the application of moral theory to the domain of data and automation. Ethics can be a powerful tool for making decisions that meet our formal and implied obligations to society and the biosphere in a manner that supports a thriving world. While the tools of ethics may not provide specific answers to every question or ensure that every decision made will be the very best one, they can provide accessible strategies for thinking through issues and reaching thoughtful and defensible conclusions. A key goal of AI ethics is to demonstrate that the humans operating and relying upon AI for insights and decision--support are doing so in a considerate and risk-aware manner, taking the reasonable concerns and interests of everyone into account. Another key goal of AI ethics is to provide a system of accountability that identifies both the parties responsible for important decisions and they key stakeholders whose lives are affected by AI, and providing avenues for meaningful and respectful processes of information sharing, feedback, and recourse.

Why AI ethics is important

Artificial intelligence is a highly impactful technology that is being implemented in an increasing number of use domains and contexts that can have major effects on people's lives and well-being. The predictive and analytical capabilities of AI systems are being implemented in public services to support decision-making, including in high-stakes and sensitive areas, such as to support healthcare decisions and to determine who is eligible for public benefits. In other words, AI is being positioned to participate in an increasing number of relations between individuals and society. However, AI does not enter into a perfected society or do so in a completely neutral way. It is also not a panacea technology that can, on its own, resolve societies many problems including those that are the consequence of legacies of discrimination, inequality, and injustice. AI is a human technology designed to carry out human-defined goals in the midst of human interpersonal and social politics. As such, AI and those who wield it, are ethical "actors" in the human drama. AI ethics is a set of strategies that can illuminate important rights and obligations and help to chart a path to their fulfilment.

We describe the complex relationship between people and their technologies in a four-quadrant model:



The World: AI technologies enter a complicated world shaped by history and culture, and characterised by social struggles including discrimination, inequality, and social strife.

Data: The data processed by AI technologies emerges from human activities, and reflects human biases, beliefs, and preferences. It is also shaped by the people who collect, share, and label it.

Design: The work to design, develop, and implement every AI technology requires numerous human decisions that are informed by a mix of worldviews, objectives, and priorities.

Ecosystem: AI technologies enter into ecosystems of people, markets, other technologies, and political/policy decisions, which shape how each technological system is adopted and its wider effects on the world.

The four-quadrant model is intended to demonstrate how AI, like many impactful technologies, is “socio-technical”, meaning that technologies do not stand apart as neutral and isolated from human activities and dramas. Rather, they are integrated participants in our social world. This understanding of AI underscores the importance of taking the time to reflect and deliberate on the ethics of an AI project from end-to-end; from its inception through its design, development, and implementation until its eventual retirement or replacement.

The SSAFE-D Principles

This guidance is oriented around a set of ethical principles we call the SSAFE-D Principles. The SSAFE-D Principles are a set of ethical principles that serve as starting

points for reflection and deliberation about possible harms and benefits associated with data-driven technologies. As a preparatory step for engaging with the governance of data-driven systems, project teams should become familiar with the principles to inform the governance activities that follow.

The acronym, ‘**SSAFE-D**’ stands for **Sustainability**, **Safety**, **Accountability**, **Fairness**, **Explainability**, and **Data-Stewardship**.⁴

The SSAFE-D Principles



Sustainability

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

Safety

- From a technical perspective, sustainable AI projects should be safe, secure, robust, and reliable. For example, for a system that supports forecasting a future trade surplus, safety as *reliability* may depend on the availability, relevance, and quality of data.

Accountability

- Transparency of processes and associated outcomes coupled with processes of clear communication that enable relevant stakeholders to understand how a project was conducted or why a specific decision was reached (e.g., project documentation) and,
- The establishment of clear roles and duties to ensure that the project is governed and conducted in a responsible manner. Establishing a single point of contact or ownership for a project is a means of ensuring accountability. In coding environments, formal version control practices are central to establishing accountability for aspects of a system.

Fairness

- Determining whether the design, development, and deployment of data-driven technologies is fair begins with recognising the full range of rights and interests likely to be affected by a particular system or practice.
- From a legal or technical perspective, projects outcomes should not create impermissible forms of discrimination (e.g. profiling of people based on protected characteristics, disparate treatment of members of protected groups) or give rise to other forms of adverse impact (e.g. negative effects on social equality). Statistical metrics of fairness may be relevant here.
- Second, there are implications that fall within broader conceptions of justice, such as whether the deployment of a technology (or use of data) is viewed by impacted communities as disproportionately harmful (e.g. contributing to or exacerbating harmful stereotypes)
- While statistical approaches to fairness may be useful, social awareness and stakeholder consultation are also important considerations.

Explainability

- Explainability refers to a property of a data-driven technology (e.g. AI system) to support or augment an individual's ability to explain the behaviour of the respective system. It is related to but separate from interpretability.
- For instance, whereas a ML algorithm may be more or less interpretable based on underlying aspects of its architecture (e.g. simple to understand decision trees versus a complex convolutional neural network), the ability to explain how an algorithm works depends in part on properties of the wider system in which an algorithm is deployed.
- The expertise of system producers and users is also a factor; sometimes the even the people who choose or develop a model are challenged to understand it completely. Auxiliary tools, such as dashboards or feature selection tools, may be required.
- The principle of explainability can often conflict or be in tension with other principles, such as confidentiality or safety, requiring careful balancing of interests.

Data Stewardship

- The principle of Data Stewardship is intended to focus an ethical gaze onto the data that undergirds AI projects.
- 'Data Quality' captures the static properties of data, such as whether the contents of a data set are a) relevant to and representative of the domain and use context, b) balanced and complete in terms of how well the dataset represents the underlying data generating process, and c) up-to-date and accurate as required by the project.
- 'Data Integrity' refers to more dynamic properties of data stewardship, such as how a dataset evolves over the course of a project lifecycle. In this manner, data integrity requires a) contemporaneous and attributable records from the start of a project (e.g. process logs; research statements), b) ensuring consistent and verifiable means of data analysis or processing

during development, and c) taking steps to establish findable, accessible, interoperable, and reusable records towards

i Is this a *complete* list of principles and considerations?

While our list of SSAFE-D principles is meant to cover a broad range of ethical considerations, no list of principles can account for everything. Users of this guidance are likely to encounter or have encountered other lists. A notable example is the government's [Data Ethics Framework](#) which includes three core principles that correspond to elements of the SSAFE-D principles: Transparency (which maps to our *accountability* and *explainability* principles), accountability, and fairness. We add to these *sustainability*, *safety*, and *data stewardship*.

In addition to the principles listed here, there are other considerations, such as adherence to data protection and human rights law, the Public Sector Equality Duty, and other legal obligations.

The SSAFE-D Principles should first be reviewed by the project team. [During Scoping and Anticipatory Reflection](#), each principle should be evaluated against what is known about the AI system, service, or component, and documented.

The AI Project Lifecycle

The Project Lifecycle Model is a heuristic model for structuring reflection, deliberation, and practical decision-making across all stages of an AI project's lifecycle.⁵ The project lifecycle delineates those stages where important ethical questions may be raised and decisions and actions may be required. The model is intended to be used by project teams and other decision-makers to support the adoption and implementation of safe and ethical AI and other data-driven technologies.

It is a *heuristic model* because it represents the typical stages and tasks of a project that are undertaken to design, develop, and deploy a data-driven technology. However, like all models, it is an abstraction from the actual day-to-day practices that are carried out by a team:

“All models are wrong, but some are useful.”

— George Box

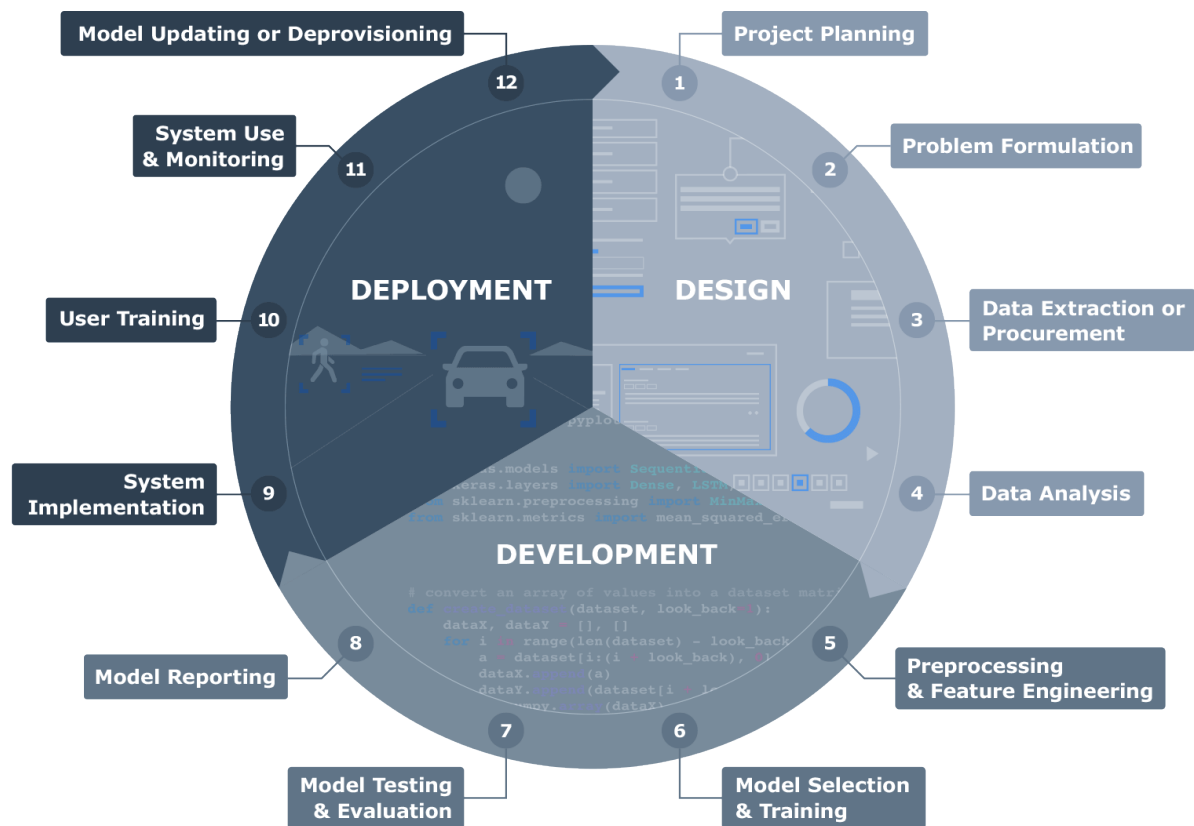
Why is the project lifecycle important?

The Project Lifecycle Model⁶ provides support by serving as a scaffold to help determine where the activities, tasks, roles, skills, and resources, and other things necessary to the project, ought to be located.

In [Section 4](#), we demonstrate how the project lifecycle model can be used to support a team's:

- Initial *reflection* about the tasks or actions that should be undertaken at the respective stages,
- *Deliberation* about how the tasks and actions may undermine or promote relevant project goals and objectives (e.g., developing a fair classifier) and,
- Ongoing *decision-making* as the project unfolds and actions are documented.

Project Lifecycle Model Overview



This model shows the typical stages of a project, which involves the design, development, and deployment of some data-driven technology, such as a ML algorithm or an AI system.



Model layers

There are two layers to the model:

- **Three overarching stages**
 - Project Design
 - Model Development
 - System Deployment
- **Twelve lower-level stages**
 - Project Planning
 - Problem Formulation
 - Data Extraction & Procurement
 - Data Analysis
 - Pre-processing & Feature Engineering
 - Model Selection & Training
 - Model Selection & Validation
 - Model Reporting
 - System Implementation
 - User Training
 - System Use & Monitoring
 - Model Updating & Deprovisioning

Let's start with the three overarching stages:

1. **Project Design**

- Preliminary tasks and activities that set the foundations for the development of the model and system (e.g., impact assessments, data extraction and analysis).


2. **Model Development**

- Technical and computational tasks associated with machine learning (e.g., training, testing, validation, and documentation), which are necessary to ensure the model is appropriate for its intended use with the target system.

3. **System Deployment**

- Tasks that ensure the safe and effective deployment and use of the system (and underlying model) within the target environment by the intended users. This stage includes ongoing monitoring, as well as tasks associated with updating or deprovisioning.

-

 Important

The Project Lifecycle Model is presented as a linear model, but in reality, it is not.

In practice, the stages of a project are often iterative, tasks within each stage are often undertaken in parallel, and actions made at a downstream stage have often been pre-determined by choices made upstream.

For instance, consider the following relationships between a series of tasks and decisions:

- When designing a project, your team may consider what sorts of data need to be collected and processed. It may not be clear at the earliest instance of the *Data Extraction & Procurement* stage what sorts or types of data may be available. Additionally, how your data should be processed may depend on the *Model Selection & Training* stage in light of your team's evolving expertise, which in turn may affect how data is cleaned and labelled at the *Pre-processing & Feature Engineering* stage.
- Looking at the overarching system deployment stage of the project lifecycle, when it comes to *System Implementation* and *User Training*, multiple iterative loops may be required between stages which include figuring out the interface design through A/B testing, beta private testing, beta public testing, and live testing depending on the nature of the project.

Procurement

The Project Lifecycle model is designed to account for different scenarios of system design, development, and deployment, including scenarios in which all or part of a system or service come from third-party providers. Where this is the case, deliberation over lifecycle stages that pertain to design and development remain important because project teams must be accountable for the systems they deploy and should have as much a view into the lifecycle stages of the system as is feasible to reputational risk and to uphold the department's public service obligations. We have prepared guidance specific to the challenges of ensuring that procured systems are sufficiently evaluated as part of an AI project. We provide a Procurement Guidance tool as [Appendix E](#).

Section 1.2 - The PBG Framework

In this Section

- ✓ Identifying Artificial Intelligence
- ✓ Process Based Governance log (PBG log)
- ✓ Project Summary Report (PS Report)
- ✓ Context Based Risk Analysis (COBRA)
- ✓ Stakeholder scoping and identification

The purpose of this guidance is to provide a framework for AI governance. The recently adopted standard, ISO 37000, defines governance as ‘the system by which the whole organisation is directed, controlled, and held accountable to achieve its core purpose in the long run’. Establishing a diligent and well-conceived governance framework that covers the entire design, development, and deployment process will provide the foundation for your teams to effectively establish needed practical actions and controls, exhaustively distribute roles and responsibilities, and operationalise answerability and auditability throughout the lifecycle of an AI project. By organising all of your governance actions into a PBG Framework, you will be better able to accomplish this task.

The purpose of the PBG Framework is to facilitate the integration of ethical norms, values, and principles, which motivate and steer responsible innovation, with the actual processes that characterise the AI adoption and deployment pipeline. A helpful framing to get started is to conceptualise the adoption and use of an AI technology as a *project* undertaken by a team within an organisation. An AI project has three major phases: design, development, and deployment. These phases can describe the adoption of a system or service whether it is purchased whole or in-part from suppliers or produced entirely within your organisation. Within every AI project is the *process* through which each phase occurs, channelling human and technical systems towards a particular goal. To maximise the ethical accountability of an AI project, this guidance employs *process-based-governance*, which is a method for putting ethical principles into practice by clearly articulating and documenting the decision-making process throughout the lifecycle of the project.



The PBG Framework should give you a landscape view of the governance actions that are organising the control structures of your project workflow. Constructing a good PBG Framework will provide you and your team with a big picture of:

- The relevant stages of the project workflow in which actions are necessary to meet governance goals.
- The relevant team members and roles involved in each governance action.
- Explicit timeframes for any necessary follow-up actions, re-assessments, and continual monitoring.
- Clear and well-defined protocols for logging activity and for instituting mechanisms to assure end-to-end auditability and appropriate documentation.

The PBG framework asks that teams not only outline the governance actions established for individual projects, but also roles involved in each action, timeframes for follow-up actions, and logging protocols.

The PBG process

Once you have fully implemented your PBG framework for an AI project, you should have the following information collected, which is tracked in a **process log**.

Is it AI?	An evaluation of the product or service under review that concludes it is AI, with justification drawn from your AI definition.
Project Summary Report	A Project Summary Report that includes preliminary information about the project, data, intended uses, preliminary risk analysis, ethical deliberation, and relevant stakeholders.
Roles and Responsibilities	A record of the team members in the AI project including each person's role in the project and their responsibilities for its ethical design, development, and deployment.
Timeframes	Explicit timeframes for actions, follow-ups, reassessments, and continual monitoring.
Data Factsheet	Documentation of the data that will be processed by the system, including what is known about training data and the data the system will act upon and produce.
Context-Based Risk Assessment	A more complete analysis of risk factors and their anticipated scale, scope, and duration.
Stakeholder Engagement Plan (SEP)	A plan for engaging with stakeholders who will design, use, or and/or are affected by the AI system.
Stakeholder Impact Assessment (SIA)	Details of the ethical and other risks and harms that emerge from engaging with stakeholders.
Readiness Self-Assessment	Responses to the Readiness Self-Assessment tool.
SSAFE-D Core Attributes Identification	An inventory of the SSAFE-D Principles broken down and operationalised as Core Attributes.
Bias Self-Assessment	Responses to the Bias Self-Assessment tool.
Data Protection	A Data Protection Impact Assessment that highlights privacy and transparency protections and obligations.
Intellectual Property Assessment	A review of copyright and/or patent issues raised by the AI project.
Monitoring and Evaluation Plan	Details of plans and schedule for monitoring the AI system in use and periodic re-evaluations of its impacts.

Scoping and Anticipatory Reflection



There are four steps in *scoping and anticipatory reflection*.

1. Identifying and describing the technology under review.
2. Launching governance documentation.
3. Conducting a Context-Based Risk Analysis.
4. Preliminary identification of stakeholders whose perspective will enrich your understanding of the risks and benefits of the technology.

Step 1: Identify & Describe

The first step to scoping and anticipatory reflection is to ensure that decision-makers governing a project have as complete a picture as possible of the technology in question. Employing a definition of AI, first we must assess if the system, service, or component is in fact AI.

The AI pioneer Marvin Minsky (1968) defined AI as follows: '*Artificial Intelligence is the science of making computers do things that require intelligence when done by humans.*' This is useful starting point, but as AI has developed, more nuanced understandings have arisen and should be considered as part of a working definition.

A glossary of AI and related terms is found in [Appendix A](#). For the purposes of this guide, we begin with three additional definitions.

💡 AI Definitions

AI systems are algorithmic models that carry out cognitive or perceptual functions in the world that were previously reserved for thinking, judging, and reasoning human beings.⁷

...a **machine-based system** that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.⁸

Generative AI, as the name suggests, generates images, music, speech, code, video or text, while it interprets and manipulates pre-existing data. Generative AI is not a new concept: machine-learning techniques behind generative AI have evolved over the past decade.⁹

Some additional considerations:

- Some, not all, AI systems can *classify* information—sorting it into categories. For example, a system that can identify tumours by analysing MRI scans.
- Some, not all, AI systems can *make predictions* about people, things, or events. For example, a system that assigns a risk category to incoming hospital patients by identifying similar patterns in data from other patients with known outcomes.
- Some, not all, AI systems (GenAI) can generate new content, including plausible text, images, audio, video, and computer code. Large language models (LLMs) are a type of GenAI but there are other types. For example, a system that can generate a video of an actual person speaking dialogue they have never spoken.
- *As a general rule*, AI systems typically process **data** though not all data-processing systems are AI. AI systems may use sensors to generate new data by sampling the physical world or act upon stored textual, numerical, and other data represented in digital form.

While these definitions and explanations are accepted and useful, it is important to recognise that clearly identifying AI is a challenging task. First, AI it is not a single

technology; it is more of a discipline or practice that aims to create a range of computer-based systems that perform complex tasks.

- **Many people or companies claim their systems are AI but those claims should not be taken at face value.**

The marketing and branding of technology as AI is often seen as a selling point. Vendor claims about their systems or software development expertise should be scrutinised closely and with a sceptical eye. The more astonishing the claim, the greater scrutiny it merits.

Another challenge for defining AI is that it is a concept in a continuous state of evolution. Many technologies we take for granted but that we would not likely call AI today are only possible because of prior AI research. The optical character recognition (OCR) technology that identifies letters and numbers in printed or image-based text and is built into document software, scanners, and copy machines is based on key computer vision techniques that were once considered foundational AI research. This is just one example of a sophisticated technology that few would describe as AI today.

The challenge of defining AI or settling on a definition that will remain sufficiently inclusive of emerging innovations means that the decision to label something as AI may come down to a judgement call. What's important is to document this decision and provide the reasoning behind it. For additional guidance, we provide a list of current AI systems in use in the public sector:

AI in the Public Sector	
Health and social care	Predicting development of pandemics and epidemics to inform preventative interventions. ¹⁰
	Categorising children as 'at-risk', to inform decisions about the safety of the home environment. ¹¹
	Predicting patients' risk in emergency rooms to triage patients or inform patient wait times. ¹²
Education	Automating assignment evaluations to save teachers' time and ensure consistency. ^{13, 14}
Local government	Predicting population trends (i.e., births) to inform development plans according to local need. ¹⁵
	Predicting individual's behaviour within services to inform interventions (i.e., encourage individuals to save, to pay council taxes, to reduce antisocial behaviour). ^{16, 17, 18}
	Identifying suitable sites for housing development. ¹⁹
Energy and utilities	Predicting households' energy usage to inform personalised tariffs.
	Predicting maintenance needs and errors within energy generation and distribution systems to inform preventative action. ²⁰
Transport	Predicting road maintenance needs to inform preventative action. ²¹
	Predicting traffic and controlling traffic signals to reduce congestion. ²²

	Predicting long-term passenger needs across modes of travel to inform transport infrastructure plans. ^{23, 24}
Environment and agriculture	Identifying sources contributing to air pollution to inform policy interventions. ²⁵
	Predicting crops at risk of disease to inform appropriate treatment. ²⁶
Defence and security	Predicting vulnerabilities within cybersecurity systems to inform preventative action. ²⁷
	Predicting battlefield conditions and optimising military effectiveness by running simulations. ²⁸
Criminal justice	Identifying individuals suitable for rehabilitation services to inform court decisions between rehabilitation and custody. ²⁹
	Predicting individuals' risk of re-offence within the criminal justice system. ³⁰
	Automating the analysis of digital evidence within court cases to streamline process. ³¹
Immigration and policing	Predicting areas likely to have heavy criminal activity to inform policy deployment. ³²
	Categorising immigration applications (i.e., visa applications, residential status applications, citizenship applications) as low, mid, or high risk of being fraudulent to inform the level of human oversight over applications to streamline processes. ^{33, 34}
Digital markets and communications	Categorising online content to remove misinformation, misleading advertisements, and scams. ^{35, 36}
Across domains	Automating service provision via conversational AI (i.e., licence approvals, customer service). ³⁷
	Categorising users of digital public services to automate personalised content delivery (i.e., recommending relevant help articles to individuals using government platforms). ³⁸
Government	Predicting optimal budget allocations and policies aimed to meet national strategy objectives. ³⁹

Step 2: Documentation

The Process Log

Documentation is the key to effective governance. Governance actions will, in practice, be tailored to the needs of specific projects as teams are to establish governance actions that are proportional to the potential risks and hazards presented by their project. In all instances, however, the effective implementation of the PBG Framework will result in the production of a **process log** (PBG log).

The PBG Log is a place for indexing the documentation of key decisions, activities, and justifications involved in the adoption and use of AI systems and services. The process

log functions as a detailed register of governance actions that stewards the end-to-end transparency and accountability of AI projects, providing a documentary touchpoint for ensuring that AI systems are produced and used ethically, safely, and responsibly. It is also a means of providing accountability for tracking and communicating about the essential decisions that were made and the evidence used to justify them. A sample PBG log is provided as [Appendix B](#).

A PBG log:

- Consolidates and articulates information about the system or service.
- Outlines governance actions across the project lifecycle.
- Identifies relevant team members and roles involved in each governance action.
- Explicit timeframes for follow-up actions, re-assessments, and continual monitoring.
- Clear and well-defined protocols for logging activity and instituting mechanisms for end-to-end audibility.
- A record of key questions and answers during ethical deliberation.

Implementing a process log requires two steps:

- Establishing appropriate governance controls and actions for a project.
- Logging project activities and ethical deliberation based on the established governance controls.

Effective implementation of the PBG Framework will result in the production of a process log. The process log functions as a detailed register of governance actions that stewards the end-to-end transparency and accountability of AI projects, providing a documentary touchpoint for ensuring that AI systems are produced and used ethically, safely, and responsibly. A template for a PBG log is attached as [Appendix B](#).

Project Summary Report (PS Report)

The first document in your PBG log is the Project Summary Report, or **PS Report**. Having determined that the project under review meets your working definition of AI, the next step in the governance of an AI system or service is to gather and document basic information about your project and to begin consideration of potential risks and ethical issues and how they may be mitigated (a report template is provided as [Appendix C](#)). To begin a PS report, members of the project team conduct desk-based research to assemble information about the technologies and data that are elemental to the project.

A PS report provides a starting point from which practitioners and decision-makers establish a proportionate approach to the remaining steps of ethical governance of AI and to determine the level of stakeholder engagement that is needed across the project lifecycle.

The work to produce the PS Report includes a set of project scoping and planning activities in which responsible parties (a) consolidate information about the project, use context, domain, and data-contexts of the prospective system, (b) identify relevant stakeholders, (c) begin to scope the ethical principles implicated, and (d) map a governance workflow.

Scoping and anticipatory reflection can be conducted by responding to a series of questions and prompts that help decision-makers pinpoint ethical principles, mitigation strategies, and an overall governance plan for the target system. During the initial project scoping activity, you should draw on organisational documents (i.e., the project business case, proof of concept, or project charter), project team collaboration, and desk research (if necessary) to complete the initial reporting. Some example questions to be addressed appear below:

Example questions for the PS Report. See the PS Report template for more.

PROJECT	USE CONTEXT	DOMAIN	DATA
Is it AI (refer to AI definitions here and in Appendix A)	What features of the system meet an accepted definition of AI or closely related technology?		
What AI system is being built (or acquired) and what type of product or service will it offer?	What is the purpose of this AI system and in which contexts will it be used? (Briefly describe a use-case that illustrates primary intended use)	In what domain will this AI system operate?	What datasets are being used to build this AI system?
Which organisation(s)—yours, other suppliers, or other providers—are responsible for building this AI system?	Is the AI system's processing output to be used in a fully automated way or will there be some degree of human control,	Which, if any, domain experts have been or will be consulted in designing and developing the AI system?	Will any data being used in the production of the AI system be acquired from a vendor or supplier? (Describe)

	oversight, or input before use? (Describe)		
Which parts or elements of the AI system, if any, will be procured from third-party vendors, suppliers, sub-contractors, or external developers? ⁴⁰	Will the AI system evolve or learn continuously in its use context, or will it be static?		Will the data being used in the production of the AI system be collected for that purpose, or will it be re-purposed from existing datasets? (Describe)
Which algorithms, techniques, and model types will be used in the AI system? (Provide links to technical papers where appropriate)	To what degree will the use of the AI system be time-critical, or will users be able to evaluate outputs comfortably over time?		What quality assurance and bias mitigation processes do you have in place for the data lifecycle—for both acquired and collected data?
In a scenario where your project optimally scales, how many people will it impact, for how long, and in what geographic range (local, national, global)? (Describe your rationale)	What sort of out-of-scope uses could users attempt to apply the AI system, and what dangers may arise from this?		

Roles and Responsibilities

As you collect information about the AI system or service, you should also be mapping out a governance workflow, to be updated throughout the project lifecycle. The governance workflow is the assigning of roles and responsibilities to individual team members and ensuring they understand those roles and responsibilities. Using the project lifecycle model as a guide, assign tasks and responsibilities to individual team members for each phase of the project and add this information to the relevant section of the PS Report.

Data Factsheets

In addition to gathering general information about the project and its technical components, the PS Report will not be complete without accounting for the data that will underpin the AI project. In the PS Report template, there are general questions about the origins, quality, and integrity of the data to be used in your project. However, a fuller accounting should be performed because of the importance and centrality of data to most AI projects. There are several widely cited approaches to creating a Data Factsheet, such as *Datasheets for Datasets*,⁴¹ which poses qualitative questions about data origins and purposes, and *The Dataset Nutrition Label*,⁴² which poses more technical questions.

We provide a Data Factsheet template as [Appendix D](#). We urge project teams to add their own questions in addition to those provided that pertain to the requirements of the project under review.

Training Data

While it is important to account for and evaluate the data the AI system or service is meant to act upon, the data used to *train* the system during system development is also important to account for to ensure the system meets ethical and legal obligations. Biases or quality issues in training data can shape the system's outputs and pose potential ethical problems.

Where the data used for training a system originates with DBT or a trusted partner, accounting for it on a factsheet should be relatively easy to straightforward. However, where AI systems or services are procured or licensed from providers, as is likely the case for Generative AI (GenAI) systems, acquiring information about training data may be more challenging. Providers may be unwilling to provide the complete details of their training data, posing a dilemma for teams attempting to govern their AI systems. Many GenAI systems are trained using proprietary methods to crawl online sources the provider considers to be “public”, but they may withhold a full accounting of those sources for business secrecy reasons. Systems may also be trained using proprietary data held or licensed by the provider, and there too, providers may not be forthcoming with the details. Nevertheless, teams should make every effort to get as much information as possible and consider this aspect as a matter of procurement ethics (see the Procurement Guidance tool in Appendix E).



Step 3: Context-Based Risk Analysis (COBRA)

Introduction

Having reviewed the SSAFE-D Principles, the Project Lifecycle, and conducted desk research, the project team should move towards accounting for the risk factors and anticipated effects of the AI system through a context-based risk analysis, or COBRA. The purpose of the COBRA is to aid project teams in identifying risk factors that may be elemental to a data-driven system or that are present in the system's

context of use. Risk analysis is a form of *anticipatory reflection* in which project teams seek to get out ahead of possible risks and make plans for mitigating them.

The COBRA is a method for assessing whether and to what extent the deployment of an AI system could pose ethical risks for the team, the organisation, or to external stakeholders. By identifying the extent of potential risks, project teams can identify a proportionate response necessary to ensure responsible system design, development, and deployment.

The COBRA is constructed as a series of questions whose responses help to identify risks potentially arising from the AI system itself, as well as the risk factors already present in the specific context (circumstantial risk factors) in which an AI system is deployed.

- A COBRA template is provided as [Appendix H](#).

What is a risk factor?

The characteristics or properties of an AI innovation context that contribute to some outcome (or outcomes) that is harmful to the well-being of individuals or groups or that negatively impacts their fundamental rights and interests. While a risk refers to the potential negative harms that an AI system could pose, a *risk factor* more generally refers to any contributor to a negative or harmful outcome.

Risk factors arising in the practical context of the AI project lifecycle

Building on a frequently cited classification scheme provided by the [Organisation for Economic Co-operation and Development \(OECD\)](#), we offer a model to describe the spectrum of risk factors that surround the practical contexts of the AI project lifecycle. These include the application context in which the system is conceived and built, the data lifecycle context, the project design context, the model development context, and the system deployment context.

Risk factors pertaining to AI application contexts

Adverse impacts can be identified when an AI system is developed or deployed without consideration of the socio-cultural and legal factors surrounding its application in a given sector. With a focus on the social and legal landscape of deployment, the harms to individuals and communities can be identified through risk factors arising from their use in the following areas:

A system deployed within high impact, safety critical, or historically highly regulated sectors or domains.

- For example, a risk assessment tool that used by social workers to identify children in need of state-assisted care operates within the social care system, where children (a vulnerable group) and their families are significantly impacted.

A system repurposed or used in prohibited ways within existing statute and regulation.

- For example, a bank using a risk-assessment tool to predict borrower default operates within the financial sector, where extensive regulations pertaining to market abuse, risk management, equity law, and competition law have historically been in place.

A system with a significant scope of deployment (including breadth and temporality) and number of stakeholders affected.

- For example, a classification system used to categorise immigration applications (visa, residential status, and citizenship applications) as low, mid, or high risk of being fraudulent may have lifetime effects on individuals' immigration status.

A system not being wholly based on existing and externally validated techniques for a similar purpose and in the same sector.

- Here, consider the example of technological immaturity that is both a circumstantial and a modifiable risk factor.

A system replacing existing flawed or harmful (human or technological) systems.

- For example, an energy supplier adopting a reinforcement learning system within a power grid to predict energy demand curves and trigger responsive energy distribution does not complete and make public an assessment outage risks.

A system built without consideration for existing legacies of bias and discrimination in the sector or domain context.

- For example, a classification system used to identify individuals eligible for cervical screenings fails to identify eligible transgender individuals. The data used to train the system did not represent this population, replicating historic inequalities within healthcare resulting in the unequal treatment of a marginalised population.
- The data use to train the system did not represent this population, replicating historic inequalities within healthcare resulting in the unequal treatment of a marginalized population. classification system used to identify individuals eligible for cervical screenings fails to identify eligible transgender individuals. The data use to train the system did not represent this population, replicating historic inequalities within healthcare resulting in the unequal treatment of a marginalized population.

A system deployed in the absence of effective and transparent compliance and reporting mechanisms for environmental protection.

- For example, an AI system used to scout and identify potential natural gas or petroleum reserves for extraction ought to be subjected to evaluation for its compliance with environmental protection regulations.

A system built without consideration for cybersecurity conditions including any opportunities for third party hacking or corruption, and/or the absence, or a lack, of system testing for vulnerabilities or other proportional cybersecurity measures.

- For example, a system used in the banking and financial sector does not routinely conduct resilience tests thereby leaving it vulnerable to data leaks or Denial-of-Service attacks.

Risks pertaining to the project design context

Within the design phase, the decision to develop an AI system can be determined without considerations of factors such as the available data, resources, existing technology or with an absence of transparent processes or stakeholder input. Consider the following examples:

- A pre-trial risk assessment tool that uses data about a person's history, education, and living conditions to predict whether they will appear for trial presents a risk of denying persons their freedom based on predicting their non-appearance.
- A screening algorithm analyses data collected during recorded job interviews and ranks candidates as "successful employees" by comparing their properties of those of existing employees identified by management as "successful". The algorithm has the potential to unfairly label candidates whose demographic or cultural identity does not closely align with the existing workforce or the assumptions of system developers.

Risks pertaining to the model development context

At the phase of development, factors include the model's characteristics, model selection, pre-processing and feature engineering, and the need for privacy-preserving methods. Here, the suitability of models and methods may be jeopardised by factors such as the inferences from the learning mechanisms, accuracy and performance metrics, transparent reporting and external reviews for verification and validation. Consider, for example:

- An algorithm used to determine if a child should go into care uses data collected about the family and held within the care system as well as medical data as inputs into a neural network which presents a risk to interpretability and explainability. If the model fails to include privacy-preserving methods, sensitive personal data of the individual may be susceptible to risks on the fundamental rights to privacy. Additionally, consideration may also be given to the methods involved in models, for instance, where data is limited, and minority ethnic groups may be grouped as "Unknown" leading to instances of bias.


Risks pertaining to the model deployment context

At the phase of model deployment, factors include potential harm to the physical, psychological, or moral integrity of implementers or adverse impacts to their dignity, autonomy, and ability to make free, independent, and well-informed judgements. At this phase, absence of measures to ensure competent involvement of human implementers or users who understand the strengths and limitations of the system and outputs, as well as conditions for exercise of human judgement based on contextual complexities, anomalies, or system failure, may be correlated with an increased chance of harm. Consider, for example:

- A social services system that determines the eligibility of claimants for benefits and services based on income and other personal data, may make determinations that conflict with those that would be made by human case workers, leaving them challenged to assert their independent judgement and discretion.
- An AI robot that interacts with children both physically and conversationally in a nursery setting for educational purposes may expose children to harm if permitted to act without monitoring.

Risk Factor Ratings

As you consider the risk factors present in the use and domain contexts of the AI system or service, you should document the degree of each risk in the PS report.

 Risk Factor Ratings	
Prohibitive risk factor	Prohibitive risk factors indicate the presence of determinants of potential harms that trigger the precautionary principle ⁴³ and precipitate pre-emptive measures to prevent adverse impacts on fundamental rights and freedoms. Pre-emptive measures are appropriate where the severity, scale, and irremediableness of the potential harm outweigh levels of risk reduction and mitigation.
Major risk factor	Major risk factors indicate the presence of determinants of potential harms that are directly or indirectly associated with significant risks of adverse impacts fundamental rights and freedoms but that provide opportunities for risk reduction and mitigation that make the risks posed tolerable.
Moderate risk factor	Moderate risk factors indicate the presence of determinants of potential harms that are directly or indirectly associated with risks of adverse impacts on fundamental rights and freedoms but that provide opportunities for risk reduction and mitigation that make the risks posed broadly acceptable.


Quantifying Risk

The goal of a COBRA is to direct project teams towards a proportionate governance response including the extent of stakeholder engagement required. Low-stakes AI

applications that are not safety-critical, do not directly impact the lives of people, and do not process potentially sensitive social and demographic data may need less proactive stakeholder engagement than high-stakes projects. You and your project team will need to carry out an initial evaluation of the scope of the possible risks that could arise from your project and of the potential hazards it poses to affected individuals and groups. You will have to apply reasonable assessments of the dangers posed to individual wellbeing and public welfare in order to formulate proportionate approaches to stakeholder involvement.

Because risk can be classified across categories, it is challenging to provide a comprehensive and conclusive quantification metric. We provide a means for categorising risk and a suggested method for evaluating each category. Regardless of the potential impacts of a project, involving affected individuals and communities in stakeholder analysis (and, later, in stakeholder impact assessment) should, in all cases, be a significant consideration. Stakeholder involvement ensures that your project will possess an appropriate degree of public accountability, transparency, legitimacy, and democratic governance, and it recognises the important role played in this by the inclusion of the voices of all affected individuals and communities in decision-making and policy articulation processes.

The COBRA assigns risk categorically in terms of *scope*, *scale*, and *likelihood*.

 Risk Categories		
Scope		
How many people will be adversely affected?	Calculated as a percentage of overall persons estimated to be affected by the AI system or service.	Teams should determine a percentage threshold to determine a proportionate approach to stakeholder engagement and subsequent action. This will be balanced with <i>scale</i> as a high degree of harm to even just a few may be an unacceptable degree of risk.
Scale		
How severe is the harm?	<p>Catastrophic Harm: Potential deprivation of the right to life; irreversible injury to physical, psychological, or moral integrity; deprivation of the welfare of entire groups or communities; catastrophic harm to democratic society, the rule of law, or to the preconditions of democratic ways of life and just legal order; deprivation of individual freedom and of the right to liberty and security; harm to the biosphere.</p> <p>Critical Harm: Significant and enduring degradation of human dignity, autonomy, physical, psychological, or</p>	Assign a score from 1 to 4 where 4 is <i>catastrophic</i> . The score indicates the degree of <i>proportionate</i> response required. Generally, any score greater than 1 should prompt in-depth stakeholder engagement and subsequent action.

	<p>moral integrity, or the integrity of communal life, democratic society, or just legal order</p> <p>Serious Harm: Degradation of human dignity, autonomy, physical, psychological, or moral integrity, or the integrity of communal life, democratic society, or just legal order or that harm to the information and communication environment.</p> <p>Moderate or Minor Harm: Does not lead to any significant, enduring, or temporary degradation of human dignity, autonomy, physical, psychological, or moral integrity, or the integrity of communal life, democratic society, or just legal order.</p>	
Likelihood		
<p>How likely is the harm to occur?</p>	<p><i>Unlikely</i></p> <ul style="list-style-type: none"> The risk of adverse impact is low, improbable, or highly improbable. <p><i>Possible</i></p> <ul style="list-style-type: none"> The risk of adverse impact is moderate; the harm is possible and may occur. <p><i>Likely</i></p> <ul style="list-style-type: none"> The risk of adverse impact is high; it is probable that the harm will occur. <p><i>Very Likely</i></p> <ul style="list-style-type: none"> The risk of adverse impact is very high; it is highly probable that the harm will occur. <p><i>Not Applicable</i></p> <ul style="list-style-type: none"> It can be claimed with certainty that the risk of adverse impact indicated in the prompt does not apply to the AI system. 	<p>Assign a score from 1 to 4 where 4 is <i>Very Likely</i>. The score indicates the degree of proportionate response required. Generally, any score greater than 1 should prompt in-depth stakeholder engagement and subsequent action.</p>

Step 4: Preliminary Stakeholder Identification

In the next section, detailed guidance for engaging with stakeholders is provided. However, here, during scoping and anticipatory reflection, a preliminary stakeholder identification process is recommended. This list of guiding questions can help to determine the appropriate stakeholders to consult:

Key Question	Sub-questions
<p>Who will use the system or service?</p>	<p>Who are the likely users within your organisation?</p> <p>Who are the likely non-public users outside of your organisation? (e.g. other government departments or similar)</p>

	Who are the likely public users outside of your organisation?
Who may be affected by the use of the system or service?	Who is likely to benefit most from the use of the system or service? Who is potentially harmed by the use of the system or service? Whose benefits or harms are uncertain?
Who is responsible for the system or service?	Who within the organisation is responsible for the system or service? Who is responsible outside the organisation? (e.g. third party provider)
Who is represented in the data used by the system?	Who can speak for those whose data was used to train the system or service? Who can speak for those whose data will be acted upon by the system or service?

Section 2 - Deliberation and Engagement

In this Section

- ✓ From ethical principles to core attributes
- ✓ Bias Self-Assessment
- ✓ Stakeholder Engagement Process (SEP)
- ✓ Stakeholder Impact Assessment (SIA)
- ✓ Readiness Self-Assessment

This stage is concerned with the deliberation about the ethical priorities for the project and for conducting engagement with relevant stakeholders.

From Principles to Core Attributes

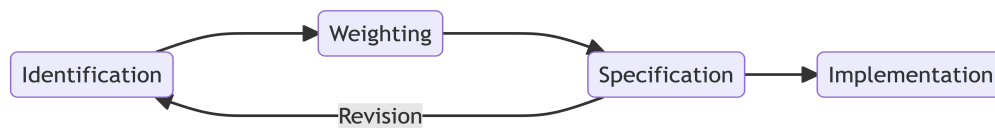
Teams and stakeholders need to have a starting point for reflection, deliberation, and engagement, and this starting point should be grounded in a shared vocabulary that recognises the unique needs and challenges of designing, developing, and deploying data-driven technologies in an ethical, trustworthy, and responsible manner.

The **SSAFE-D Principles** address this need of establishing a shared vocabulary for anticipatory reflection and deliberation, but on their own they are insufficient to guide practical decision-making and action.

To move from principles to practice is a process of specification and operationalisation, which we can summarise as follows:

- Identify the ethical principles that are relevant to the specific project.
- **Specification:** Draft a loose definition of the principle, which can be used to support preliminary forms of reflection and deliberation and also support initial stakeholder engagement and communication.
- Further specify this principle through consideration of the relevant **core attributes** and any additional attributes that help contextualise the principle.
- **Operationalise** the principle (and core attributes) by identifying practical tasks and guidelines that are embedded within the project lifecycle.
- Through additional ethical deliberation and stakeholder engagement, agree on mechanisms for monitoring and evaluation the project and

system, including practical steps for ensuring that the principles are being upheld and what to do if they are not.



Process for operationalising the SSAFE-D Principles

Before we can break down these steps further, we need to understand what we mean by several of the terms—most notably, the term **core attributes**.

! Specification

Specification is a process of clearly defining and articulating a principle or value that a project or system should uphold and embody. A high-level principle, such as ‘Fairness’ is not *specific* enough to address contextual challenges and dilemmas. It needs to be *specified* by identifying which of the core attributes are most significant. For example, in one project, ‘fairness’ may emphasise attributes such as ‘promoting equal access to services’, whereas another project may emphasise attributes such as ‘ensuring non-discrimination’. While related, these two attributes may be interpreted differently in different contexts.

! Operationalisation

Operationalisation is a process of putting a principle into practice by making it *operational* or *actionable*. This can be achieved by developing specific guidelines, policies, procedures, or mechanisms that both embody the principle in question and help implement the principle into the design, development, and deployment of a technology or system. Operationalisation depends on initial specification, but it also involves additional and iterative reflection and deliberation to ensure that the operationalisation is appropriate for the specific context (e.g. the proposed guidelines are appropriate for the specific project).

! Core Attributes

The core attributes are the set of attributes that are deemed most relevant to the corresponding principle and help guide the process of specification and operationalisation. The core attributes are designed to a) reflect the challenges of putting ethical principles into practice across the project lifecycle, and b) respond to the unique risks and opportunities of designing, developing, and deploying data-driven technologies within your department. The core attributes this user guide describes are not necessarily the only attributes that are relevant. Additional context-specific attributes may be established through anticipatory reflection and deliberation and diverse forms of stakeholder engagement.

With these definitions in mind, let’s look at how the five steps can be carried out.

Step 1: Identifying Relevant Principles

Of the five steps, this one is perhaps the easiest, as SSAFE-D Principles are a clear set of principles that are relevant to the entire project lifecycle. However, it is important to note that they will not all be relevant to every project. Therefore, this first step is simply a process of identifying which principles are relevant to the project at hand. This decision may need to be revisited as the project progresses and the project's scope or context change.

As always, meaningful forms of stakeholder engagement are critical to this process.

Important

In addition to different SSAFE-D Principles being relevant to different projects, the application of some principles may be in tension with others. For example, there may be challenges balancing the need for openness and transparency within the SSAFE-D Principles and protecting sensitive information.

Step 2: Draft a Loose Definition of the Relevant Principles

The next step is to draft a loose definition of the relevant principles. The definitions provided above can be used as a starting point but should be adapted to reflect the specific context of the project.

Initial communication and engagement with stakeholders can help you determine a) if the principle's you have selected are the most relevant ones, and b) if the definitions you have drafted are clear and accessible.

Step 3: Specify the Principle by Considering Relevant Attributes

This is the first of the steps that requires a more systematic approach.

Before we look at a prescriptive set of attributes for each principle, let's look at a simple activity that could be done to arrive at a set of relevant attributes assuming that no prior set exists.

Activity: Identifying Core Attributes

In a group, ask the following question, replacing the {principle} variable with one of the SSAFE-D Principles and then repeating the process for each one:

What does {principle} mean to you in the context of designing, developing, and deploying data-driven technologies within your organisation?

The answers to these questions represent a rough set of attributes, although they will likely require some form of synthesis and refactoring (e.g., reducing the number of attributes, or grouping attributes into broader categories). One way to capture these responses is to use a word cloud generator, to display the answers in a visual format that shows the frequency of each word (assuming that the frequency of answers captures something about the perceived importance of the word for the group).

Step 4: Operationalise the Principle(s) and Core Attributes

Once the relevant attributes have been identified (i.e. set of core attributes and any additions or revisions), the next step is to operationalise them.

To assist this process, the [AI Project Lifecycle](#) can be used to help scaffold a structured assessment of the choices or tasks that might occur (or should occur) at each stage, and whether these choices or tasks are aligned (or, conversely, undermining) the principle(s) and core attributes.

For example, for the principle of *fairness*, you might have identified that ‘Ensuring diverse and meaningful opportunities for stakeholder involvement’ is a relevant attribute. As such, you might set aside time to ensure intended users of an algorithmic system have a chance to work with members of the project team to ensure that their concerns are taken into account (e.g. reduced autonomy in professional decision-making, or reduced privacy). Alternatively, for the principle of *explainability*, you may have identified ‘Clear and accessible documentation’ as a relevant attribute. For this attribute, you might consider the need to ensure that your documentation is free of technical jargon, properly referenced, and accessible to the relevant parties through a centralised knowledge repository. Alternatively, you may choose to use a template for reporting on your data processing and analysis activities or the evaluation and testing of your model, in an attempt to ensure consistency with shared standards.

Tailored guidance

The choices or tasks that will be relevant to operationalising the principle is highly dependent on the context of the project and the nature of the principle’s specification, however, to support your individual and context-specific efforts, we provide a set of Core Attributes for the SSAFE-D Principles as [Appendix I](#).

The format of the core attributes illustrated in the Appendix serves as a template for developing a list of core attributes for your AI project.

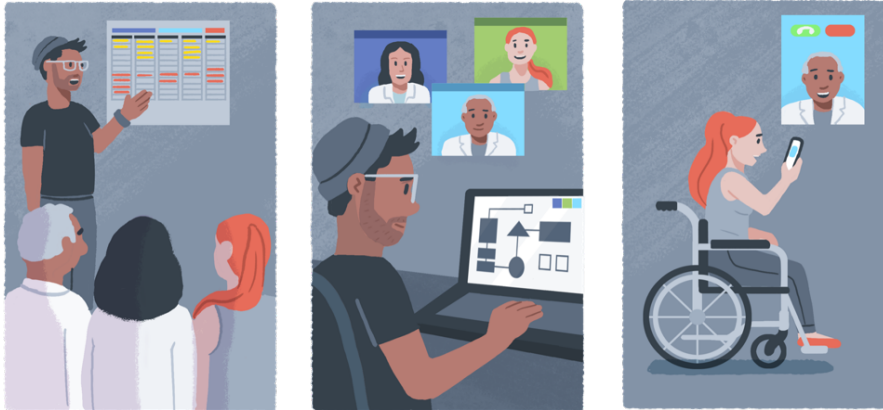
Step 5: Bias Self-Assessment

Armed with information about your AI project, an inventory of risk factors, the insights gleaned from stakeholder engagement, and an operationalised set of ethical principles in the form of core attributes, a next step is to conduct a Bias Self-Assessment that can shed light on the sources of bias and misalignment between your ethical goals and the details of the project. The assessment includes mitigation strategies for addressing biases discovered. The Bias Self-Assessment is optional but recommended. If conducted, the Bias Self-Assessment should be noted on the **PBG log**.

The Bias Self-Assessment is a set of deliberative prompts that can help guide a project team through a series of questions to help identify and mitigate biases in the project. The bias self-assessment is particularly valuable for investigating the principle of fairness and its core attributes, but it is also useful for investigating other ethical principles and attributes. The AI Project Lifecycle model is a useful navigation tool: project teams can iterate through the project lifecycle to identify where various types of bias are likely to occur.

The Bias Self-Assessment is provided as [Appendix J](#).

Stakeholder Engagement Process (SEP)



The purpose of this stage is to enhance the understanding of contextual and ethical issues surfaced during the previous stage.

A key goal is to obtain trust for the project and to ensure that the views, experiences, and perspectives of relevant stakeholders are meaningfully reflected upon throughout the project lifecycle. Importantly, the SEP is an iterative and ongoing activity that may have varying levels of stakeholder engagement depending on the specific context of the AI/ML project. **The SEP should be documented as part of Process-Based Governance.**

A *Stakeholder Impact Assessment (SIA)* is the documentation product of stakeholder engagement. The form and content of the SIA is described below. First, we describe the process of identifying and consulting with relevant stakeholders.

Stakeholder Engagement

Stakeholder engagement is a core component of any governance process. It is not a one-off activity, but an ongoing set of activities that occur throughout the project lifecycle.

Stakeholder engagement is a valuable mechanism to ensure that stakeholders' views, experiences, and perspectives are reflected in the design, development, and deployment stages. This may lead to greater support for the project and can also

KEY CONCEPT

Stakeholder

Scholars and practitioners from areas as diverse as public policy, land use, environmental and natural resource management, international development, and public health have offered many different definitions of "stakeholders" over the past several decades.¹⁹ Even so, these definitions have converged around a few common characteristics. Stakeholders are individuals or groups that:

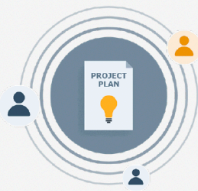
- 1 have interests or rights that may be affected by the past, present, and future decisions and activities of an organisation;
- 2 may have the power or authority to influence the outcome of such decisions and activities; and/or
- 3 have relevant characteristics that put them in positions of advantage or vulnerability with regard to those decisions and activities.

The Stakeholder Engagement Process focuses on impacted communities and groups.

lead to more beneficial, socially acceptable, and sustainable products or services. Stakeholder engagement is also vital to identify and anticipate potential impacts that might be experienced differently by different stakeholder groups, as well as developing appropriate approaches to mitigate negative impacts.

It is essential to ensure that stakeholders' views are incorporated at all stages and that any potential risks or adverse impacts are identified and mitigated. An iterative approach to stakeholder engagement is essential. It is important to revisit and revise the stakeholder analysis to ensure that approaches taken continue to reflect the perspectives and interests of salient stakeholders.

Three steps of stakeholder engagement process



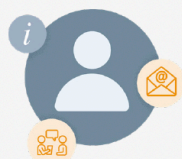
1 Preliminary Project Scoping and Stakeholder Analysis

Outline key project components, identify individuals or groups who may be affected by, or may affect, your innovation project, scope potential stakeholder impacts, and evaluate the salience and contextual characteristics of identified stakeholders.



2 Positionality Reflection

Evaluate team positionality as related to that of stakeholders. Consider strengths and limitations presented by team positionality.



3 Stakeholder Engagement Objectives and Methods

Establish engagement objectives that enable the appropriate degree of stakeholder engagement and co-production in project evaluation, and methods that support the achievement of defined objectives.

Preliminary Project Scoping and Stakeholder Analysis

This first activity of the SEP process involves identifying and evaluating the salience and contextual characteristics of individuals or groups who may be affected by, or may affect, the AI/ML project. It aims to help project teams understand the relevance of each identified stakeholder to the project and to its use contexts.

Project scoping and stakeholder analysis is comprised of four sub-steps:

1. Outlining project, use context, domain, and data: Drawing from the PS Report, consider the domain in which it will operate, the contexts on which it will be used, and the data on which it will be trained.
2. Identifying stakeholders: Building on the contextual understanding developed during project scoping, identify who may be affected by, or may affect, your project. This includes organisational stakeholders, consumers of the potential service, or large business partners, whose input can likewise strengthen the openness, inclusivity, and diversity of the project.
3. Scoping potential stakeholder impacts: Carry out a preliminary evaluation of the potential impacts of the prospective AI/ML project on relevant stakeholders. At this initial stage of reflection, members of your project team should review the ethical concerns related to the project, as well as the Departmental mission and objectives, and then consider which of these could be impacted by the design, development, and deployment of the prospective AI system. Also consider the social environment and human factors that may be affected by, or may affect, the AI model or tool.
4. Analysing stakeholder salience: Assess the relevance of each identified stakeholder group to the project and to its use contexts. Assess the relative interests, rights, vulnerabilities, and advantages of identified stakeholders as these may be impacted by, or may impact, the AI/ML project.

Stakeholder salience questions

- What stakeholder groups are most likely to be impacted by the system or tool?
- What stakeholder groups have the greatest needs in relation to potential benefits/applications of the system or the domain in which it will be deployed?
- What stakeholder groups are most and least powerful? What stakeholder groups' influence is limited?

Determining a Proportionate Approach to Stakeholder Involvement

Stakeholder analyses may be carried out in a variety of ways that involve more or less stakeholder involvement. This spectrum of options ranges from analyses carried out exclusively by a project team without active community engagement to analyses built around the inclusion of community-led participation and co-design from the earliest stages of stakeholder identification. The degree of stakeholder involvement will vary from project to project based upon a preliminary assessment of the potential risks and hazards of the model or tool under consideration.

Low-stakes AI applications that are not safety-critical, do not directly impact the lives of people, and do not process potentially sensitive social and demographic data may need less proactive stakeholder engagement than high-stakes projects. You and your project team will need to carry out an initial evaluation of the scope of the possible risks that could arise from your project and of the potential hazards it poses to affected individuals and groups. You will have to apply reasonable assessments of the dangers posed to individual wellbeing and public welfare in order to formulate proportionate approaches to stakeholder involvement.

Regardless of the potential impacts of a project, involving affected individuals and communities in stakeholder analysis (and, later, in stakeholder impact assessment)

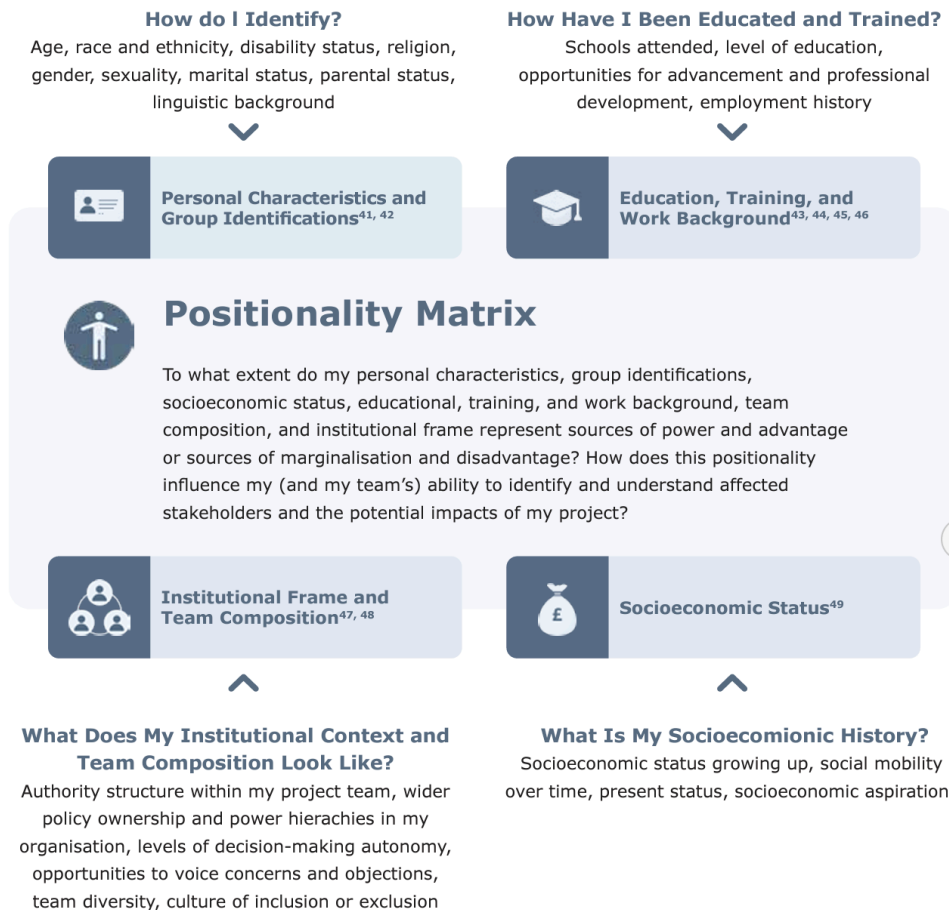
should, in all cases, be a significant consideration. Stakeholder involvement ensures that your project will possess an appropriate degree of public accountability, transparency, legitimacy, and democratic governance, and it recognises the important role played in this by the inclusion of the voices of all affected individuals and communities in decision-making and policy articulation processes.

In addition to providing these important supports for building public trust, stakeholder involvement can help to strengthen the objectivity, reflexivity, reasonableness, and robustness of the choices your project team makes across the project lifecycle. This is because the inclusion of a wider range of perspectives (especially of those who are most marginalised) can enlarge a project team's purview and expand its domain knowledge as well as its understanding of citizens' needs. It can likewise unearth potential biases that may arise from limiting the standpoints that inform decision-making to those of team members. Public engagement and community involvement, however, are only one part of the measures your team needs to take to ensure the objectivity, reflexivity, reasonableness, and robustness of its stakeholder analysis, impact assessment, and decision-making more generally. Apart from outward-facing community participation, processes of inward-facing reflection should also inform the way your team approaches to these challenges.

Positionality Reflection

All individual human beings come from unique places, experiences, and life contexts that have shaped their thinking and perspectives. Reflecting on this variation can help us understand how our viewpoints might differ from those around us, and from those who have diverging cultural and socioeconomic backgrounds and life experiences.

Social scientists have long referred to this kind of self-locating reflection as 'positionality'. When project team members consider their own positionalities, and make them explicit, they can better grasp how the influence of their respective social and cultural positions may affect how they engage with others.



At the centre of a positionality reflection is the question “How does your positionality influence my (and my team’s) ability to identify and understand affected stakeholders and the potential impacts of your project?”. These may factors include:

- personal characteristics
- cultural context
- group identifications
- socio-economic status
- education, training, and work background,
- team composition
- institutional frame

After reflecting individually on positionality, the project team should collaboratively answer the following questions:

- How does the positionality of the team members relate to that of affected stakeholders?
- Are there ways that your position as a team could lead you to choose one option over another when assessing the risks posed by the prospective AI system?
- Are there missing stakeholder viewpoints that would strengthen your team’s assessment of this system’s potential impact?

Stakeholder Engagement Objectives and Methods

Determining your objectives

The final step is to establish engagement objectives that enable the appropriate degree of stakeholder engagement in project evaluation, and methods to support the achievement of defined objectives.

The use of these goals to support the identification of engagement objectives should also be informed by a) the following variations on participation, and b) the methods of participation available to you and those who you are engaging.



Degrees of participation during stakeholder engagement

Stakeholder Engagement Objective	Level of Agency
<p>Inform</p> <p>Stakeholders are made aware of decisions and developments.</p>	<p>LOW</p> <p>Stakeholders are considered information subjects rather than active agents.</p>
<p>Consult</p> <p>Stakeholders can voice their views on pre-determined areas of focus, which are considered in decision-making.</p>	<p>LOW</p> <p>Stakeholders are included as sources of information input under narrow, highly controlled conditions of participation.</p>
<p>Partner</p> <p>Stakeholders and teams share agency over the determination of areas of focus and decision-making.</p>	<p>MODERATE</p> <p>Stakeholders exercise a moderate level of agency in helping to set agendas through collaborative decision-making.</p>
<p>Empower</p> <p>Stakeholders are engaged with as decision-makers and are expected to gather pertinent information and be proactive in co-operation.</p>	<p>HIGH</p> <p>Stakeholders exercise a high level of agency and control over agenda-setting and decision-making.</p>

The table above indicates the spectrum of stakeholder engagement that may be carried out. This ranges from analyses conducted exclusively by a project team without active internal community engagement to analyses built around the inclusion of community-led participation and co-design from the earliest stages of stakeholder identification. Low-stakes AI applications that are not safety-critical, do not directly impact the lives of people, and do not process potentially sensitive social and demographic data may need less proactive stakeholder engagement than high-stakes projects.

Stakeholder engagement objectives questions

To help project teams determine the degree of participation that is most relevant to a stakeholder engagement objective, the following questions should be answered:

- Why are you engaging stakeholders?
- What are the expected outcomes of engagement activities?
- How will stakeholders be able to influence the engagement process and the outcomes?
- What participation goal do you believe would be appropriate for this project considering challenges or limitations to assessments related to positionality, and proportionality to the project’s potential degree of impact?
- Will the stakeholders feel valued and heard through your SEP?

Once you have established your engagement objective, you are in a better position to assess which method or methods of engaging stakeholders are most appropriate in conducting your Stakeholder Impact Assessments. Determining appropriate engagement methods for conducting this process necessitates that you: 1. evaluate and accommodate stakeholder needs; and 2. pay attention to practical considerations of resources, capacities, timeframes, and logistics that could enable or constrain the realisation of your objective:

Factors Determining the Objectives of Engagement

Evaluation and Accommodation of Stakeholder Needs

- Identification of potential barriers to engagement. For instance, constraints on the capacity of vulnerable stakeholder groups to participate, difficulties in reaching marginalised, isolated, or socially excluded groups, and challenges to participation that are presented by digital divides or information and communication gaps between public sector organisations and impacted communities.^{53, 54, 55}
- Identification of strategies to accommodate stakeholder needs, such as catering the location or media of engagement to difficult to reach groups. Provision of childcare, compensation, or transport to secure equitable participation.⁵⁶
- Tailoring the information and educational materials to the needs of participants.⁵⁷
- Consideration of engagement objectives.



Practical Considerations of Resources, Capacities, Timeframes, and Logistics




- The resources available for facilitating engagement activities.
- The timeframes set for project completion.
- The capacities of your organisation and team to properly facilitate public engagement.
- The stages of project design, development, and implementation at which stakeholders will be engaged.





When weighing these factors, project teams should prioritise the establishment of a clear and explicit stakeholder engagement objective and document this.




Stakeholder Engagement Methods Summary

Deciding on the best method requires awareness of your audience and the resources that are available to your team. The following table summarises a selection of salient methods:

Engagement Method	Practical Strengths	Practical Weaknesses
 <p>Newsletters (email)</p> <p>Regular emails (e.g., fortnightly or monthly) that contain updates, relevant news, and calls to action in an inviting format.</p> <p>Engagement method:</p> <p>Inform</p>	<p>Can reach many people; can contain a large amount of relevant information; can be made accessible and visually engaging.</p>	<p>Might not reach certain portions of the population; can be demanding to design and produce with some periodicity; easily forwarded to spam/junk folders without project team knowing (leading to overinflated readership statistics).</p>
 <p>Letters (post)</p> <p>Regular letters (e.g., monthly) that contain the latest updates, relevant news and calls to action.</p> <p>Engagement Method:</p> <p>Inform</p>	<p>Can reach parts of the population with no internet or digital access; can contain large amount of relevant information; can be made accessible and visually engaging.</p>	<p>Might not engage certain portions of the population; slow delivery and interaction times hampers the effective flow of information and the organisation of further engagement.</p>

Engagement Method	Practical Strengths	Practical Weaknesses
<p> App Notifications</p> <p>Projects can rely on the design of apps that are pitched to stakeholders who are notified on their phone with relevant updates.</p> <p>Engagement Method:</p> <p>Inform</p>	<p>Easy and cost-effective to distribute information to large numbers of people; rapid information flows bolster the provision of relevant and timely news and updates.</p>	<p>More significant initial investment in developing an app; will not be available to people without smartphones.</p>
<p> Community Fora</p> <p>Events in which panels of experts share their knowledge on issues and then stakeholders can ask questions.</p> <p>Engagement Method:</p> <p>Inform</p>	<p>Can inform people with more relevant information by providing them with the opportunity to ask questions; brings community together in a shared space of public communication.</p>	<p>More time-consuming and resource intensive to organise; might attract smaller numbers of people and self-selecting groups rather than representative subsets of the population; effectiveness is constrained by forum capacity.</p>
<p> Online Surveys</p> <p>Survey sent via email, embedded in a website, shared via social media.</p> <p>Engagement Method:</p> <p>Consult</p>	<p>Cost-effective; simple mass-distribution.</p>	<p>Risk of pre-emptive evaluative framework when designing questions; does not reach those without internet connection or computer/smartphone access.</p>

Engagement Method	Practical Strengths	Practical Weaknesses
 <p>Phone interviews</p> <p>Structured or semi-structured interviews held over the phone.</p> <p>Engagement Method:</p> <p>Consult Partner</p>	<p>Opportunity for stakeholders to voice concerns more openly.</p>	<p>Risk of pre-emptive evaluative framework when designing questions; might exclude portions of the populations without phone access or with habits of infrequent phone use.</p>
 <p>Door-to-door interviews</p> <p>Structured or semi-structured interviews held in-person at people's houses.</p> <p>Engagement Method:</p> <p>Consult Partner</p>	<p>Opportunity for stakeholders to voice concerns more openly; can allow participants the opportunity to form connections through empathy and face-to-face communication.</p>	<p>Potential for limited interest to engage with interviewers; time-consuming; can be seen by interviewees as intrusive or burdensome.</p>
 <p>In-person interviews</p> <p>Short interviews conducted in-person in public spaces.</p> <p>Engagement Method:</p> <p>Consult Partner</p>	<p>Can reach many people and a representative subset of the population if stakeholders are appropriately defined and sortition is used.</p>	<p>Less targeted; pertinent stakeholders must be identified by area; little time/interest to engage with interviewer; can be viewed by interviewees as time-consuming and burdensome.</p>
 <p>Focus groups</p> <p>A group of stakeholders brought together and asked their opinions on a particular issue. Can be more or less formally structured.</p> <p>Engagement Method:</p> <p>Consult Partner</p>	<p>Can gather in-depth information; can lead to new insights and directions that were not anticipated by the project team.</p>	<p>Subject to hazards of group think or peer pressure; complex to facilitate; can be steered by dynamics of differential power among participants.</p>

Engagement Method	Practical Strengths	Practical Weaknesses
<p> Online Workshops</p> <p>Workshops using digital tools such as collaborative platforms.</p> <p>Engagement Method:</p> <p>Consult</p>	<p>Opportunity to reach stakeholders across regions, increased accessibility depending on digital access.</p>	<p>Potential barriers to accessing tools required for participation, potential for disengagement.</p>
<p> Citizen Panel or Assembly</p> <p>Large groups of people (dozens or even thousands) who are representative of a town/region.</p> <p>Engagement Method:</p> <p>Inform Partner</p> <p>Empower</p>	<p>Provides an opportunity for co-production of outputs; can produce insights and directions that were not anticipated by the project team; can provide an information base for conducting further outreach (surveys, interviews, focus groups, etc.); can be broadly representative; can bolster a community's sense of democratic agency and solidarity.</p>	<p>Participant roles must be continuously updated to ensure panels or assemblies remain representative of the population throughout their lifespan; resource-intensive for establishment and maintenance; subject to hazards of group think or peer pressure; complex to facilitate; can be steered by dynamics of differential power among participants.</p>
<p> Citizen Jury</p> <p>A small group of people (between 12 and 24), representative of the demographics of a given area, come together to deliberate on an issue (generally one clearly framed set of questions), over the period of 2 to 7 days.</p> <p>Engagement Method:</p> <p>Inform Partner</p> <p>Empower</p>	<p>Can gather in-depth information; can produce insights and directions that were not anticipated by the project team; can bolster participants' sense of democratic agency and solidarity.</p>	<p>Subject to hazards of group think; complex to facilitate; risk of pre-emptive evaluative framework; small sample of citizens involved risks low representativeness of wider range of public opinions and beliefs.</p>

Stakeholder Impact Assessments

Core to the work of stakeholder engagement is the conduct of regular and rigorous Stakeholder Impact Assessments (SIAs).

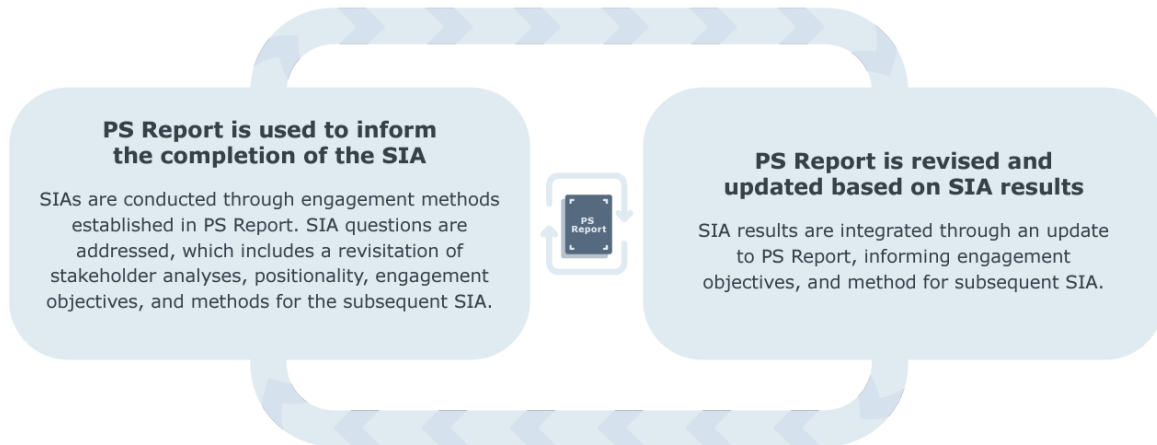
In recent years, several different types of “impact assessment” have become relevant for public sector AI innovation projects. Data protection law requires data protection impact assessments (DPIAs) to be carried out in cases where the processing of personal data is likely to result in a high risk to individuals. DPIAs assess the necessity and proportionality of the processing of personal data, identify risks that may emerge in that processing, and present measures taken to mitigate those risks. Equality impact assessments (EIAs) assist public authorities to fulfil the requirements of the equality duties, specifically regarding race, gender, and disability equality. They identify the ways government can proactively promote equality.

DPIAs and EIAs provide relevant insights about the ethical stakes of AI innovation projects. However, they go only part of the way in identifying and assessing the full range of potential individual and societal impacts of the design, development, and deployment of AI and data-intensive technologies. Reaching a comprehensive assessment of these impacts is the purpose of Stakeholder Impact Assessments (SIAs). SIAs are tools that create a procedure for, and a means of documenting, the collaborative evaluation and reflective anticipation of the possible harms and benefits of AI innovation projects. SIAs are not intended to replace DPIAs or EIAs, which are obligatory. Rather, SIAs are meant to be integrated into the wider impact assessment regime. This demonstrates that sufficient attention has been paid to the ethical permissibility, transparency, accountability, and equity of AI innovation projects.

The goals of the SIA include:

- Help to build public confidence that the design and deployment of the AI system by the public sector agency has been done responsibly.
- Facilitate and strengthen your accountability framework.
- Bring to light unseen risks that threaten to affect individuals and the public good.
- Underwrite well-informed decision-making and transparent innovation practices.
- Demonstrate forethought and due diligence not only within your organisation but also to the wider public.

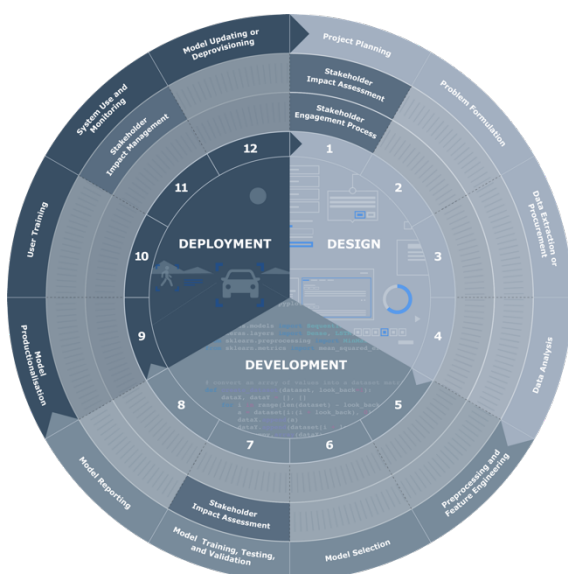
Stakeholder Impact Assessments (SIAs) provide you with the opportunity to draw on the learning and insights you have gained in your Stakeholder Engagement Processes (SEPs), and on the lived experience of engaged persons, to delve more deeply into the potential impacts of your project.



Your SIAs should enable you:

- To re-examine and re-evaluate the potential impacts you have already identified in your PS Report. •
- To contextualise and corroborate these potential impacts in dialogue with stakeholders, when appropriate. •
- To identify and analyse further potential impacts by engaging in extended reflection and by giving stakeholders the chance (when appropriate) to uncover new harms that have not yet been explored and to pinpoint gaps in the completeness and comprehensiveness of the previously enumerated harms.

Stakeholder engagement should be conducted during each major phase of the [AI Project Lifecycle](#), from Design through Development, and Deployment. To illustrate how to implement an SIA, we have provided a 3-part template with each phase appropriate to the stage of project development, from Design through to Deployment. This is provided as [Appendix G](#).



Section 1 guides the **Design Phase**, addressing project planning, problem formulation, as well as revisit of stakeholder analysis, positionally, and engagement objectives and methods.

Section 2 provides a touchpoint for evaluation and reflexivity during **Development Phase** of models and outputs, and facilitates ongoing model reporting.

Section 3 supports ongoing ethical deliberation and reflection during the **Deployment Phase** of resultant project outputs, recording relevant changes from earlier iterations of the SIA.

Readiness Self-Assessment



For new technologies to be successfully adopted and to serve the organisation’s mission while also respecting fundamental rights, decision makers at every level must ensure that new systems are feasible, useful, properly understood, and appropriately valued.

Achieving this may include ensuring that there is sufficient understanding about new technologies but also that organisations and teams involved in design, development, and deployment activities create the necessary pre-conditions of trust.

Where new AI/ML systems are particularly impactful on familiar or established working methods and performance standards, on the practice or rule of law, or upon people’s fundamental rights, ensuring readiness for a project is an essential component of ethical project design, development, and deployment. This tool supports project teams and other decision-makers in capturing the upskilling, preparation, and communication needs required to ensure a project’s success.

The [Readiness Self-Assessment tool](#) complements the activities of other tools, including the Stakeholder Engagement Process.

The Readiness Self-Assessment is provided as [Appendix H](#).

Section 3: Data Protection and Intellectual Property Considerations

In this Section

- ✓ Data protection considerations
- ✓ Intellectual property considerations

This section provides basic guidance for meeting data protection obligations and ensuring compliance with intellectual privacy standards in AI projects. We recommend adding your Data Protection Impact Assessment and an Intellectual Property Impact Assessment to your PBG log.

Data protection considerations

Your organisation should already be familiar with the key rights and obligations enshrined in UK data protection regulations. These include:

- Consent
- Data security
- Data minimisation
- Transparency
- Purpose limitation
- Accountability
- Lawfulness, fairness, and transparency
- Respect for the rights of data subjects

Fairness and transparency are foundational AI ethics principles and also to the use of data under the UK GDPR. Understanding that you are responsible for ensuring fairness and transparency will help you to ensure your general compliance with data protection law, which aims to protect people's rights and freedoms in relation to the processing of their personal data. The framework includes the UK General Data Protection Regulation (UK GDPR) and the Data Protection Act 2018 (DPA 2018). If you use personal data during AI development or deployment you need to comply with data protection. The data protection legislation is overseen by the Information Commissioner's Office (ICO), which is the UK's independent data protection authority. It has produced a suite of products a suite of products on AI on AI to assist developers and users of AI systems.

A core requirement and best practice for data protection compliance is to complete a Data Protection Impact Assessment (DPIA) for every AI project. The ICO provides an AI [a data protection toolkit](#) to assist with DPIAs and other recommended safeguarding and documentation.

As part of complying with data protection you will need to comply with its **fairness principle**. In simple terms, fairness in data protection means that organisations should only process personal data in ways that people would reasonably expect and not use it in any way that could have unjustified adverse effects on them. Organisations should not process personal data in ways that are unduly detrimental, unexpected, or misleading to the individuals concerned.

GenAI and Data Protection

GenAI poses particular challenges to data protection principles. The training generative models, including LLMs and image generators, may involve accessing data originally produced by natural persons with data protection rights. GenAI model training may be in conflict with the data protection right of consent. Additionally, it is the standard practice of GenAI companies to reveal very little about the data they use for training. This means that data subjects are unlikely to be aware when their data is being processed. This is a potential violation of notification obligations and it may limit the ability of rights-holders to conduct [subject-access requests](#) or otherwise to assert their rights under data protection law.

Furthermore, many consumer-facing GenAI systems operate by responding to the input of users. The inputted information may be subject to data protection rights that would be difficult to assert. ChatGPT, for example, generates text based on natural language inputs from users, such as “write a cover letter for the position of sales manager based on this information from my CV”. At this time, it is not known how or if a data subject can make a subject-access request regarding the information she has sent to such a system.

Intellectual property considerations

Intellectual property (IP) generally refers to a set of legal and moral rights and obligations related to any and all creative works and inventions. In the United Kingdom, the official government body responsible for IP is the Intellectual Property Office (IPO).

AI has posed new questions about two IP types: patents and copyright. Patents are granted by the IPO to the creators of useful inventions for a limited period of time. A patent reserves the use of the invention to the inventor during the patent length. Copyright applies to creative works that are not inventions, such as literature and music. A copyright applies automatically to all creative works produced by natural persons. Copyright generally includes a creator’s right to copy, distribute, rent/lend, perform/show, adapt, or publish/post their work, whether or not done for profit. The IPO provides general information on its [website](#).

Legal experts and moral rights scholars have historically debated many questions about intellectual property, including its status as “intangible” property and the moral questions of raised by creative rights. Public sector organisations are generally advised to abide by the legal definitions and obligations of intellectual property law.

GenAI and intellectual property

Generative AI poses new challenges to our understanding of IP rights and obligations. There have been developments in the use of GenAI to create patentable inventions. For the moment, regulators maintain that inventions produced with AI are only patentable by natural persons, but this could change. DBT will likely need to keep abreast of court decisions and regulatory action in this area, as it could have direct implications for UK businesses and international trade involving countries with diverging IP frameworks regarding GenAI work products.

Copyright applied to “text data mining”

GenAI systems must be trained on very large datasets. Typically, these include copyrighted material. There is a limited exemption to copyright law in the UK for ‘text data mining’ (TDM) performed to for non-commercial AI research, but the exemption is quite limited. The IPO has indicated that it may broaden the exemption to support commercial AI development in the UK, but there is no immediate plan to do so and rights-holders are not in favour of such an expansion.

Copyright protection for computer-generated works:

GenAI systems are already being used to support the creation of new artistic works and inventions. Relevant authorities have, so far, have concluded that software systems do not themselves have intellectual property rights they can assert, but some questions remain about who, if anyone, can assert such rights over generated content. At present, all IP rights remain with natural persons.

Related Considerations

The Central Digital & Data Office (CDDO) has provided [guidance](#) that cautions about the risks of submitting information to GenAI systems.⁴⁴ In brief, the CDDO suggests at all time to be mindful of *the three Hows*:

- **How your question will be used by the system**
 - Cautions about sending confidential or sensitive information to the system.
- **How answers from GenAI can mislead**
 - Cautions about the problem of GenAI “hallucinations” and factually wrong or dangerous outputs
- **How GenAI operates**
 - Cautions about the problem of GenAI failing to understand bias or context)

The National Cyber Security Centre has also provided guidance regarding GenAI that includes some data protection and related considerations. In addition to similar guidance as above, they recommend:

- **Visibility:** queries will be visible to the organisation providing the system or service. Those queries are stored and will almost certainly be used for developing the LLM service or model at some point

The NCSC also provides additional non-data protection recommended considerations:

- Consideration of the **compute resources** required to train GenAI systems.
 - AI project teams may wish to consider the carbon footprint of these systems as part of their overall ethical obligations.
- **Toxic content:** GenAI systems can be coaxed into creating toxic content and are prone to ‘injection attacks’

To date, current guidance by relevant authorities does not prohibit the use of GenAI systems by government personnel but does provide many cautions. Legal counsel should be consulted regarding your organisation’s data protection and intellectual property obligations in regards to AI. Legal opinion and official guidance about the permitted use of GenAI systems in addition to ethical considerations should guide its use by your organisation.

Data protection and intellectual property governance actions

Each of the below actions should be documented as part of the PBG log:

- **A Data Protection Impact Assessment (DPIA)** should be conducted for every AI project. For procured systems and datasets, providers should share their DPIAs. A DPIA template should be provided by your organisation’s data protection team.
- **An Intellectual Property Impact Assessment** may also be recommended. Consult with legal counsel with specific expertise in IP law to conduct such an assessment.
- **User training:** Before implementing an AI system, users should be made well-aware of their data protection, intellectual property, and general confidentiality obligations.

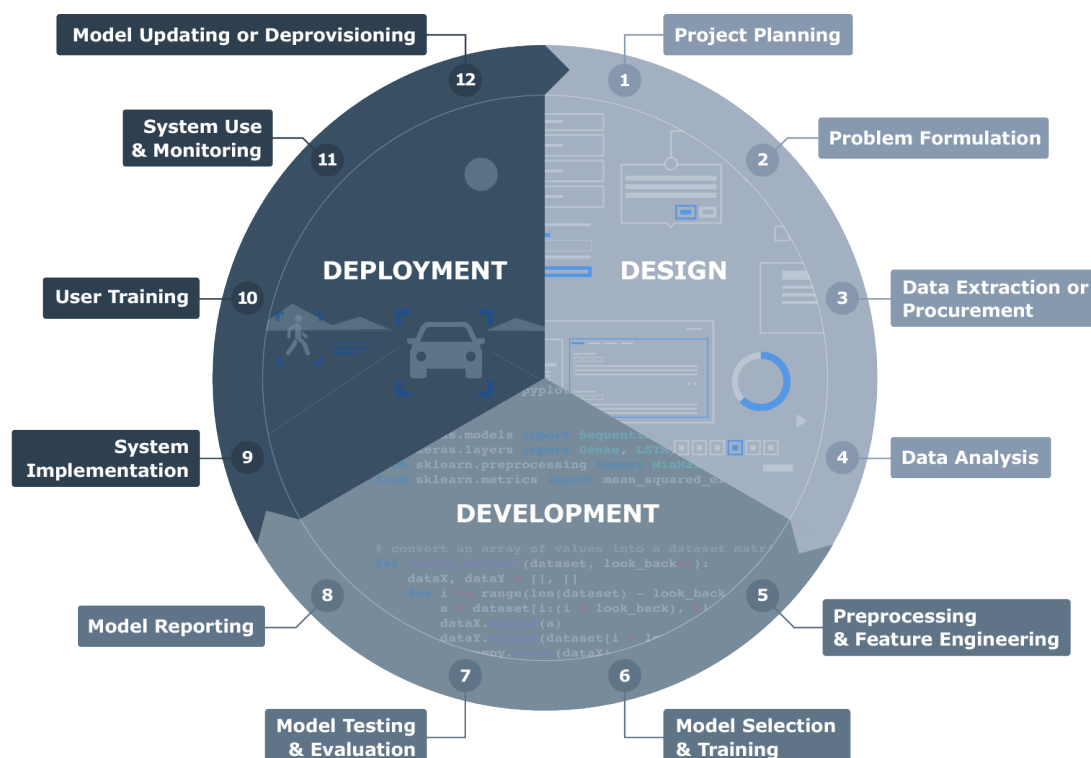
Section 4. Action and Decision-Making

In this Section

- ✓ Making decisions through the project lifecycle
- ✓ Ongoing governance

Using the Project Lifecycle Model

Having conducted the recommended deliberation and engagement activities and having integrated data protection and intellectual property considerations into your governance framework, the next step is to channel the learnings into ethically-guided action. We revisit the stages of the AI Project Lifecycle to identify ethical risks and opportunities for mitigation. The AI Project Lifecycle model provides the conceptual basis for this governance stage.



The table below contains detailed descriptions of the stages of the Project Lifecycle.

Lower-level lifecycle stage	Details
Design	
Project Planning	<p>Description</p> <p>The project planning task encompasses the preliminary activities that are intended to help determine the aims, objectives, scope, and processes associated with the project, including an assessment of the potential risks and benefits.</p> <p>Ethical Significance</p> <p>Creates a space for anticipatory and reflective activities that help create a stable foundation for the project.</p> <ul style="list-style-type: none"> • Offers and opportunity for the team to agree on any “red lines” (e.g. contexts or domains in which a system should not be used, data types that are not permissible to collect or use). • Allows project team to set milestones and objectives that can be used throughout the project to determine if their original goals have been achieved.
Problem Formulation	<p>Description</p> <p>This task involves the formulation a clear statement about the overarching problem the target system or project seeks to address (e.g. a research statement or system specification) and a lower level description of the computational procedure that instantiates it (e.g. a functional mapping from input to output variables and explanation about why it is appropriate).</p> <p>Use of the term ‘problem’ is intended to focus attention on the fact that the project team is attempting to solve a problem, rather than just build a novel system. This helps to avoid the bias of ‘Maslow’s Hammer’, in which you have a pre-existing solution (the hammer) and go looking for a problem (a nail) to solve, regardless of whether it is the right tool for the job.</p> <p>Ethical Significance</p> <p>The importance of this stage is split across the two interlocking understandings of the term “problem”:</p>

	<ul style="list-style-type: none"> • As a statement about a well-defined computational process (or a higher-level abstraction of the process), this task helps identify threats to the validity and legitimacy of the project. For example, an algorithmic system that attempts to predict a job candidate's 'employability' (the target variable) on the basis of a model trained on biased data from historical hiring practices may be perceived as unjust. • As a statement about how the system attempts to address a wider practical, social, or policy issue, this task helps the project team determine if their goal is valid and if the target system is sufficient to achieve their goal. It can also support stakeholder engagement and project communication activities.
Data Extraction (Or Procurement)	<p>Description</p> <p>By 'data extraction' we refer to both the design of an experimental method or decisions about data gathering and collection, based on the planning and problem formulation from the previous steps, as well as the actual extraction and storage of novel data or the procurement of existing data.</p> <p>Ethical Significance</p> <p>The well-known principle of 'garbage-in, garbage-out' summarises the importance of this task nicely.</p> <p>As data-driven technologies, ML algorithms or AI systems depend on the data fed into them. However, due diligence at this stage is important for reasons other than statistical validity. Responsible data extraction is, among other reasons, vital for the design of accountable and trustworthy services, the development of safe, fair, and explainable algorithms, and the deployment of sustainable and privacy-preserving systems.</p>
Data Analysis	<p>Description</p> <p>Data analysis is typically split into two types: exploratory and confirmatory analysis:</p> <ul style="list-style-type: none"> • Exploratory data analysis allows analysts to better understand the structure and content of the dataset, and identify possible associations between data types and variables. • Confirmatory data analysis is where initial hypotheses that are developed in the previous stage are evaluated using a variety of statistical methods (e.g. significance testing). <p>Ethical Significance</p> <p>In the context of responsible research and innovation, data analysis is vital to the assessment of myriad biases that can negatively</p>

	<p>impact a project, many of which are most obvious at this stage in a project.</p> <p>Identifying and dealing with missing data is particularly important during this task. Although upstream stakeholder engagement activities can help mitigate the impact of this bias, identifying the scope of its impact and determining how effectively it can be addressed (e.g. using various methods imputation, collecting additional data), will largely depend on the quality of the data analysis task.</p>
<p>Pre-processing and Feature Engineering</p>	<p>Description</p> <p>Whereas data analysis can give rise to valuable insights (e.g. business intelligence), the data structures are not always appropriate to train ML algorithms. Therefore, 'pre-processing and feature engineering' is required to clean, normalise, or otherwise refactor data into the features that will be used in model training and testing, as well as the features that may be used in the final system.</p> <p>Ethical Significance</p> <p>Features are dependent upon, but separate from, the raw data that are collected in the prior stages. They can be engineered by hand or by using algorithmic techniques to improve the performance of subsequent ML processes.</p> <p>However, the features that are used in the process of model training, for instance, do not only affect the model's accuracy or predictive power; they also impact the ethical consequences of the project (e.g. reducing explanatory potential of system, creating discriminatory outcomes). Therefore, selecting the best features is a vital, albeit often time-consuming and complicated task can involve trade-offs about which parameter to optimise for (e.g. predictive power versus interpretability).</p>
<p>Development</p>	
<p>Model Selection & Training</p>	<p>Description</p> <p>Simply put, this stage involves the selection of a particular algorithm (or multiple algorithms) for training a model.</p> <p>There are many factors that feed into this decision, including (but not limited to):</p> <ul style="list-style-type: none"> • Access to computational resources (some learning algorithms require vast levels of computational power) • Predictive performance of model (as compared to other models)

	<ul style="list-style-type: none"> • Properties of underlying data (e.g. is the size of the dataset sufficient) <p>Ethical Significance</p> <p>There are, of course, many technical and logistical reasons for the responsible selecting and training of a model (e.g. ensuring parsimony, optimising performance).</p> <p>However, a key concept in the responsible development of a model is the inherent interpretability and post hoc explainability of the model and the behaviour of the system into which it is implemented. Although there are nuances and exceptions, it is generally the case that more complex models are harder to interpret and explain (e.g. linear regression versus convolutional neural networks). Selecting the right technique, therefore, depends on the ultimate use case of the model and system.</p>
<p>Model Testing & Validation</p>	<p>Description</p> <p>Model testing and validation involves the assessment of a model against a variety of metrics, which may include the evaluation of the model's accuracy as applied to novel data (held out from the original training data).</p> <p>Ethical Significance</p> <p>Where a dataset is split into testing and training (i.e. internal validation), or where a model's performance is evaluated against wholly new data (e.g. external validation from a separate trial or project team), there are options to assess more than just the model's performance.</p> <p>Testing the generalisability of a model to a new domain or context can also help ensure the model is both sustainable and fair (e.g. has similar levels of accuracy or performance when validated externally).</p>
<p>Establishing Monitoring Tests</p>	<p>Description</p> <p>Successful monitoring of the performance of a model in its runtime environment depends on the prior establishment of metrics to test whether the model is still operating within the respective parameters.</p> <p>While these tests will invariably measure properties such as accuracy or global interpretability, there could also be a need to establish tests that address the performance of the model at a system level (e.g. efficiency, compute resources). Those who were responsible for the model will likely have expertise that is required for this stage, alongside the involvement of downstream team members, such as systems and software engineers or end users.</p> <p>Ethical Significance</p>

	<p>This stage is a vital component of establishing a collective responsibility over the lifecycle of a project. This is because such a task exists at a key juncture for most projects, and depends on and in turn supports clear and accessible forms of communication (see next stage).</p>
<p>Model Documentation</p>	<p>Description</p> <p>This task involves the documentation of both formal and non-formal properties of the model and the processes by which it was developed. This includes (but is not limited to):</p> <ul style="list-style-type: none"> • Data sources and summary statistics • Model used (e.g. proprietary model purchased from vendor) • Evaluation metrics (e.g. model performance) <p>Ethical Significance</p> <p>Clear and accessible documentation is an important form of responsible project governance for the following reasons:</p> <ul style="list-style-type: none"> • In research projects it ensures reproducibility and replicability of results, as well as other values associated with open research, such as public accessibility. • In development projects it ensures accountability and transparency of decision-making. • In all projects it can help affected individuals seek redress for any harms that may arise from the design, development, or deployment of data-driven technologies.
<p>Deployment</p>	
<p>System Design & Implementation</p>	<p>Description</p> <p>System design and implementation is the process of putting a model into production, and implementing the resulting system into an operational environment. The system enables and structures interaction with the model, within the environment (e.g. a recommender system that converts movie ratings into recommendations for future movies of interest for a specific user).</p> <p>Ethical Significance</p> <p>Regardless of how well the preceding stages have gone, unless the encompassing system is implemented effectively, the model's performance will be impacted. Here, we can note the importance of two forms of implementation:</p> <ul style="list-style-type: none"> • Technical implementation: designing and building the hardware and software infrastructure (e.g. server,

	<p>interfaces) that will host the model. Among other things, it is important to ensure the technical system is secure, performant, accessible.</p> <ul style="list-style-type: none"> • Social or organisational implementation: how the technical system is situated within broader social and organisational practices is also important when considering the project's goals and objectives (e.g. appropriately informed users, complementarity with organisational practices).
<p>User Training</p>	<p>Description</p> <p>'User training' includes any form of support or upskilling that is offered and carried out by the individuals or groups who are required to operate a data-driven technology (perhaps in a safety-critical context), or who are likely to use the system (e.g. consumers).</p> <p>Ethical Significance</p> <p>User training is rarely carried out by the same team members who designed and developed the system. While developers may produce documentation for the model (see above), this is often insufficient as a form of user training—additional forms of formal training workshops or courses may be required depending on the complexity of the system.</p> <p>Insufficient or inadequate training can create conditions in which cognitive biases such as algorithmic aversion thrive (e.g. users do not trust the performance or behaviours of a trustworthy algorithmic system, or users trust the outputs of an untrustworthy system).</p>
<p>System Use & Monitoring</p>	<p>Description</p> <p>Depending on how an AI system has been designed, its deployment and use in an environment (physical or virtual) can create conditions for ongoing feedback and learning (e.g. robotic systems that employ reinforcement learning, digital twins linked to a monitored counterpart). Regardless, the use of metrics and evaluative methods are commonly used to monitor the performance of a system and ensure that it retains (or ideally improves on) the same level of performance that it had when first validated.</p> <p>Ethical Significance</p> <p>The potentially dynamic (and sometimes unpredictable) behaviour of machine learning models and AI systems means that ongoing monitoring and feedback of the system, either automated or probed, is important to ensure that issues such as model drift have not affected performance or resulted in harms to individuals or groups.</p>
<p>Model Updating & De-commissioning</p>	<p>Description</p> <p>If the use and monitoring of a model or system identifies vulnerabilities or inadequate levels of performance, it may be</p>

	<p>necessary to either update the model through retraining (i.e. looping back through some of the model development tasks) or deprovision the system if it is no longer fit for purpose.</p> <p>Where the latter option occurs, the project team or organisation may need to commence a new project lifecycle to address any gaps in their business or organisation that arise because of the deprovisioning of the present system.</p> <p>Ethical Significance</p> <p>An algorithmic model that adapts its behaviour over time or context may require updating or deprovisioning (i.e. removing from the production environment). While this can include elements such as improvements to the system’s architecture (e.g. for speed or security), the more important component here is the model itself (e.g. the model parameters, the features used).</p> <p>An important issue to address is model drift, which can arise because of changes to the underlying data distribution used to train the original model (e.g. average values of house prices in a fluctuating property market) or because the semantic meaning of the features has changed due to shifting societal practices or norms (e.g. political or geographic boundaries).</p>
--	--

Iterative Development in the Project Lifecycle

It is important to note that, in practice, a project lifecycle is non-linear. During ethical reflection and deliberation, the lifecycle of a specific project may be addressed in the order of actual practice. The lifecycle should be considered iteratively to encompass the project’s own design-development-implementation cycles.

For example, system implementation comes before user training because there has to be a system in place for users to be trained on. However, this does not mean the system goes live prior to training or that each step is only done once.

Additionally, data analysis presumes some initial cleaning and pre-processing, so there will be iteration between these stages.

How different stages of the project lifecycle is iterated will depend on your project and project team.

To give one illustrative example, consider the following hypothetical scenario. The estates team for the Department for Business and Trade’s buildings are exploring the use of AI-enabled facial recognition systems or other forms of biometric data to automate the identification of staff and visitors and ensure authorised access to their building. In designing and developing this system, the project team carry out an assessment of the possible factors that could prevent authorised people from accessing the building. The factors they identify include:

- **Variation in available hardware and software across the buildings (e.g. variations in IT capacity)**
- **Potential unequal distributions of false negatives from the AI system due to how well the system can handle variations in biometric data (e.g. problems identifying users with darker skin colours through facial recognition)**
- **Inability of staff to explain to visitors any false negatives in a transparent and accessible manner due to a lack of training**

Each of these factors serve as a source of uncertainty across the project lifecycle, which could undermine the project’s goals and objectives (e.g. *sustainability* of the system, *fairness* of the classification algorithm, *explainability* of the system). As they are factors that could negatively impact the realisation of the project’s goals and objectives, or increase the uncertainty of achieving them, we can call such factors ‘risk factors’.

In contrast, factors that minimise these risks and increase the likelihood of achieving the project’s goals and objectives are enabling factors or opportunities.

Therefore, when we speak about risk management, we are referring to processes that address the identified risks and associated risk factors through regulatory, policy, or technological means

Consider the table below, which shows possible risks and opportunities for a hypothetical project involving a predictive algorithm that is used to identify criminal defendants who are at risk of reoffending.

Risks and opportunities associated with each of the project lifecycle stages

Project Lifecycle Stage	Risk	Opportunity
Project Planning	Underestimating the amount of resources or technical skills required to complete the projects and achieve objectives.	Early identification of potential stakeholders to obtain their buy-in and ensure the project aligns with their needs and goals.
Problem Formulation	Defining the problem (social and technical) too narrowly, resulting in a model and system that is not fit-for-purpose.	Identifying additional outcomes of interest, which go beyond the original target variable (i.e. risk of recidivism).
Data Extraction or Procurement	Obtaining biased or incomplete data that could lead to inaccurate predictions and perpetuate	Obtaining additional data that could improve the accuracy and robustness of the model.

Project Lifecycle Stage	Risk	Opportunity
Data Analysis	existing biases (e.g. racial profiling). Failing to fully explore the data and missing key insights or patterns because of a lack of awareness of cognitive or social biases.	Identifying unexpected correlations or patterns in the data, which could help inform feature engineering or the identification of new predictors.
Pre-processing and Feature Engineering	Overfitting the model by creating features that are too specific to the training data and do not generalize well to new contexts (e.g. other prisons or different cities).	Exploiting computational techniques to help engineer features that could improve the accuracy of the model and reduce biased judgement or decision-making.
Model Selection and Training	Selecting a model that is too complex and has low levels of interpretability, while also offering marginal improvements in performance.	Selecting a model that is both accurate and interpretable, allowing for greater insight into the factors that contribute to recidivism as well as supporting decision-making.
Model Testing and Validation	Not using robust forms of external validation (e.g. new datasets from different cohorts or contexts) to ensure robustness and generalisability of the model.	Performing rigorous testing and validation—perhaps supported by “red teams”—to ensure that the model is reliable, fair, and explainable.
Establishing Monitoring Tests	Failing to choose appropriate metrics to monitor the model over time, resulting in a degradation of performance beyond standard metrics (e.g. increasing levels of discrimination for sub-groups).	Identifying broader and sustainable metrics, and ensuring complementary organisational processes, which collectively help detect changes in the data or the performance of the model in good time to allow for proactive adjustments.
Model Documentation	Failing to document the model clearly enough to ensure all stakeholders can access and understand how the model was developed and why certain choices were made, resulting in a lack of trust and use of the model.	Releasing reproducible, replicable, and usable code and documentation to allow others to support ongoing development and improvements.
System Design and Implementation	Failing to consider readiness barriers or obstacles for downstream teams (e.g. reliance on legacy systems and incompatible data pipelines).	Early stakeholder engagement to ensure that the design and implementation of the system takes into account the needs and perspectives of all stakeholders (e.g. bringing stakeholders into upstream conversations).
User Training	Failing to provide adequate training to users of the system,	Providing ongoing and comprehensive training to users

Project Lifecycle Stage	Risk	Opportunity
	leading to inappropriate or ineffective use of the model that exacerbate biases (e.g. decision-automation bias).	of the system to ensure that they understand the model, how to use it effectively, and have had a chance to shape its design through participatory design workshops.
System Use and Monitoring	Failing to monitor how the model adapts to deployment in a runtime environment (i.e. deployed in the real world), leading to unintended consequences and possible harm to people.	Effective and broad monitoring to ensure sustainability and early identification of opportunities for improvement (e.g. addition of new features).
Model Updating or Decommissioning	Failing to have a plan for decommissioning that creates “technological lock-in” and degradation in the system’s performance.	Updating the model in a timely manner to adjust to changes in social patterns or behaviour that were not present in the original training data.

The above table gives a broad range of risks and opportunities, but if a project is only considering one principle (e.g. fairness), then it is suggested that the risks and opportunities that a project team identifies during anticipatory reflection are focused on those that are relevant to that principle (e.g. those that undermine or support the principle and its core attributes). In addition, emphasis should be placed on the *preliminary* part of this risk management plan, because during scoping and anticipatory reflection it is likely that a team will not have been able to sufficiently identify the project’s stakeholders.

As such, without the diverse and meaningful engagement of stakeholders it is likely that a project team will have missed some of the risks and opportunities associated with a project. This is where the preliminary forms of anticipatory reflection bleed into inclusive and diverse forms of deliberation and engagement.

Return to Deliberation and Engagement

In the [Deliberation and Engagement](#) section, you were introduced to participatory processes that enable a project team, with the involvement of stakeholders (and affected users) to refine the preliminary risks and opportunities identified during scoping and anticipatory reflection stage.

There are several tools, mechanisms, and processes that can be used to support deliberation and engagement. The most significant is a robust plan for diverse and inclusive stakeholder engagement.

Beyond stakeholder engagement, there are myriad tools that can assist with deliberation and engagement once relevant goals and objectives have been established, such as:

- **Bias Self-Assessment:** As part of this tool a series of deliberative prompts are provided that can help guide a project team through a series of questions to help identify and mitigate biases in the project.
- **Readiness Self-Assessment:** This tool is an extension of the Stakeholder Engagement Process and has been designed to help individuals and teams understand the degree of preparedness to accept AI/ML systems or practices amongst key internal and external stakeholders.
- **Worked examples:** Worked examples are provided to help project teams gain insight into ethical deliberation.

Supplementary Questions to Consider

In addition to the material presented in this section, you and your project team may have other questions related to wider project lifecycle processes that you may want to address. Some of these questions include:

- What precedes the project lifecycle (e.g. commissioning, minister approval, etc.)?
- Should iterative cycles exist between the stages (e.g. outputs from training and testing may feed back into data analysis)?
- How can this model be used by people with different roles and responsibilities? How should disagreement about the ordering of the stages or their labelling be addressed?
- Automated checks may be used to monitor system performance. Where do these get decided? Documentation should also include clear indication of who is responsible for monitoring the system.
- Some exploratory projects may not complete the full lifecycle. For example, a project may be abandoned if it is not feasible. Should anything be added to address this?
- How do you monitor and respond to changes to the regulatory environment or ministerial priorities?

Map Governance Workflow

At this stage, a project's design, development, and deployment activities are underway. Process-based governance implies that governance activities take place during the processes that bring an AI project into production and use, rather than retrospectively. At each major phase of design, development, and deployment, project governors should hold meetings, workshops, and other engagements to explore ethical issues, mitigation strategies, and onward monitoring and re-evaluation, documenting it in the PBG log.

To ensure accountability through the PBG framework, you should routinely revisit the PS Report and PBG log to update the governance workflow. In so doing, respond to the following questions. Where there are identified gaps, return to earlier steps in the framework.

The table below contains a set of ongoing governance questions to be considered for each AI project. Project teams may wish to supplement this list with their own

questions that are specific to team practices, organisational policies, and the specific requirements of the AI project.

- a. Do established governance actions proportionally mitigate possible harms to stakeholders posed by this project? If not, how can your PBG framework be rectified to address these potential harms?
- b. Does this distribution of responsibilities outlined in the PBG Framework establish a continuous chain of human accountability throughout the design, development, and deployment of this project? If not, how can any identified gaps or breaks in the chain be rectified in the PBG Framework?
- c. How will you ensure that all team members, who are assigned roles/responsibilities understand the roles/responsibilities that have been assigned to them?
- d. If you are procuring parts or elements of the system from third-party vendors, suppliers, sub-contractors, or external developers, how are you instituting appropriate governance controls that will establish end-to-end accountability, traceability, and auditability for these procured parts or elements?
- e. If any data being used in the production of the AI system will be acquired from a vendor, supplier, or third party, how are you instituting appropriate governance controls that will establish end-to-end accountability, answerability, and auditability across the data lifecycle?

Governance workflow questions

Section 5: Ongoing Governance



Mechanisms for monitoring, evaluation, and communication

Once an AI system is in production, the team should identify mechanisms for monitoring and evaluating whether the ethical principles identified and operationalised through this governance framework are being upheld, and to ensure that risks to fundamental rights and interests are minimised. This is an ongoing process that begins during initial project scoping and continues until the AI system or service has been deprovisioned.

Stakeholder Impact Assessments

After your AI system has gone live, your team should iteratively revisit and re-evaluate your **Stakeholder Impact Assessment (SIA)**. These check-ins should be logged on the Deployment Phase section of the SIA with any applicable changes added and discussed. Deployment-Phase SIAs should focus both on evaluating the existing SIA against real world impacts and on considering how to mitigate the unintended consequences that may have ensued in the wake of the deployment of the system. As with each SIA iteration, the PS report is revisited at this point, when objectives, methods, and timeframes for the next Deployment Phase SIA are established. You should keep in mind that, in its specific focus on social and ethical sustainability, your Stakeholder Impact Assessment constitutes just one part of the governance platform for your AI project and should be a complement to your accountability framework and other auditing and activity-monitoring documentation.

Your SIA should be broken down into four sections of questions and responses.

- In the 1st section, there should be general questions about the possible big-picture social and ethical impacts of the use of the AI system you plan to build.
- In the 2nd section, your team should collaboratively formulate relevant sector-specific and use case-specific questions about the impact of the AI system on affected stakeholders.

- The 3rd section should provide answers to the additional questions relevant to pre-implementation evaluation.
- The 4th section should provide the opportunity for members of your team to reassess the system in light of its real-world impacts, public input, and possible unintended consequences.

Metrics and Indicators

The project team should conduct regular performance tests in addition to impact assessments to test if the model is responding well to real-world data, it is serving its intended purpose, and it is being used responsibly. The users of the model will report back on whether they find the system to be useful, reliable, and accurate, amongst other metrics. The team will monitor and evaluate the results and feedback received and decide if any adjustments are necessary (e.g. retraining the model or changing the system according to new policy requirements). If the model is not performing to standards, revisit earlier phases and make adjustments or deprovision the model if needed.

Where qualitative or quantitative measures can be established to evaluate the project's continual adherence to ethical principles, these can serve as indicators or thresholds for monitoring whether the project is still aligned with each principle (e.g. number of safety incidents for *sustainability*, amount of project documentation openly available for *accountability*, sub-group accuracy rates for *fairness*, feedback from users about a model's interpretability for *explainability*, and timeliness or recency of the training data for *data stewardship*).

Using metrics and indicators responsibly

It is important to note that the use of metrics and indicators can be problematic if they are used without awareness of their limitations. The examples given above, for instance, only capture a small part of the complexity for each principle. In the case of *sustainability*, for example, the number of safety incidents may be a useful indicator of the principle or a specific core attribute, but it is not a complete measure of the principle as a whole.

In addition, the over-reliance on metrics and indicators can create a poorly calibrated incentive structure where people shift their attention solely to meeting some narrow metric, rather than focusing on the broader goal. This is often known as 'Goodhart's Law', which states that "when a measure becomes a target, it ceases to be a good measure".

Peer Review

Regular forms of peer review can be a reliable and robust mechanism for monitoring and evaluating whether ethical principles are being upheld. This can include the following:


- **Committee review:** A committee of experts can be established to review the project at regular intervals.
- **Red teams:** A team that is specifically tasked with finding problems can be a useful mechanism for identifying potential issues and can also help address cognitive biases such as confirmation bias or self-assessment bias (see the [Appendix J: Bias Self-Assessment](#) tool for more information).
- **External stakeholder engagement:** Ongoing engagement with external stakeholders (e.g. through evaluative workshops) can help ensure that the project is aligned with the relevant principle(s), and also help to ensure diverse perspectives are taken into account.
- **Code review and auditing:** Having an independent expert review the code and algorithms used in the data-driven technology to identify potential ethical concerns, such as bias or discrimination.

Communication

An important component of ethical practice is to establish routines of communication about the project's governance and accountability. Documentation will have been produced at various stages of the project's lifecycle. However, at key points, the team should review the project's governance and communicate the outcomes to relevant or requesting stakeholders. This includes understanding and communicating how the outcome of the system may impact different systems and decision-making processes within the Department.

This communication can achieve several goals, including:

- Reaching out to new stakeholders or parties for the specific purpose of conducting an impartial or independent review of the project.
- Obtaining feedback from relevant communities who may be interested in contributing (e.g. other teams within the organisation, or external stakeholders involved in similar work).
- Building awareness of the project and its results with a wider audience
- Ensuring compliance with relevant regulations and standards (e.g. quality assurance auditing).
- Building trust among stakeholders and users.
- Responding to enquiries about the governance of your project.

 **Non-linear processes**

It is important to reiterate that these processes are presented in a linear fashion for the sake of clarity, because they are high-level or macroscopic processes. In reality, the picture is more complex, and the tasks that fall within each stage are likely iterative and non-linear (e.g. reviewing a preliminary model may result in unexpected (and insufficient) levels of performance that require a project team to go back and re-evaluate their initial plans).

Section 6: Worked Examples

GenAI: To Use or Not to Use?

ChatGPT is a chat agent built by OpenAI and is a form of Generative AI (GenAI). Generative AI is a set of relatively new technologies that leverages large (very large) volumes of data along with some machine learning (ML) techniques to produce content based on inputs from the users known as prompts.

ChatGPT brought significant attention to the capabilities of GenAI and has accelerated the development and its appearance in consumer-facing apps and systems. This form of GenAI is also known as a large language model (LLM) because it mainly uses and acts upon data based on human language. ChatGPT is being used in a range of ways, including to create textual content, summarise information, and generate code. Other companies, including GoogleAI have also produced GenAI/LLM systems with similar capabilities. These systems are built on sophisticated machine learning models that operate on vast amounts of data to generate text by recognising statistical patterns in the input data.

Although GenAI systems can be used to speed up tasks such as the ones listed above, this technology illuminates a range of ethical considerations. First, the training data used in the development of GenAI models typically relies on data largely “scraped” from the Internet. As revealed by researchers looking into other large data sets used to train different types of AI, the textual training data used for LLM systems like ChatGPT likely includes contains toxic, racist, misogynistic, and conspiracy theoretical material. This contributes to the risk that the systems will produce biased and inaccurate information. The characteristics of the training data shape the outputs of the systems in unpredictable and can produce results with significant consequences on people’s lives and wellbeing.

In addition to the potential for bias and inaccuracy, there are additional risks in relying on the content produced by the latest wave of LLMs. ChatGPT has been shown to ‘hallucinate’ responses that have little to no basis in reality.⁴⁵ This is due in part to the ways LLMs are optimised for producing *plausible* outputs that resemble human speech but lack awareness of the real world and the context of information. These hallucinations exacerbate the risk of misinformation, which can have far-reaching individual and societal implications if people place too much trust in such systems.

Caution advised

Human oversight over any technology that automatically generates information is crucial – especially with LLM-powered chat agents like ChatGPT. As ChatGPT has been shown to render inaccurate, misleading, and harmful results, its outputs must be thoroughly reviewed by the user. In many cases, the risk and effort required to fact-check the outputs of any chat agent may outweigh the benefits of using the chat agent.⁴⁶

The lack of transparency from the companies who build chat agents like ChatGPT about how they function and the details of their training data makes it challenging to trust these technologies for ethically sensitive uses.

Traffic Light System

To evaluate the risks of using particular forms of AI, we recommend using the traffic light system approach detailed below. This is aligned with the recent UK Government 'Guidance to civil servants on use of generative AI', the procurement decision tool, and the PBG framework.

! Consider context

This traffic light system is best understood as a deliberative support for decision-making. The decision to use any AI system is *context dependent*. Each system adoption should be considered independently with an appreciation of the specific case context.

The traffic light system approach is explored further in the three worked examples below. The indicators are explained as follows:

Green:

System use is appropriate. Low risk of ethical or reputational harm. Anticipatory and ongoing human oversight and accountability protocols should be followed using PBG Framework guidelines and assessments.

Amber:

Caution advised, potential risks of using target system identified. In-depth risk analysis and stakeholder impact assessments leading to mitigation measures required before use. Anticipatory and ongoing human oversight and accountability protocols should be followed using PBG Framework guidelines and assessments.

Red:

Use not advised. Identified adverse impacts and risks outweigh potential benefits.

Green scenario

Using a fully licensed generative AI system to produce graphics or audio for a report, publication, or presentation.

- No sensitive or confidential information should be input into a system not wholly owned and operated by a UK government entity.
- Human review and oversight are mandatory to ensure that all generated material is accurate and appropriate for sharing with the public.

- Visual or audio content that represents actual persons, places, or events must be properly identified as “computer-generated” or similar to avoid misleading audiences.
- The use of generated content should be made clear to all audiences.
- A living person should bear responsibility to respond to reports of copyright infringement, misrepresentation, or other ethical concern.

Amber scenario

Training a fully licensed generative AI system on financial data to conduct analysis of economic trends for use in a report or publication.

- No sensitive or confidential information should be input into a system not wholly owned and operated by a UK government entity.
- Human review and oversight are mandatory to ensure that the resulting analysis is accurate and appropriate for sharing with audiences.
- Generated content or analysis should be identified as such in the publication.
- A living person should bear responsibility to respond to reports of inaccuracy, misrepresentation, or other ethical concern.

Red scenario

Using a generative AI system to respond to public enquiries without human supervisions on factual matters about which the requestor could be held liable.

- Generative AI systems are at risk of producing inaccurate content due their inability to understand the context of their source data (i.e. cannot tell fact from fiction).
- As generative AI systems are not entirely reliable, their use to provide information to members of the public for which they may be held liable does not meet the ethical obligations of government.

Case Study 1: Green

Case Study Description

A government department has proposed an AI-enabled chatbot for individual consumers and businesses to navigate and identify reports, guidelines, protocols, advisory notices, and best practices crafted by government bodies. The chatbot includes an additional feature to translate items into select languages, including Arabic, Bengali, French, German, Hindi, Punjabi, and Spanish. A chat-based application was suggested to improve efficiency in public administration tasks such as responding to individual requests for information as well as user experience on government portals.

Technology Description

The chatbot is a software programme capable of interacting with users using natural human languages in a conversational manner that simulates human conversation. The chatbot service includes predefined menu-based options for users to pinpoint their queries to specific domains or the nature of information (i.e. report or guidance). The chatbot was pre-trained by a third-party provider to provide its basic functions and then further trained on publicly available publications hosted on Gov.UK. The chatbot responds to queries by assigning probability scores to words and determining the final structure of a plausible sentence.

Key Issues

The developed chatbot was trialled on a group of small business owners and individuals from different departments of the civil service. The chatbot performed with a high rate of accuracy on prompts in the English language. In most cases, the users were able to identify and access the specific documents which they required. When the queries were more general, further prompting was necessary to locate the required documents.

The chatbot's performance metrics varied across other languages. In the case of German, French, and Spanish, the performance was nearly but not quite as accurate as its performance in English. Users noted that the responses from the chatbot failed to provide accurate information wherein the output text followed English language grammar rules rather than specific syntax used in German, French, and Spanish. For Bengali, Punjabi, and Hindi, the chatbot struggled to understand prompts which used transliterations, alongside poor performance on syntax and semantics.

Additionally, the chatbot did not reliably direct users to all pertinent resources. In some cases, a query that included the title of a publication did not identify the primary publication but instead pointed to other publications that mentioned it. This result was found to occur across language types.

In sum, for English-speaking users, the model was satisfactory despite requiring additional information and sometimes failing to identify all of the responsive material. The users noted that the chatbot was easier to use than navigating the government websites. On the other hand, users who interacted in non-English languages preferred interacting in English (if that was an option) to avoid the burdens of additional explanations and misleading information.

Deliberative Prompts

- What ethical principles and core attributes are raised by the use of an LLM for this purpose?
- What are the risk factors and potential harms that could occur?
- What is the scale, scope, likelihood of potential harm?
- Which groups and communities will be affected by the use of your model?

- Are there groups or communities that will be excluded from your model or experience barriers to using your system? If so, why?
- What have you learned from stakeholder groups to understand to verify that your system is responsive to the broadest set of needs?

Stakeholders Affected

- Fluent English-speaking users
- Non-fluent English-speaking users
- Non-English-speaking users

Final Use Case Status & Rationale

GREEN: The model may be approved for use after the following recommendations and oversight mechanisms are incorporated:

1. Translation feature is limited to languages which achieve a 95 per cent or higher rate of accuracy.
2. Inclusion of helpline numbers and additional contact details to be shared before the chat is closed.
3. Users must accept terms and conditions of using the chatbot before interactions begin. These terms will include warnings of generated content, liability waivers, and links to further resources such as a fair data use policy and guidance on chatbot interactions.
4. Regular auditing of the system required including routine evaluation of system functionality across supported languages.
5. Efforts should be made to improve non-English language capabilities.
6. Personnel shall be assigned responsibility for monitoring performance and responding to user and other feedback.

Case Study 2: AMBER

Case Study Description

Users experimenting with and LLM have found it can be a useful writing companion. Government agencies often rely on accurate and unbiased reports to make informed policy decisions. LLMs can generate text content, including entire reports or report sections. As the tasks to conduct research, aggregate findings, and write reports can take up considerable time, the use of an LLM could increase a worker's efficiency by automating much or all of the process.

Technology Description

The LLM-based chatbot is a software programme capable of interacting with users using natural human languages in a conversational manner that simulates human conversation. The chatbot responds to queries by assigning probability scores to words and determining the final structure of a plausible sentence. The chatbot service accepts user prompts and can generate report-length strings of text while processing additional data that it has been directed to.

Key Issues

LLM systems can be inaccurate. LLMs do not understand context and may misinterpret information or use information intended for a different purpose.

LLM systems learn from historical data that may be reflect human bias, that can in turn be expressed in the report's content.

LLMs are powerful data processors but have been known to manufacture non-existent citations. They also may not be able to distinguish between high and low quality data (including recency or completeness) increasing the risk of inaccurate findings and conclusions.

LLM systems are typically offered by third-parties and hosted on privately-held domestic and overseas servers. The security of information sent to such systems cannot be guaranteed.

The credibility of the department may rest on the perceived accuracy and objectivity of the reports it produces. Quality standards must be sustained to maintain this credibility.

Deliberative Prompts

- How are the ethical principles of *sustainability, accountability, and explainability* implicated by this use of an LLM?
- What are the risk factors and potential harms that could occur by using this system?
- What is the scale, scope, likelihood of potential harm?
- Who is most likely to experience the harms and to what extent?
- What insights have been provided by stakeholders who may be affected by inaccurate or biased reports?
- What would be the reputational risk to DBT in producing an inaccurate, biased, or out of date report?

Stakeholders Affected

- DBT personnel and management.
- Report audiences.
- Other interested parties who may be affected by policy guidance or advice.

Final Use Case Status & Rationale

AMBER: The model is approved for use with the following cautions and safeguards incorporated:

An LLM can be used to generate first-drafts of report content, which may save an author time, but authors should not become over-reliant on such a system.

1. Human oversight is required. All content must be verified for accuracy by experienced personnel.
2. Citations and other data sources must be verified before publication.
3. Any findings or conclusions generated by the LLM must be verifiable and explainable.
4. Report readers must be made aware that an LLM was used to support the writing of the report.
5. Human authors must assume final responsibility for the report's content.
6. Sensitive data should never be sent to privately-held servers and systems.

Case Study 3: RED

Case Study Description

To meet government imperatives to overcome human resource and costs constraints in public administration, DBT proposes using an LLM to automate some data analysis tasks and help staff reduce their workload. The proposed LLM would be used to synthesize domestic and international information discovered online to predict international supply chain fluctuations to produce a dashboard of guidance for policy-makers and investors.

Technology Description

LLM systems are capable of synthesising numerical and textual data through simplified commands. Project teams can interact with the LLM and use their outputs to further analyse market outcomes, employment trends, public expenditure, and consumer preferences, likely completing tasks faster than manual calculations. Furthermore, LLMs can summarise research from diverse sources and convert it into accessible formats.

Key Issues

LLMs are trained on *historical* data and require regular retraining and fine tuning to be up to date and to accurately reflect contemporary developments and policy priorities.

LLMs are not actually 'intelligent'. They are complex algorithms designed to mimic language use by predicting what string of words should follow a prompt. LLMs do not understand context and may assemble sentences that sound factually correct but are in fact fabrications.

LLMs can be very persuasive because they are designed to master language and mimic many speaking and writing styles. This can lead to over-reliance and misplaced trust in system outputs.

LLMs (and GenAI) are very complex systems that can process many forms of data from a large number of sources. It can be difficult to determine how a given output was produced.

The stakes of providing inaccurate or misleading policy or investment advice could result in significant financial losses, reputational harm, and other undesirable results.

Third-party LLM systems are triggered by user prompts. Information submitted as prompts is sent to the system, which may mean sending information to privately-held domestic and overseas servers. Sensitive information could be shared with unauthorised people and companies.

Deliberative Prompts

- How are the ethical principles of *safety*, *accountability*, and *explainability* implicated by this system?
- What are the risk factors and potential harms that could occur by using this system?
- What is the scale, scope, likelihood of potential harm?
- Who is most likely to experience the harms and to what extent?
- What do people who may be affected by inaccurate or hallucinated outputs think about this system and what safeguards would they demand for its use?
- What is the threshold of risk for system users and DBT from reputational or informational harms?

Stakeholders Affected

- DBT staff.
- Consumers of DBT policy and investment advice.
- Persons affected by decisions supported by the outputs of this system.

Final Use Case Status & Rationale

RED: The model is not recommended for this use.

- Potential harms to affected users are not fully understood and/or cannot be mitigated.
- System's black box nature and complex logic make it unlikely its outputs can be fully audited. Accuracy and desired content cannot be assured.
- System outputs cannot be made sufficiently explainable to affected parties.
- High risk of sensitive information being shared with unauthorised persons or companies.

Appendixes

[Appendix A: Glossary](#)

[Appendix B: Process-Base Governance log template](#)

[Appendix C: Project Summary Report template](#)

[Appendix D: Data Factsheet](#)

[Appendix E: AI Procurement Guidance Tool](#)

[Appendix F: COBRA Worksheet](#)

[Appendix G: Stakeholder Impact Assessment](#)

[Appendix H: Readiness Self-Assessment](#)

[Appendix I: SSAFE-D Principles Core Attributes](#)

[Appendix J: Bias Self-Assessment](#)

Appendix A: Glossary

This section provides basic definitions of key terms in AI and other data-driven technologies. These definitions can aid decision-makers in identifying and classifying technologies whose production and use requires governance at different degrees.



The Challenge of Defining AI

There are many ways that AI has been defined over the last several decades, but for the purposes of this guide, we will stick to defining it by describing what it does, i.e. what role it plays in the human world. While more detailed definitions appear below, we start with this:

AI systems are algorithmic models that carry out cognitive or perceptual functions in the world that were previously reserved for thinking, judging, and reasoning human beings.

While AI has existed for some time, recent advances in computing power, coupled with the ever-expanding availability of big data, and the advancement of increasingly sophisticated machine learning algorithms mean that AI designers are able to build systems capable of undertaking increasingly complex tasks.

AI can be difficult to define with precision because it is not a single technology. It is more of a discipline or practice that aims to create computer-based systems that perform complex tasks. Another challenge for defining AI is that it is a concept in a continuous state of evolution. Many technologies we take for granted but that we would not likely call AI today are only possible because of prior AI research. The optical character recognition (OCR) that identifies letters and numbers in printed or image-based text and is built into document software, scanners, and copy machines is based

on key computer vision techniques that were once considered foundational AI research. Similarly, the features of email and messaging systems that offer quick text responses to incoming messages employ natural language processing, which is another long-running AI research topic. Many would not define these and other sophisticated technologies as AI today.

Using definitions to keep track of AI

Amongst the reasons to identify working definitions for AI technologies is to enable those responsible for governance to identify, ethically evaluate, and log AI and ML technologies. This form of documentation is discussed in the Process-Based Governance (PBG) section.

Detailed Definitions

Algorithm

A set of steps that are performed to complete a task or solve a problem. An algorithmic model is a formal representation (e.g. mathematical or logical) of the steps to be undertaken. Algorithms follow steps to process inputs into desired outputs. Humans could also follow an algorithmic process.

Artificial Intelligence Terminology

Artificial Intelligence (AI)

Any algorithmic system or a combination of such systems that uses computational methods derived from statistics or other mathematical techniques and that generates text, sound, image or other content or either assists or replaces human decision-making.

Additional definitions are as follows:

AI systems are algorithmic models that carry out cognitive or perceptual functions in the world that were previously reserved for thinking, judging, and reasoning human beings.⁴⁷

...a **machine-based system** that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.⁴⁸

Generative AI, as the name suggests, generates images, music, speech, code, video or text, while it interprets and manipulates pre-existing data. Generative AI is not a new concept: machine-learning techniques behind generative AI have evolved over the past decade.⁴⁹

While the above definition is a starting point, a more complete understanding of AI comes from detailing its component methods and technologies. A Parliament Select Committee paper on AI provides this explanation:

AI is underpinned by several technologies that enable its adaptiveness and autonomy. Algorithms, which are simply sets of rules and instructions that a system follows in order to perform a certain task, form the programming that tells the system how to operate on its own. Machine learning (ML), which refers to the use of statistical methods to leverage (typically large quantities of) data to evaluate and improve a system's performance in a supervised and/or unsupervised manner (i.e., where data is labelled by a human or unlabelled), allows a system to learn from "experience". Deep learning is a more modern type of machine learning, using artificial neural networks, where processors are linked together like neurons and synapses in the human brain. There have been recent breakthroughs in types of deep learning methods called large language models (LLMs), which use powerful neural networks called transformer models that learn context and meaning by tracking relationships in sequential data (such as relationships between words in a phrase or sentence).⁴⁴ LLMs can recognise, summarise, translate and generate content from massive (internet-scale) datasets with hundreds of billions of parameters; it is OpenAI's LLMs, GPT-3.5 and GPT-4, that are the foundation for its generative AI chatbot ChatGPT.⁵⁰

Being more precise

The main definition above, while widely accepted, can be interpreted quite broadly. It could include technologies such as a spreadsheet that includes formulas and produces graphs—something not usefully defined as AI. For this reason, we provide definitions of key AI technologies and related concepts in the remainder of this glossary.

Artificial General Intelligence (AGI)

A form of AI that is capable of any task a human could undertake, and perhaps more. This is sometimes called “strong AI”. AGI is purely theoretical and there are no currently extant examples. Whether it will ever be possible to develop an AI whose capabilities can be said to be “fully human”, with capabilities including true creativity, contextual understanding, and critical reasoning, is a contested topic in cognitive theory, computer science, and philosophy.

Foundation Model

Foundation models are complex AI models trained on broad data that can be used as standalone systems but are also sufficiently flexible and adaptable to be extended to different types of AI systems. By example, the GPT-3 and 4 models developed by the

company OpenAI can be accessed through consumer-facing applications like ChatGPT but also provide functionality for other AI systems like Microsoft's enhancements to its Bing search tool. The table (Figure 1) is a list of current foundation models and applications at the time of writing.

Frontier AI

Some authors have begun referring to cutting edge innovation in AI, such OpenAI's language model GPT-4 model and Google's BERT, as well as the techniques that underpin them, as 'frontier AI'. Currently this term describes recent work in computer science characterised by the combination of very large data sets, industrial strength compute power, and neural network architectures to autonomously generate text, imagery, sound, and other material of a type formerly produced exclusively by humans.

Generative AI (GenAI)

Computing techniques and tools that can be used to create new content, including text, speech and audio, images and video, computer code, and other digital artifacts by interpreting and manipulating pre-existing data. This is in contrast to AI systems that perform other functions, such as classifying data, grouping data, or choosing actions.

GenAI is not new

The machine-learning techniques behind generative AI have been around for some time but have evolved in the past decade. The latest approach is based on a neural network architecture, coined 'transformers'. Combining transformer architecture with unsupervised learning, generative AI models with more advanced outputs emerged.

Large Language Model

A language model is a type of AI system that works with and represents language. Given a corpus of language, a language model takes as an input a string of words (which may include punctuation and other tokens) and predicts the next word. By repeating this process (autoregression), LLMs can output coherent and plausible statements.

The 'large' part of the term describes the trend towards training language models with more parameters. Recent research has shown that using more data and computational power to train models with more parameters consistently results in better performance. Accordingly, cutting-edge language models trained today might have thousands or even millions of times as many parameters as language models trained ten years ago, hence the description 'large'.

Not all LLMs are GenAI

ChatGPT (the well-known chat agent by OpenAI) and Starcoder (by Hugging Face which produces computer code) are examples of LLMs that are also considered generative AI. However, not all LLMs are considered as such. Contrary examples include models that convert oral speech to text and those that provide direct translation between languages. As these do not produce entirely new arrangements of content, they are not considered generative.

Narrow AI

This term describes all currently possible forms of AI. It is distinguished from AGI in that, at its most sophisticated, it describes AI that can *mimic* human behaviour and carry out tasks described by human operators but cannot be said to be an independent initiator of action. Another way of stating this is to say that currently possible AI lacks independent *agency*. Some computer scientists argue that there is already AI that demonstrates characteristics associated with independent agency, but this is contested.

Data Terminology

Data

The representation of information about the world recorded through observations (qualitative data) or measurements (quantitative data). In the context of AI, data is digitally recorded. In the context of AI, qualitative and quantitative data is digitally recorded.

Data Point

A discrete unit of information (or singular item of data).

Dataset

A collection of data that can range in type (e.g. numbers, words, images) and be either specific to a purpose or general and varied.

Training Data

A subset of the dataset that is used to initially develop the model, by feeding the data into an algorithm.

Structured Data

A collection of data that is specific to a purpose and organised into tables with clearly defined categories (e.g., official government statistics, spreadsheets).

Unstructured Data

a collection of general and varied data with no formatting or defined categories (e.g. collections of emails, text messages, CCTV footage, social media use).

Annotated Data

Data that has tags or labels added to it, which provide context (or metadata).

Features

Features are characteristics within data that are organised into categories. Features are also referred to as variables and are the inputs that AI models use to generate outputs (e.g., organised survey results).

Data Protection Terminology

Personal Data

Data that can be used to identify a living individual. Some personal data need more protection because they are particularly sensitive or revealing (e.g. data about a person's medical condition or finances). This is referred to as "special category data".

Data Subject

A person whose personal data is collected, stored, processed, or used and who is identified or can be identified, directly or indirectly, by information such as a name or identity number, or by a combination of characteristics specific to that individual.

Datafication

A process associated with the increasing use of technology, by which individuals, objects, and actions generate digital records (data). The interactions of individuals with digital public-facing services generate data, such as demographic information, usage patterns, and user behaviour. Organisations that employ digital technologies to transform public facing services may collect, analyse, and use these data to inform future service delivery.

Data Controller

A natural or legal person, public authority, agency or other body that, alone or jointly with others, exercise overall control over the purposes and means of the processing of personal data. This includes decisions over what data to process and why. In an AI project, data controllers play a crucial role in ensuring that an AI system is built and operated in a responsible and ethical manner. They must ensure data is collected, stored, processed, and used in accordance with relevant data protection regulations. Data processors are accountable for their own compliance and that of the processors.

Data Processor

A natural or legal person, public authority, agency or other body that does not have any purpose of their own for processing the data and only act on the data controller's instructions.

Data Protection Impact Assessment (DPIA)

A Data Protection Impact Assessment (DPIA) describes a process designed to identify risks arising out of the processing of personal data and to minimise these risks as far and as early as possible.⁵¹

Degrees of Automation

An AI system is fully automated if when used, its output and any action taken as a result are implemented without any human involvement or oversight. In lower degrees of automation, a human can oversee the AI system to ensure it is producing the intended outcomes and/or use the outputs as part of a wider process in which they consider the output of the AI model, as well as other information available to them, and then acts based on this. This is often referred to as having a 'human-in-the-loop'. Degrees of automation can be seen as a spectrum rather than a binary concept and can vary depending on the specific context.

Full automation

A ML model used to prioritise patients for emergency care within an overcrowded hospital categorises a patient as ineligible for emergency care, automatically barring them access.

Human-on-the-loop

A ML model categorises a patient as ineligible for emergency care. A nurse, who is aware of the patient's medical history and the patient's account of their current symptoms, considers the model's output and their professional judgement. They decide the model's output is erroneous and override the model's decision.

Human-in-the-loop

A ML model categorises a patient as ineligible for emergency care. A nurse considers the model's output, the patient's medical history, the patient's account of their current symptoms, and their professional judgement to determine the patient's eligibility.

Human-over-the-loop

A ML model makes a preliminary determination about patients' eligibility for emergency care. Before any decision is made, the ML model output is automatically routed to a nurse, who reviews and approves or overrides the model's output before any action is taken.

Machine Learning terminology

Machine Learning (ML)

A branch of artificial intelligence that allows computer systems to learn directly from examples, data, and experience. Through enabling computers to perform specific tasks intelligently, machine learning systems can carry out complex processes by learning from data, rather than following pre-programmed rules.

Supervised learning

Supervised learning is the most widely used type of machine learning. Supervised ML models differ from unsupervised models because they are trained using labelled data. This means that the dataset contains 'labels' of the desired output or target variable that the model is trying to predict. Therefore, using this data, the model can find patterns between specific features in the dataset and the defined target variable (i.e., linking the inputs to the outputs). After the model is trained on this data, it can then be used to

predict future outcomes by applying the model to new, unseen data that represents real-world scenarios.

Unsupervised learning

While supervised learning models map relationships between features in datasets that contain labels, unsupervised ML models identify patterns within unlabelled data by determining similarities and differences amongst the unlabelled data points. There are various applications of unsupervised learning, but one of the most common is clustering.

Reinforcement learning

These models ‘learn’ on the basis of their interactions with a virtual or real environment rather than existing data. Reinforcement learning ‘agents’ search for an optimal way to complete a task by taking a series of steps that maximise the probability of achieving the given task. Depending on the success or failure of the steps they take, their actions are iteratively rewarded or penalised to maximise rewards. Reinforcement learning models improve with multiple iterations of trial and error to change their given ‘state’. A ‘state’ is the position of the agent at a specific point in time. It represents the current environment the agent is in and includes the collection of all relevant information that the agent needs to make decisions about what action to take next. Reinforcement learning models may be designed to develop long-term strategies to maximise their reward overall rather than looking only at their next step.

ML or AI?

The terms “machine learning” and “artificial intelligence” are often used interchangeably. Indeed, a significant amount of contemporary AI technology has ML at its core. However, rules-based AI systems that do not require training data may not be considered ML systems. In practice, complex AI systems incorporate combinations of techniques such that they commonly include at least some form of ML.

Neural Network

Neural networks are a type of ML algorithm that is used to model complex patterns in data. Neural networks are similar to other machine learning algorithms, but they are composed of a large number of interconnected processing nodes, or neurons, that can learn to recognize patterns of input data. The mathematical constructs behind the interaction of the various nodes of a neural network are modelled on an understanding of how biological neurons interact and filter information in animal brains.

Like many AI techniques, neural network techniques are not new. The first computer to employ the technique was developed in 1950. Neural networks were further developed in the 1960s but declined in use and attention in subsequent AI research. Yoshua Bengio, Geoffrey Hinton, and Yann LeCun (2019) are credited for making multilayer neural networks (aka “deep learning”) a critical part of modern computing.

A current wave in neural network development is a class of multi-layered networks called “transformers”. The transformer architecture is foundational to generative AI and many of the large language models we hear about today, including BERT and ChatGPT.

Appendix B: Process-Based Governance log template

Project Name _____

Directorate/team _____

Governance Action	Logged by	Artefact	Log Date	Last Revisitation Date
Project scoping	M Katell	PS Report	04/06/2023	
Stakeholder Engagement Process (SEP)	M Katell	PS Report updated	30/06/2023	31/10/2023
Stakeholder Impact Assessment (SIA)		SIA created	04/07/2023	
Safety Self-Assessment and Risk Management		Safety Self-Assessment Risk Management Plan		
Bias Self-Assessment and Risk Management		Updated Bias Self-Assessment and Risk Management Plan		
Data Factsheet		Updated Data Factsheet		
Roles and Responsibilities	L Singh	Updated roles and responsibilities list		
PBG Framework		Updated PBG Framework		
Explanation-Aware Design Self-Assessment and Risk Management		Updated Explanation-Aware Design Self-Assessment and Risk Management Template		

Appendix C: Project Summary Report template

PROJECT Questions	Response
What is the AI system or service being built or acquired and what type of product or service will it offer?	
Who are the users or customers of the system or service?	
What benefits will the system bring to its users and customers, and will these benefits be widely accessible?	
Which organisation(s)—yours, other suppliers, or other providers—are responsible for building this AI system? ⁵²	
Which parts or elements of the AI system, if any, will be procured from third-party vendors, suppliers, sub-contractors, or external developers?	
Which algorithms, techniques, and model types will be used in the AI system? (Provide links to technical papers where appropriate)	
In a scenario where your project optimally scales, how many people will it impact, for how long, and in what geographic range (local, national, global)? (Describe your rationale)	

USE CONTEXT Questions	Response
What is the purpose of this AI system and in which contexts will it be used? (Briefly describe a use-case that illustrates primary intended use)	
Is the AI system's processing output to be used in a fully automated way or will there be some degree of human control, oversight, or input before use? (Describe)	
Will the AI system evolve or learn continuously in its use context, or will it be static?	

To what degree will the use of the AI system be time-critical, or will users be able to evaluate outputs comfortably over time?	
What sort of out-of-scope uses could users attempt to apply the AI system, and what dangers may arise from this?	

DOMAIN Questions	Response
In what domain will this AI system operate? (e.g. international trade, goods manufacturing)	
Which, if any, domain experts have been or will be consulted in designing and developing the AI system?	

DATA Questions	Response
What datasets are being used to build this AI system?	
Will any data being used in the production of the AI system be acquired from a vendor or supplier? (Describe)	
Does the system or service require that user data be transmitted off-site to a third-party system? ⁵³	
Will the data being used in the production of the AI system be collected for that purpose, or will it be re-purposed from existing datasets? (Describe)	
Are there any anticipated data protection or intellectual property considerations about the data?	
If the system is procured from a third-party, what is known about the data used to train the system?	

STAKEHOLDERS Questions	Response
Who are the stakeholders (including individuals and social groups) that may be impacted by, or may impact, the project?	
Do any of these possess sensitive or protected characteristics that could increase their vulnerability to abuse or discrimination, or for reason of which they may require additional protection or assistance with respect to the impacts of the project? If so, what characteristics?	
Could the outcomes of this project present significant concerns to specific groups of affected users given vulnerabilities caused or precipitated by their distinct circumstances?	
If so, what vulnerability characteristics expose them to being jeopardized by project outcomes?	

ETHICAL PRINCIPLES	Response
What are the top-line ethical considerations for this system in regards to <i>Sustainability</i> ?	1.
	2.
	3.
What are the top-line ethical considerations for this system in regards to <i>Safety</i> ?	1.
	2.
	3.

What are the top-line ethical considerations for this system in regards to <i>Accountability</i> ?	1.
	2.
	3.
What are the top-line ethical considerations for this system in regards to <i>Fairness</i> ?	1.
	2.
	3.
What are the top-line ethical considerations for this system in regards to <i>Explainability</i> ?	1.
	2.
	3.
What are the top-line ethical considerations for this system in regards to <i>Data Stewardship</i> ?	1.
	2.
	3.

MAP GOVERNANCE WORKFLOW

What roles are involved in each of the project phases?	Design Roles
	Development Roles

	Deployment Roles
What are the responsibilities of each of these roles?	Design
	Development
	Deployment
How are each of these duty bearers assigned responsibility for the system's potential impacts? Does this distribution establish a continuous chain of human accountability throughout the design, development, and deployment of this project? If so, how?	
What logging protocol is established for documenting workflow activities? Does this protocol enable auditing and oversight of the design, development, and deployment of this project? If so, how?	
Can responsible duty bearers be traced in the event that stakeholders are harmed by this system? If so, how do the project's distribution of responsibilities and logging protocol enable this?	
If you are procuring parts or elements of the system from third-party vendors, suppliers, sub-contractors, or external developers, how are you instituting appropriate governance controls that will establish end-to-end accountability, traceability, and auditability?	

Appendix D: Data Factsheet

The **Data Factsheet** is a series of questions about the data that will interact with the AI system or service. We recommend adding facts to this sheet that are known about the data that add to the team’s understanding of the data in addition to what is revealed by the queries. It may be desirable to maintain separate factsheets for data used at different stages, such as training, testing, and production. It may also be useful to add factsheets where data is from multiple, diverse origins.

Category	Query	Response
Features	What is the dataset? What’s in it?	
	Is the data about people? ¹	
	How large is the dataset? (bytes, rows, other measures)	
	How and where is the data stored and accessed?	
	Are there any copyright or licensing concerns?	
Provenance	What is the origin of the data?	
	For what purpose was it originally collected?	
	Who collected it?	
	Were the methods of data collection exploitative of data workers or data subjects?	
	What was the timeframe of the data collection?	
	Who funded the collection?	
	What else is know of the dataset’s origin?	
Procurement	Who is providing the dataset?	

¹ If the data is about people who may be identifiable, ensure that data protection and other obligations are met.

	Are there any concerns about the ethics or legality of the provider's methods or business practices?	
	What are the restrictions on its use, if any?	
	What has the provider disclosed/not disclosed about it?	
	What recommendations or warnings have been provided about the data?	
Quality	Has the data been evaluated for accuracy? (describe)	
	Is there a sufficient amount of data for the task?	
	Is the data complete and integral?	
	Has the data been evaluated for representativeness of relevant groups or categories? (describe)	
	Has the data been evaluated for relevance to the intended task? (describe)	
	To what extent has the data been cleaned or otherwise prepared?	
	Is the labelling/annotation sufficient for the purpose?	
Use and Impact	What is the intended use for the data?	
	What uses should be avoided?	
	How long will the data be relevant, accurate for the purpose?	
	What types of bias are reflected in the data?	
	Is sufficient bias mitigation possible to avoid data harms?	
	How will the limitations and contraindications about the data be communicated?	

	How will any harms from the use of this data be avoided, minimised, or managed?	
Security and safety	Is the data held securely?	
	What are the risks (and to whom) if the data were to be breached or corrupted?	
	Who is responsible for ensuring data security?	
Other information	What else should be noted about this dataset?	

Appendix E: Procurement Guidance Tool

The purpose of this tool is to provide guidance for identifying ethical issues in the procurement of AI/ML systems produced from third parties. This section includes a decision matrix for reasoning through procurement decisions.

Introduction to this section

Technical systems, including many forms of AI and useful data sets may be more practically acquired from external suppliers than produced within government organisations. This includes contracting with software developers to support the creation of bespoke systems and also working with companies who license or otherwise provide fully-built systems or key components of larger systems to government. While such arrangements can be beneficial for meeting public service objectives, there are feasibility and ethical considerations for the acquisition of any technical system or component and some additional considerations for LLMs and GenAI systems.

The base AI models that undergird the most powerful GenAI systems, which are known as “foundation models”, are not currently feasible to be built by system developers within government ministries and departments due to the significant compute, data, and human resources they require to develop, train, and maintain. While individual GenAI technologies may soon be built by government personnel, they are likely to continue to be provided by third-parties or using component parts provided by third-parties. This raises procurement risks due to a lack of transparency by the companies who produce the models, such as opacity about the sources of their training data and other facets of their design and functioning. Most leading GenAI companies are based outside the UK, raising risks of cross-border data flows that may be required to use these systems. Data protection, national security, and general confidentiality concerns follow from having to send data to remote systems in order to take advantage of their functions and outputs.

As the UK and other trusted governments work to develop sovereign capacity the development of advanced technologies, including GenAI, many safety and confidentiality concerns may be resolved, but at the time of writing of this guidance, such capacity is not yet available. It is also true that less intensive AI systems may be more feasibly and economically produced in-part or entirely by external developers. In any case, when government personnel are considering the procurement of AI systems and services, there are a number of important considerations, some of which are unique to the context of adopting technology for government.

“Black-box” systems

Wherever possible, government ministries and departments should use their significant buying power to compel third-party providers to disclose as much as possible about their systems and to otherwise comply with agreed-upon ethical principles. Where this is not possible or only possible in a minimal way, such systems should be considered to be “black-box” systems. It may be ethically permissible to acquire and use black-box systems, but project teams should:

(1) Thoroughly weigh up impacts and risks: Your first step in evaluating the feasibility of using a complex AI system should be to focus on issues of ethics and safety. As a general policy, you and your team should utilise ‘black box’ models only:

- where their potential impacts and risks have been thoroughly considered in advance, and you and your team have determined that your use case and domain specific needs support the responsible design and implementations of these systems;
- where supplemental interpretability tools provide your system with a domain appropriate level of semantic explainability that is reasonably sufficient to mitigate its potential risks and that is therefore consistent with the design and implementation of safe, fair, and ethical AI.

(2) Consider the options available for supplemental interpretability tools: Next, you and your team should assess whether there are technical methods of explanation-support that both satisfy the specific interpretability needs of your use case and are appropriate for the algorithmic approach you intend to use. You should consult closely with your technical team at this stage of model selection. The exploratory processes of trial-and-error, which often guide this discovery phase in the innovation lifecycle, should be informed and constrained by a solid working knowledge of the technical art of the possible in the domain of available and useable interpretability approaches.

The task of lining up the model selection process with the demands of interpretable AI requires a few conceptual tools that will enable thoughtful evaluation of whether proposed supplemental interpretability approaches sufficiently meet your project’s explanatory needs. First and most importantly, you should be prepared to ask the right questions when evaluating any given interpretability approach. This involves establishing with as much clarity as possible how the explanatory results of that approach can contribute to the user’s ability to offer solid, coherent, and reasonable accounts of the rationale behind any given algorithmically generated output. Relevant questions to ask that can serve this end are:

- What sort of explanatory resources will the interpretability tool provide users and implementers in order (1) to enable them to exercise better-informed evidence-based judgments and (2) to assist them in offering plausible, sound, and reasonable accounts of the logic behind algorithmically generated output to affected individuals and concerned parties?
- Will the explanatory resources that the interpretability tool offers be useful for providing affected stakeholders with a sufficient understanding of a given outcome?
- How, if at all, might the explanatory resources offered by the tool be misleading or confusing?

You and your team should take these questions as a starting point for evaluating prospective interpretability tools. These tools should be assessed in terms of their capacities to render the reasoning behind the decisions and behaviours of the

uninterpretable 'black box' systems sufficiently intelligible to users and affected stakeholders given use case and domain specific interpretability needs.

(3) Formulate an interpretability action plan: The final step you will need to take to ensure a responsible approach to using 'black box' models is to formulate an interpretability action plan so that you and your team can put adequate forethought into how explanations of the outcomes of your system's decisions, behaviours, or problem-solving tasks can be optimally provided to users, decision subjects, and other affected parties.

This action plan should include the following:

- A clear articulation of the explanatory strategies your team intends to use and a detailed plan that indicates the stages in the project workflow when the design and development of these strategies will need to take place.
- A succinct formulation of your explanation delivery strategy, which addresses the special provisions for clear, simple, and user-centred explication that are called for when supplemental interpretability tools for 'black box' models are utilised.
- A detailed timeframe for evaluating your team's progress in executing its interpretability action plan and a role responsibility list, which maps in detail the various task-specific responsibilities that will need to be fulfilled to execute the plan.

Specific guidance

The Office of AI published [a guide](#) for AI procurement that sets out key considerations for government participation in a robust and responsible market for AI. While this guide provides valuable advice, it was published in 2020 and does not emphasise key risks involved in procuring GenAI.

Inspired by the OAI's guidance and other government guidance we provide these procurement principles (in brief):

- **Be strategic:** Consider how an AI system or service fits into an overall strategy of technology adoption for the DBT. Coordinate with and learn from other government organisations.
- **Be multidisciplinary and diverse:** Decision-making about AI should involve many perspectives, many disciplines (including ethics!).
- **Know your data landscape:** A complete data assessment should be conducted that includes a thorough understanding of data on hand (or to be acquired). Prepare a "data position" that sets out requirements before engaging with vendors. Ensure that data protection, confidentiality, and national security obligations can be met.

- Assess and analyse risks and impacts: Employ the COBRA process provided in this guide to assess the potential risks of the system and forecast its impacts.
- Promote UK leadership in responsible AI: Use the power of government spending to shape the market for AI with public benefit and respect for fundamental rights foremost.
- AI should never be the first step: avoid “solution looking for a problem” scenarios.
- Engage with vendors with a clear problem statement and/or requirements documents.
- Perform a complete data assessment that includes a
- Use government’s significant purchasing power to ensure that procured systems and services meet public standards.
- Accountability: Insist on maximum transparency from vendors.
- Risk-benefit analysis: Always consider alternatives to procurement.

Principles to guide the use of third-party systems

- Designate a responsible person who will supervise the procured system, service, or component once it is in production. Communicate this responsibility to system users and other relevant parties.
- Employ filtering on input. Monitor for confidential or sensitive information and add warnings to user interface. Require users to assert their understanding of ethics and regulatory requirements (per CDDO guidelines).
- Log all use. Regularly and repeatedly audit. Adjust input filtering based on audit results.
- Create cautionary use cases for training.
- Automated decisions: What are the stakes? What is the worst outcome? Work backwards from there.

There is a set of questions for each principle below. The answer to each question should be answered with a yes or no and the evidence to support the answer should be documented. A “no” response to any question should serve as a warning to decision-makers, triggering closer examination and mitigation strategies to prevent undesirable outcomes. The decision to accept a procured system that does not meet these principles should be clearly assigned to a senior manager.

Data-driven systems are frequently embedded into other systems and practices. Their inclusion may alter and extend the capabilities of existing systems. They may also present new risks and concerns that were not considered previously. For each principle and question below, it is important to consider the target system both from the standpoint of its stand-alone features and its potential effects on the status other systems in which it is embedded or otherwise affects.

Principle	Key Questions	Response
Risk-benefit	<p>Does the choice to purchase a system or system component provide a non-generic public benefit?</p> <p>Have the potential risks and harms of the system been anticipated, explored, and mitigated where needed?²</p>	
Interpretability	<p>Do you have an interpretability action plan to make explicit how the system's outputs will be explained and justified to affected users and other stakeholders?³</p>	
Data sufficiency	<p>Is sufficient high-quality data to make the procurement useful currently available or readily accessed?⁴</p>	
Problem definition	<p>Is there a clearly defined problem the purchase is meant to solve?</p> <p>Is the problem definition specified in a manner suppliers can understand and respond to?</p> <p>Does the problem definition originate with the department (rather than the supplier)?</p>	
Openness	<p>Will the procured system be open, transparent, and auditable?</p> <p>Have the limits on openness been fully considered?</p>	

² The COBRA tool may be useful for answering this question.

³ We recommend the ICO/Turing publication "Explaining decisions made with AI" to assist with this step: <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/>

⁴ The Bias Self-Assessment tool may be useful for answering this question.

	Does the supplier assume a sufficient degree of risk and liability for systems that are not open, transparent, and auditable?	
Security	Does the purchase present security risks to the department and other system users? Is there a clearly defined mitigation strategy for security risks, if any? Does the supplier accept responsibility for security? Is the supplier's security assurance auditable?	
Privacy & Confidentiality	Will the purchased system sufficiently protect users from privacy risks? Can system users fully exercise their data protection rights under UK law? Does the system provide safeguards for confidential and otherwise sensitive information? ⁵	

⁵ See 'Guidance to civil servants on use of generative AI' (29 June 2023)

References

- [1] H. Toner, "What Are Generative AI, Large Language Models, and Foundation Models?" *Center for Security and Emerging Technology*. May 2023. Accessed: Aug. 05, 2023. [Online]. Available: <https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/>
- [2] E. ACM Technology Council, "Principles for the Development, Deployment, and use of Generative AI Technologies," ACM, Jun. 2023.
- [3] M. Goyal, S. Varshney, and E. Rozsa, "What is generative AI, what are foundation models, and why do they matter?" *IBM Blog*. Mar. 2023.
- [4] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean, "Large Language Models in Machine Translation."
- [5] Royal Society, "Machine learning: The power and promise of computers that learn by example," The Royal Society, 2017.

Scrutiny	<p>Will the use of the procured system withstand the scrutiny of internal and external sceptics, including the public and regulators?</p> <p>Are procurers prepared to defend the choice to purchase and use this system?</p>	
Cost-benefit	<p>In addition to base procurement costs, have any additional costs to make the system comply with these principles been incorporated into the overall cost-benefit analysis?</p>	

[6] A. Wang, S. Kapoor, S. Barocas, and A. Narayanan, “Against predictive optimisation: On the legitimacy of decision-making algorithms that optimise predictive accuracy”, Oct. 2022. <https://ssrn.com/abstract=4238015>

[7] P. Hacker, A. Engel, and M. Mauer “Regulating ChatGPT and other Large Generative AI Models”, In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT ‘23). Association for Computing Machinery, New York, USA, 2023. <https://doi.org/10.1145/3593013.3594067>

[8] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, C. Biles, S. Brown, Z. Kenton, W. Hawkins, T. Stepleton, A. Birhane, L. A. Hendricks, L. Rimell, W. Isaac, J. Haas, S. Legassick, G. Irving, I. Gabriel, “Taxonomy of Risks Posed by Language Models”, 2022. <https://dl.acm.org/doi/10.1145/3531146.3533088>

Appendix F: Context-Based Risk Assessment (COBRA) worksheet

Document to be carried out as part of the Context-based Risk Assessment Tool (COBRA)

Risk factors pertaining to AI application contexts (for each risk, what is the severity of risk related to its scale, scope, and likelihood?)

Risks pertaining to the project design context (for each risk, what is the severity of risk related to its scale, scope, and likelihood?)

Risks pertaining to the model development context (for each risk, what is the severity of risk related to its scale, scope, and likelihood?)

Risks pertaining to the model deployment context (for each risk, what is the severity of risk related to its scale, scope, and likelihood?)

Appendix G: Stakeholder Impact Assessment

Project Name	
Date Completed	
Team members involved	
External stakeholders consulted	

Section 1A: Design Phase (Project Planning)

Horizon-Scanning and the Decision to Design

Have you assessed whether building an AI model or tool is the right solution to help you deliver the desired services given:

- a) the existing technologies and processes already in place to solve the problem;
- b) current user needs;
- c) the current state of available data;
- d) the resources (material and human) available to your project;
- e) the nature of the policy problem you are trying to solve; and
- f) whether an AI-based solution is appropriate for the complexity of its potential use contexts?

Do these initial assessments support the justifiability and reasonableness of choosing to build an AI model or tool to help you deliver the desired services?

For more details on “Assessing if artificial intelligence is the right solution” see guidance by the Office for AI and Central Digital and Data Office. For further details about understanding user needs, see Section 1 of the [Data Ethics Framework](#) and the user research section of the [Gov.UK Service Manual](#).

Has a thorough assessment of the human rights compliant business practices of all businesses, parties, and entities involved in the value chain of the AI product or service been undertaken? This would include all businesses, parties, and entities directly linked to your business lifecycle through supply chains, operations, contracting, sales, consulting, and partnering. If not, do you have plans to do this?

Goal-Setting and Objective-Mapping

- a) How are you defining the outcome (the target variable) that the system is optimising for? Is this a fair, reasonable, and widely acceptable definition?

- b) Does the target variable (or its measurable proxy) reflect a reasonable and justifiable translation of the project's objective into the statistical frame?
- c) Is this translation justifiable given the general purpose of the project and the potential impacts that the outcomes of its implementation will have on the communities involved?
- d) Where appropriate, have you engaged relevant stakeholders to gather input on their views about reasonableness and justifiability of the outcome definition and target variable determination?

Possible Impacts on the Individual

- a) How, if at all, might the use of your AI system impact the abilities of affected stakeholders to make free, independent, and well-informed decisions about their lives? How might it enhance or diminish their autonomy?
- b) How, if at all, might the use of your system affect their capacities to flourish and to fully develop themselves?
- c) How, if at all, might the use of your system do harm to their physical, mental, or moral integrity? Have risks to individual health and safety been adequately considered and addressed?
- d) How, if at all, might the use of your system impact freedoms of thought, conscience, and religion or freedoms of expression and opinion?
- e) How, if at all, might the use of your system infringe on their privacy rights, both on the data processing end of designing the system and on the implementation end of deploying it? This question should supplement the completion of a Data Protection Impact Assessment.

Possible Impacts on Society and Interpersonal Relationships

- a) How, if at all, might the use of your system adversely affect each stakeholder's fair and equal treatment under the law? Are there any aspects of the project that expose vulnerable communities to possible discriminatory harm? These questions should supplement the completion of an Equality Impact Assessment.
- b) Does the project aim to advance the interests and wellbeing of as many affected individuals as possible? Might any disparate socioeconomic impacts result from its deployment?
- c) How, if at all, might the use of your system affect the integrity of interpersonal dialogue, meaningful human connection, and social cohesion?
- d) How, if at all, might the use of your system affect freedom of assembly and association?
- e) How, if at all, might the use of your system affect the right to diverse and reliable information and access to plurality of ideas and perspectives?
- f) How, if at all, might the use of your system affect the right of individuals and communities to participate in the conduct of public affairs?
- g) How, if at all, might the use of your system affect the right to an effective remedy for violation of rights and freedoms, the right to a fair trial and due process, the right to judicial independence and impartiality, and equality of arms?

- h) Have the values of civic participation, inclusion, and diversity been adequately considered in articulating the purpose and setting the goals of the project? If not, how might these values be incorporated into your project design?
- i) Have you sufficiently considered the wider impacts of the system on future generations and on the planet and biosphere as a whole?
- j) How could the use of the AI system you are planning to build or acquire—or the policies, decisions, and processes behind its design, development, and deployment—lead to the discriminatory harassment of impacted individuals?
- k) How could the use of the AI system you are planning to build or acquire—or the policies, decisions, and processes behind its design, development, and deployment—lead to the disproportionate adverse treatment of impacted individuals from protected groups on the basis of their protected characteristics?
- l) How could the use of the AI system you are planning to build or acquire—or the policies, decisions, and processes behind its design, development, and deployment—lead to the discriminatory harassment of impacted individuals?

Section 1B: Design Phase (Problem Formulation)

Sector-Specific and Use Case-Specific Questions

Project Name	
Date Completed	
Team members involved	
External stakeholders consulted	

In this section you should consider the sector-specific and use case-specific issues surrounding the social and ethical impacts of your AI project on affected stakeholders. Compile a list of the questions and concerns you anticipate. State how your team is attempting to address these questions and concerns. Where appropriate, engage with relevant stakeholders to gather input about their sector-specific and use case-specific concerns.

Section 1C: Design Phase

Revisiting Project Summary Report

Project Name	
Date Completed	

Team members involved	
External stakeholders consulted	

Revisiting Stakeholder Analysis and Positionality

- a. Do the stakeholder groups outlined in the report accurately reflect current stakeholders of this project? Are there other stakeholder groups that should be considered as stakeholders for this project?
- b. Do the potential impacts outlined in the report accurately reflect current SIA results?
- c. Do the stakeholder groups currently identified as salient represent those groups that are currently likely to be most differentially impacted, vulnerable, or marginalised?
- d. Does the team positionality reflection accurately represent the relationship between team members and stakeholders at this stage in the project?

Revisiting Engagement Objectives and Methods

- a. Considering the results of the SIA, are there any new potential project impacts that may lead you to reconsider your engagement objectives and methods? If so, how?
- b. Do your chosen engagement objectives and methods seem proportional to the current identified impacts?
- c. Do any adjustments need to be made to your chosen engagement objectives and methods given the SIA results? If so, are there any additional practical considerations that need to be addressed to ensure that your engagement objectives and methods are realised?

Revisiting the Process-Based Governance (PBG) Framework

- a. Considering SIA results, does the PBG Framework for this project still accurately reflect the human chain of responsibility and create the baseline conditions for the project team to be actively accountable for system impacts?

Section 2: Development Phase

Model Reporting

Project Name	
---------------------	--

Date Completed	
Team members involved	
External stakeholders consulted	

After reviewing the results of your initial SIA, answer the following questions:

- a. Are the trained model’s actual objective, design, and testing results still in line with the evaluations and conclusions contained in your original assessment? If not, how does your assessment now differ?
- b. Have any other areas of concern arisen with regard to possibly harmful social or ethical impacts as you have moved from the Design to the Development Phase?

Re-assess questions in the Project Summary Report

You should set a reasonable timeframe for Public Consultation and Development Phase re-assessment.

Dates of public consultation on development phase impact re-visitation	
Date of planned development phase re-assessment	

Section 3: Deployment Phase

System Use and Monitoring

Project Name	
Date Completed	
Team members involved	
External stakeholders consulted	

Once you have reviewed the most recent version of your SIA and the results of the public consultation, answer the following questions:

- a. How does the content of the existing SIA compare with the real-world impacts of the AI system as measured by available evidence of performance, monitoring data, and input from implementers and the public?
- b. What steps can be taken to rectify any problems or issues that have emerged?
- c. Have any unintended harmful consequences ensued in the wake of the deployment of the system? If so, how might these negative impacts be mitigated and redressed?
- d. Have the maintenance processes for your AI model adequately taken into account the possibility of distributional shifts in the underlying population? Has the model been properly retuned and retrained to accommodate changes in the environment?

Appendix H: Readiness Self-Assessment



This tool has been designed to help individuals and teams understand the degree of preparedness to adopt AI/ML systems or practices amongst key stakeholders.

Readiness

Readiness refers to the degree of preparedness to accept AI/ML systems or practices amongst key stakeholders.

! Readiness is an exploration of organisational adaptability and change

Consider that every AI/ML project is an opportunity for institutional change; each project potentially opens up possibilities for mutually beneficial dialogues and learning about the risks and opportunities of new technologies.

Readiness Alignments

The purpose of conducting a readiness self-assessment is to understand and address deficits in the establishment and maintenance of trust, awareness, and capabilities within an organisation.

Readiness gaps are often caused by ‘misalignments’ between decision-making actors on the one hand (e.g. project teams and leadership) and relevant stakeholders (e.g. product users and affected persons and communities) on the other.

Three categories of misalignment are described here and discussed in the 'Mitigation Techniques' section below:

- **Values:** Alignment between the AI/ML project and the values, beliefs, purposes, and missions of the system producers, users, and individuals affected by its implementation.
- **Needs:** Alignment between the AI/ML project and the administrative and practice needs of users and the service needs of individuals affected by its implementation.
- **Knowledge:** Alignment between the AI/ML project's goals and requirements and a) users' cognitive needs, adaptability, skills levels, and capabilities; b) the organisations' commitments to training and development to upskill everyone and fill gaps; and c) the cognitive participation, sense-making, and informed acceptance of users and individuals affected by a project's implementation.



Readiness Taxonomy

This taxonomy of readiness types is organised as a series of prompts that question the achievement of the goals of readiness. The types are grouped into categories for *Team*, *Stakeholders*, and *Leadership* to indicate what organisational role is best positioned to identify, experience, and/or mitigate an obstacles to readiness. Readiness challenges may fall across role categories, so we recommend that all readiness prompts be reviewed by users of this tool.

i Documentation matters

Whether the response to a readiness claim is positive or negative, it is a good practice to document the answer as well as any evidence that demonstrates how the claim has been or could be met.

Team

Readiness questions for assessment within teams

Main Category	Interpretation	Deliberative Prompt
Feasibility (practical)	Reflecting on the feasibility of proposed AI/ML projects.	Has our team/organisation determined that the proposed AI/ML project is feasible to design, develop, and implement?
Feasibility (suitability)	Reflecting on the need and value of the proposed AI/ML project.	Has our team thoughtfully considered whether a proposed AI/ML project is the right solution to the stated or perceived problem or challenge?
Feasibility (agency)	Reflecting on the power to decide.	Does our team have a voice in deciding if AI/ML projects are both feasible and suitable to address a stated or perceived problem or challenge?
Absorptive Capacity (knowledge and skill base)	Existing or readily accessible internal knowledge base.	Can our team draw upon existing knowledge and skills from within our organisation that are necessary to responsibly design, develop, and deploy the AI/ML project? Can our team/organisation gain and assimilate new knowledge and skills we do not already have available but are necessary to responsibly design, develop, and deploy the AI/ML project?
Absorptive Capacity (knowledge sharing)	Sharing on knowledge internally.	Does our team have accessible and established mechanisms for sharing and disseminating the necessary skills and knowledge to

Main Category	Interpretation	Deliberative Prompt
Knowledge of the State of the Art	Technological and regulatory sophistication.	<p>others in our organisation who are affected by the AI/ML project?</p> <p>Does our team have access to expertise and knowledge relating to the state of the art in AI that is sufficient to equip us with the understandings needed to:</p> <ul style="list-style-type: none"> * Adequately and appropriately scrutinise uses of AI and claims made by vendors and decision makers? * Situate the AI/ML project within broader understandings of the technology, its capabilities, and its limitations?
Capacity for Gap Understanding and Risk Mapping	Risks of AI, gaps in legal/policy compliance, impact assessments.	<p>Does our team have internal processes in place that enable us to:</p> <ul style="list-style-type: none"> * Map and understand the risks posed by the AI/ML project? * Develop risk mitigation procedures? * Identify and meet legal obligations (e.g. data protection)? * Identify and meet policy obligations (e.g. court procedures)? * Conduct impact assessments?
Receptivity to Change (internal)	Attitude towards new technologies and technological change	Do the norms of our team encourage and nurture the acceptance of new AI/ML technologies and practices?
Receptivity to Change (organisational)	Institution encourages ethical and responsible practices.	Does our team promote and nurture the acceptance of new AI/ML technologies and practices in our organisation?

Main Category	Interpretation	Deliberative Prompt
Legal and Policy Compliance	Consulting with legal teams and policy managers to sense-check and validate legal and organisational obligations implicated by the AI/ML project.	Has our team undertaken meaningful consultation with internal and external collaboration with regulatory bodies and/or regulatory experts to ensure the AI/ML project complies with existing regulations?

Stakeholders

Readiness questions for working with stakeholders

Main Category	Interpretation	Deliberative Prompt
Stakeholder Engagement (internal)	Consulting with appropriate stakeholders to anticipate risk, harm, and opportunity.	Has our team proactively pursued and cultivated meaningful stakeholder engagement with system users and affected teams to cooperatively shape and collectively monitor the quality, effectiveness, and permissibility of AI/ML projects?
Stakeholder Engagement (external)	Working with external organisations/actors to sense-check and validate new and revised AI/ML systems.	Had our team proactively pursued and cultivated meaningful stakeholder engagement with affected persons or their representatives and/or has partnered with external domain experts to cooperatively shape and collectively monitor the quality, effectiveness, and permissibility of AI/ML projects?
Training and Skills Development (technical understanding)	Training and technical skill development processes for affected team members.	Do we have training and skills development processes in place to sufficiently prepare team members across affected departments to understand the technical dimensions of the AI project and the role it will play in the work of the Ministry?
Training and Skills Development	Training and interdisciplinary skill development processes	Do we have the training and skills development processes in place to

Main Category	Interpretation	Deliberative Prompt
(interdisciplinary understanding)	for affected team members ethics, policy, governance, market, etc.	encompass non-technical dimensions of AI such as: * Ethical dimensions? * Policy dimensions? * Governance dimensions? * Commercial dimensions?
Feedback and Perpetual Learning	Establishing clear and open lines of communication between system producers, users, and affected persons (as appropriate).	Does our team or organisation provide pathways for providing feedback about existing and proposed AI/ML systems that do not perform to an acceptable standard of fairness, quality, and accuracy? Does our team/organisation provide pathways for recourse to seek remedies when systems fall short of expectations or cause harm?

Leadership

Readiness questions for leadership

Main Category	Interpretation	Deliberative Prompt
Organisation-level Leadership	Leadership cultivates environment for change.	Do members of the leadership cultivate a cultural environment that is amenable to responsible AI/ML technologies and practices? Do members of the leadership take ownership over end-to-end best practices and responsibly implementing AI/ML technologies and practices? Do members of the leadership act as role models for bearing responsibility for the consequences of AI/ML technologies and practices?

Main Category	Interpretation	Deliberative Prompt
Resource Availability	Ability to make resources available for development, implementation, and sustainability demands.	Does our organisation make sufficient resources available for all stakeholders to engage meaningfully in the cooperative development and collective monitoring of the new AI/ML technologies we produce and deploy?
Change Readiness and Adaptability	Agency and confidence to implement change.	<p>Do members of our team/organisation thoughtfully implement changes related to new technologies or technology policy? This may include:</p> <ul style="list-style-type: none"> * Effectively conveying the importance and benefits of the AI/ML project to affected stakeholders * Posing critical questions that promote beneficial change * Anticipating, communicating, and mitigating potential consequences
Receptivity to Change (organisational)	Leadership encourages ethical and responsible change.	Does our leadership promote and nurture the acceptance of new AI/ML technologies and practices throughout the organisation?
Participant Attitudes	Cultivation of optimistic attitudes.	Does our organisation cultivate agency and optimism regarding AI/ML technologies and practices? Agency to have a say about technological change, and optimism that through inclusive collaboration, technological change can be beneficial for the organisation and affected individuals and communities?

Readiness Self-Assessment

The Readiness Self-Assessment is intended as ongoing process of deliberation and strategy that focuses attention on both the immediate concerns and challenges of a specific project and a broader analysis of an organisation's culture and technological evolution. The readiness self-assessment can be used at any point in the project lifecycle, but it is especially useful during the early stages when planning may be most flexible.

Lifecycle stages where investigating readiness is useful:

- Design: Project teams make decisions about feasibility and consider how the project might be shaped to accommodate downstream concerns and engender trust.
- Development: Making decisions about features and data use that demonstrate thoughtfulness about downstream perceptions.
- Deployment: Contributes ideas for training and other messaging that demonstrate awareness of the project’s effects on others and opportunities to promote acceptance and trust. For each claim, ask if the goal has, or can be met. If the answer is negative, strategise with your team about if it possible to overcome the barriers to readiness. In some cases, the answer will be “no”, either because achieving a type of readiness is out of the control of the person doing the assessment, or because it is not relevant to the current project.

Mitigating Barriers to Readiness

Once the project team has identified barriers to readiness, there are mitigation strategies that may help to overcome them. Each readiness type falls under an *obstacle type*, which is paired with a mitigation strategy suggestion.

- Lack of in-house technical skills and expertise
- Data challenges: quantity, quality, interoperability of data
- Low understanding of AI
- Resistance to change and/or risk aversion
- Organisation not sufficiently up to speed with new technologies

Steps

1. Teams review each readiness category to determine if a) an obstacle to readiness is present, b) its level of seriousness, and c) what mitigation strategies are available to address it.
2. Teams evaluate whether they have the knowledge, authority, and resources available to address a readiness challenge. Where there is a lack, teams report this to decision-makers.
3. Documentation of these activities are documented and made available to leadership, users, affected persons, as appropriate. Documentation is also retained in a clearly identified repository for the purpose for future investigations.

Mitigation Techniques

(Please refer back to “Readiness Alignments” above)

Technique	Description	Alignments (V)(N)(K)
Training (team)	Most readiness questions for teams are potentially addressable through targeted training, from building the team’s technical knowledge to	(V) Training curricula can convey openness to new technologies and methods. (N) Potential to demonstrate the usefulness of technologies and methods that are unfamiliar.

Technique	Description	Alignments (V)(N)(K)
	communicating effectively with stakeholders.	(K) Upskilling teams can make them more effective at building acceptable systems.
Training (users/stakeholders)	Beyond merely demonstrating how to use a new system, training about AI/ML can demystify them and prepare users and other stakeholders to be “critical friends” rather than merely critical.	(V) Training curricula to explore and address deeply held beliefs and concerns about AI/ML. (N) Potential to demonstrate the usefulness of technologies and methods that are unfamiliar. (K) Understanding AI/ML improves the user experience and the salience of their feedback about AI/ML systems.
Training (leadership)	Leadership can be a better proponent of trust and organisational culture regarding AI/ML when they themselves are well-informed.	(V) Leaders use training opportunities to resolve conflicts amongst contrasting perspectives about data-driven technologies. (N) Leaders learn to build a credible case to support adoption of novel projects and systems. (K) Gaining a deeper understanding of data-driven technologies improves leadership ability to advocate for technological change and organisational adaptation.
Resources	Identifying and securing needed resources (personnel, development tools, data) may be essential to ensuring that project teams are ready to responsibly produce and sustain an AI/ML system.	(V) Resource allocation is an expression of values. If necessary resources cannot be secured for a given project, this may raise questions about its value to the organisation. (N) Perceived need of a system may not correspond with the availability of necessary resources. If a resource gap cannot be filled, the project scope may have to be revised. (K) Additional expertise may be able to identify how to conduct a given project given existing resource constraints.

Technique	Description	Alignments (V)(N)(K)
Stakeholder engagement	Engaging with stakeholders, both within and outside of the organisation, can contribute to efforts to design, develop, and implement AI/ML that does what it claims and has the social license to do so. Stakeholder engagement within organisations can inform leaders and other decision-makers about their organisational culture.	<p>(V) Provides opportunities to understand beliefs and concerns about AI/ML and to find common ground.</p> <p>(N) Stakeholders are more likely to accept new technologies if they address the needs that matter to them.</p> <p>(K) Project teams can expand their perspectives with the <i>situated knowledge</i> (see note below) of users and other stakeholders .</p>
Risk analysis	Engaging in sincere and open risk analysis processes in which risks to all affected persons and communities are investigated and taken seriously sets the conditions for trust and acceptance of systems built with appropriate safeguards	<p>(V) Risk of harm is a serious concern. Identifying and attending to it demonstrates care and sensitivity.</p> <p>(N) The perceived necessity of system must be weighed against its costs, including the risk of harm.</p> <p>(K) Risk analysis provides opportunities for learning about limitations and alternate options, particularly for project teams.</p>
Open communication	Multi-way communication (e.g. open documentation, workshops, newsletters) between project teams, users, and other stakeholders demonstrates respect for those affected by AI/ML and opens opportunities for useful feedback for the team and recourse for affected persons when systems are perceived negatively.	<p>(V) Demonstrating a willingness to share information and to listen conveys respect. Respect promotes trust and acceptance.</p> <p>(N) When communication channels are open, more likely for teams to get useful feedback that improves the targeting of AI/ML systems.</p> <p>(K) The requirements of effective communication can prompt communicators to learn more about the technologies being discussed, ultimately improving understanding and shared language.</p>

Technique	Description	Alignments (V)(N)(K)
Strategic planning	<p>Leadership can use strategic planning activities to gain understanding about and promote appropriate AI/ML design, development, and use. Other organisational participants can advocate for prioritisation of training, communication, increased visibility and voice of domain and situated experts.</p>	<p>(V) Strategic planning is often about a recalibration of organisational values. Opportunities for focusing on the promotion of trust across teams and with affected persons.</p> <p>(N) Strategic planning is often about resourcing and clarifying present needs.</p> <p>(K) Strategic plans can be more successful when skill-building and knowledge development are emphasised.</p>

Appendix I: SSAFE-D Principles Core Attributes

Each of the SSAFE-D Principles can be broken down further into a set of attributes that reflect some common ways of understanding what the principle means in context. Being able to do this work of digging deeper into each of the principles helps to ensure that the ethical issues raised in the work of designing, developing, and deploying data-driven technologies can be fully expressed and accounted for. These *core attributes* help specify the principle and also narrow the scope of the principle to make it more actionable (i.e. to help operationalise the principle). In short, they help to decompose the high-level principle into more specific and actionable elements.

This framework includes a set of attributes for each principle that can be used as a starting point. We call these attributes the ‘core attributes’ of each principle. That is, they represent a core set of attributes that help with operationalisation for a broad range of projects. This is because, like the SSAFE-D Principles, they are built around a set of common concerns that are relevant to the design, development, and deployment of data-driven technologies. The tables below presents these attributes, along with a corresponding description, for each principle.

NOTE: This list is not meant to be exhaustive. It is up to project teams and decision-makers to consider the ethical principles that are applicable in a given project, and to operationalise with an understanding of the project and its stakeholders.

Table 1: *Sustainability Principle and Core Attributes*

Core Attribute	Description
Safety	Safety goes beyond the mere operational safety of the system. It also includes an understanding of the context of long-term use and impact of the system, and the resources needed to ensure the system continues to operate safely over time within its environment (i.e. is sustainable). For instance, safety may depend upon sufficient change monitoring processes that establish whether there has been any substantive drift in the underlying data distributions or social operating environment. Because aspects of safety may not be immediately apparent to system developers, engaging and involving users and stakeholders in the design and assessment of AI systems can help mitigate potential impacts to their human rights and fundamental freedoms.
Security	Security encompasses the protection of several operational dimensions of an AI system when confronted with possible adversarial attack. A secure system is capable of maintaining the integrity of its constitutive information. 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.

Core Attribute	Description
Robustness	The objective of robustness can be thought of as the goal that an AI system functions reliably and accurately under harsh or uncertain 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 functional integrity and the soundness of its operation in response to difficult conditions, adversarial attacks, perturbations, data poisoning, or undesirable reinforcement learning behaviour. Documentation is essential to ensure that the data, models, and systems are robust in case of changing personnel.
Reliability	The objective of reliability is that an AI system behaves exactly as its designers intended and anticipated. A reliable system adheres to the specifications it was programmed to carry out. Reliability is therefore a measure of consistency and can establish confidence and trust in the safety of a system based upon the dependability with which it conforms to its intended functionality. As part of reliability, the availability of relevant and high quality data is also important for reproduction, contextual accuracy, and ensuring continuity of the resource.
Accuracy and Performance	The accuracy of a model is the proportion of examples for which it generates a correct output. This performance measure is also sometimes characterised conversely as an error rate or the fraction of cases for which the model produces an incorrect output. Specifying a reasonable performance level for the system may also require refining or exchanging the measure of accuracy. For instance, if certain errors are more significant or costly than others, a metric for total cost can be integrated into the model so that the cost of one class of errors can be weighed against that of another. Margins of error should be communicated and rectified. It is important to note that accuracy and performance may change with scale and teams should be aware of, anticipate, and prepare for any changes where possible.

Table 2: **Accountability** Principle and Core Attributes

Core Attribute	Description
Traceability	Traceability refers to the process by which all stages of the data lifecycle from collection to deployment to system updating or deprovisioning are documented in a way that is accessible and easily understood. This may include not only the parties within the organisation involved but also the actions taken at each stage that may impact the individuals who use or are affected by the system.
Answerability	Answerability depends upon a human chain of responsibility. Answerability responds to the question of who is accountable for an

Core Attribute	Description
	automation supported outcome. There should be a point of ownership and responsibility identified, usually as a single point of contact at the first instance. Stakeholder transparency and communication is key to ensure clear lines of reporting. This includes making sure that there is a handover procedure in place where there is a shift in responsibility or ownership.
Auditability	Whereas the property of answerability responds to the question of who is accountable for an automation supported outcome, the notion of auditability answers the question of how the designers and implementers of AI systems are to be held accountable. This aspect of accountability has to do with demonstrating and evidencing both the responsibility of design and use practices and the justifiability of outcomes.
Clear Data Provenance and Lineage	Clear provenance and data lineage consists of records that are accessible. They simultaneously detail how data was collected and how it has been used and altered throughout the stages of pre-processing, modelling, training, testing, and deploying. This could include the use of version control or tracked changes as preservation of different versions at different points in time, for example, to outline how certain statistics were produced and what review processes were in place.
Accessibility	Accessibility involves ensuring that information about the processes that took place to design, develop, and deploy an AI system are easily accessible by individuals. This not only refers to suitable means of explanation (clear, understandable, and accessible language) but also the mediums for delivery.

Table 3: *Fairness Principle and Core Attributes*

Core Attribute	Description
Bias Mitigation	It is not possible to eliminate bias entirely. However, effective bias mitigation processes can minimise the unwanted and undesirable impact of systematic deviations, distortions, or disparate outcomes that arise to a project governance problem, interfering factor, or from insufficient reflection on historical social or structural discrimination.
Diversity and Inclusiveness	A significant component of fairness-aware design is ensuring the inclusion of diverse voices and opinions in the design and development process through the collaboration of a representative range of stakeholders. This includes considering whether values of civic participation, inclusion, and diversity have been adequately considered in articulating the purpose and setting the goals of the project. Consulting with internal organisational stakeholders is also necessary to strengthen the openness, inclusiveness, and diversity of the project, as well as its

Core Attribute	Description
	acceptance. External stakeholders, such as civil society, NGOs, and affected communities, may be sought directly. This ensures that a collaborative spirit is adopted within projects and services where relevant voices are in the room.
Non-Discrimination	A system or model should not create or contribute to circumstances whereby members of protected groups are treated differently or less favourably than other groups because of their respective protected characteristic.
Equality	The outcome or impact of a system should either maintain or promote a state of affairs in which every individual has equal rights and liberties, including equal treatment under the rule of law and equal access or opportunities to whatever good or service the AI system brings about.

Table 4: *Explainability Principle and Core Attributes*

Core Attribute	Description
Interpretability	Interpretability consists of the ability to know how and why a model performed the way it did in a specific context and, therefore, to understand the rationale behind its decision or behaviour. Within diverse teams, it may be the case where people who work on a specific model may not understand it fully, and so a wide range of expertise should be sought where there are knowledge gaps.
Responsible Model Section	The normal expectations of intelligibility and accessibility that accompany the function of the system, as fulfilled in the sector or domain in which it will operate. This can also necessitate the availability of more interpretable algorithmic models or techniques in cases where the selection of an opaque model poses risks to the physical, psychological, or moral integrity of rights-holders or to their human rights and fundamental freedoms. The availability of the resources and capacity that will be needed to provide responsible, supplementary methods of explanation (e.g. simpler surrogate models, sensitivity analysis, or relative feature important) in cases where an opaque model is deemed appropriate and selected.
Accessible Rationale Explanation	The reasons that led to a decision—especially one that is automated—delivered in an accessible and non-technical way. Where possible, this includes mitigating jargon and/or providing a glossary of terms to remove any assumptions regarding definitions. Ongoing communication of decisions is also important.
Implementation and User Training	Training users to operate the AI system may include: a) conveying basic knowledge about the nature of machine learning, b) explaining the limitations of the system, c) educating users about the risks of AI-

Core Attribute	Description
	related biases, such as decision-automation bias or automation-distrust bias, and d) encouraging users to view the benefits and risks of deploying these systems in terms of their role in helping humans to come to judgements, rather than replacing that judgement.
Reproducibility	Related to and dependant on the above four properties, reproducibility refers to the ability for others to reproduce the steps you have taken throughout your project to achieve the desired outcomes and where necessary to replicate the same outcomes by following the same procedure.

Table 5: *Data Stewardship Principle and Core Attributes*

Core Attribute	Description
Responsible Data Management	Responsible data management ensures that the team has been trained on how to manage data responsibly and securely, identifying possible risks and threats to the system and assigning roles and responsibilities for how to deal with these risks if they were to occur. Policies on data storage and public dissemination of results should be discussed within the team and with stakeholders, as well as being clearly documented.
Adequacy of Quantity and Quality	This attribute involves assessing whether the data available is comprehensive enough to address the problem set at hand, as determined by the use case, domain, function, and purpose of the system. Adequate quantity and quality should address sample size, representativeness, contextual relevance, and availability of features relevant to problem.
Source Integrity and Measurement Accuracy	Effective bias mitigation begins at the very commencement of data extraction and collection processes. Both the sources and instruments of measurement may introduce discriminatory factors into a dataset. When incorporated as inputs in the training data, biased prior human decisions and judgments—such as prejudiced scoring, ranking, interview-data or evaluation—will become the ‘ground truth’ of the model and replicate the bias in the outputs of the system in order to secure discriminatory non-harm, as well as ensuring that the data sample has optimal source integrity. This involves securing or confirming that the data gathering processes involved suitable, reliable, and impartial sources of measurement and sound methods of collection.
Timeliness and Recency	If datasets include outdated data, then changes in the underlying data distribution may adversely affect the generalisability of the trained model. Provided these distributional drifts reflect changing social relationship or group dynamics, this loss of accuracy with regard to the

Core Attribute	Description
Legal and Organisational Compliance	<p>actual characteristics of the underlying population may introduce bias into an AI system. In preventing discriminatory outcomes, timeliness and recency of all elements of the data that constitute the datasets must be scrutinised.</p> <p>Project teams are required to adhere to existing laws and regulations regarding data stewardship, including data protection law, privacy standards, and public sector duties.</p>

Appendix J: Bias Self-Assessment



This tool has been designed to help individuals and teams identify and reflect on a variety of biases that can impact the design, development, and deployment of AI/ML systems.

💡 What is 'bias' and why do we need to assess it?

Generally speaking, bias is a disproportionate effect on some phenomenon that is undesirable, inaccurate, or otherwise flawed.

There are many *types* of bias, which enter into the design, development, and deployment of data-driven systems. A key source of bias is that AI and related technologies are inherently 'sociotechnical systems' that emerge from and intervene in human relations and activities. AI and related technologies are products of human decisions, actions, and goals, which are reflected as patterns in data sets or as contributors to decision-making processes. **It is not possible to fully eliminate bias in technical systems**, but acknowledging and addressing bias is particularly important where such systems are used to support decision making or predictions that have significant consequences on people's lives and well-being.

This tool supports the identification and addressing of bias with a **taxonomy** of biases, a heuristic of the AI/ML **project lifecycle** to identify the entry points of bias into technical systems, a detailed '**reflect-list**' of individual biases with definitions and examples, and a table of **mitigation strategies** for addressing bias.

 **Tool contents**

Conceptual Resources:

1. Bias Taxonomy – A structure for identifying and evaluating biases according to three categories (i.e., statistical, cognitive, and social). Biases from each category are likely to require different actions or processes to mitigate, and, therefore, specialised skills or resources may be required to address the different types of bias.
2. Project Lifecycle Model – A summary of the Project Lifecycle Model, which can be used to ground the different biases in specific stages, according to a) where the biases are likely to have the most significant impact and b) where mitigation activities are most likely to be effective.
3. Bias Reflect-List – A list of biases, categorised according to the taxonomy (above), which are relevant in the context of design, development, and deployment of data-driven technologies.

Bias Self-Assessment: A process and set of best practices for using the three above elements to identify, mitigate, and document biases.

Worksheet: A template to support the operationalisation of the Bias Self-Assessment.

Conceptual Resources

Bias Taxonomy

Categorising biases through a taxonomy allows for the a) identification of how bias affects an AI/ML project and b) suggest potential mechanisms of action to counter the bias.

The range of biases that may occur throughout the project lifecycle can be understood within three main categories, as listed below.

- Social Bias
 - Refers to pre-existing or historical patterns of institutional and individual discrimination, behaviour, and social injustice, which can be drawn into the activities conducted throughout the project lifecycle. In particular, the term relates to how these patterns and attitudes can be perpetuated, reinforced, or exacerbated through the development and deployment of data-driven technologies.
- Statistical Bias
 - This term refers to a systematic deviation from an expected statistical result that arises due to the influence of some additional factor. This understanding is common in observational studies where bias can arise in the process of sampling or measurement. Statistical biases can involve errors (deviations from a true state) or differences between measured or calculated values and true values.

- Cognitive Bias
 - Refers to a deviation from a norm of rationality that can occur in processes of thinking or judgement and that can lead to mental errors, misinterpretations of information, or flawed patterns of response to decision problems.


Biases from each category can impact the various activities and stages of a project's lifecycle. Therefore, ongoing reflection and deliberation is required to minimise the possible negative impact upon downstream activities or the risk of discriminatory outcomes. The taxonomy creates a structured approach to bias identification and mitigation. It enables individuals and teams to a) identify how and whether the bias affects a specific project, b) who is responsible for mitigating the effects of the bias, and c) what actions need to be carried out to address the impacts of the bias (if any).

Because of the contextual nature of bias impact and assessment, it is not possible to give an exhaustive list of the mitigations that may be required for each bias. However, by grouping biases into these three categories, common themes and approaches can be identified to inform the development of mitigation strategies, as listed below.

Bias Mitigation Strategies

- Social Bias
 - Identifying relevant biases will require domain expertise and/or specialist knowledge to address the impact such biases could have on the project.
 - Understanding the scope and impact of a social bias will require diverse and meaningful engagement with stakeholders and communities.
 - Developing mitigation strategies will require wide-ranging discussions about the project's goals and objectives, and whether the risk of specific biases is significant enough to warrant mitigation (e.g., principle of proportionality).
 - The scope of what can be achieved from within the perspective of a project lifecycle is likely to be very limited or highly constrained. The minimal standard, therefore, should be ensuring a project does not exacerbate or perpetuate existing social biases. However, adherence to obligations such as the [Public Sector Equality Duty](#) should compel teams to go beyond this minimal duty.
- Statistical Bias
 - Speaking to a range of domain experts can help identify both the quantitative and qualitative limitations that can support seeking the appropriate technical or non-technical solutions.
 - Technical mitigation strategies will likely have a limited impact on the social biases responsible for giving rise to the statistical bias (e.g., imputation for missing data bias).
 - Although challenging and time-consuming, stakeholder engagement may be necessary to address data gaps that the project team may not have the expertise or lived experience to identify and tackle.

- Cognitive Bias
 - Individuals may be unable to identify the impact of a cognitive bias from within their team. As such, strategies like ‘red teams’ and ‘peer review’ are likely to be useful for identifying and mitigating cognitive biases.
 - Transparent documentation and project governance will help ensure long-term accountability for a project team’s decision-making and actions.
 - Meaningful inclusion and engagement from stakeholder groups can help minimise the impact of cognitive biases by creating a more diverse team with varying perspectives on problems (e.g., neurodiversity).

 **Additional themes and strategies**


What other themes or mitigation strategies can you think of for the three categories?

Project Lifecycle Model

Being able to ground biases in the Project Lifecycle Model is important because it helps to identify where the bias is likely to have the most significant impact and where mitigation activities are most likely to be effective. Therefore, for each bias (presented below) the following information is provided:

- **Lifecycle Scope:** The range of stages and typical activities in the project lifecycle that are likely to be impacted by the respective bias.
- **Significant Stage(s):** Those stages and typical activities where intervention is most likely to have an impact in mitigating the respective bias.
- **Illustrative Example:** Each bias also provides an illustrative example to help you and your team reflect upon how the respective bias may affect your own research or development.

The lifecycle scope and significant stage(s) serve as a guide to help you identify which biases are relevant in the context of your project, but also to help you identify when and where specific mitigation strategies are most likely to be effective. The purpose of grounding biases in the project lifecycle model, therefore, is also to help develop a transparent and accessible approach to accountable project governance.

 **Scope and stages**

The information provided in the ‘lifecycle scope’ and ‘significant stages’ sections should be treated as a guide only. In the context of your own project, a specific bias may be more or less significant in different stages of the project lifecycle due to the contextual nature of the bias.

You should be familiar with the individual project lifecycle stages before carrying out a self-assessment.

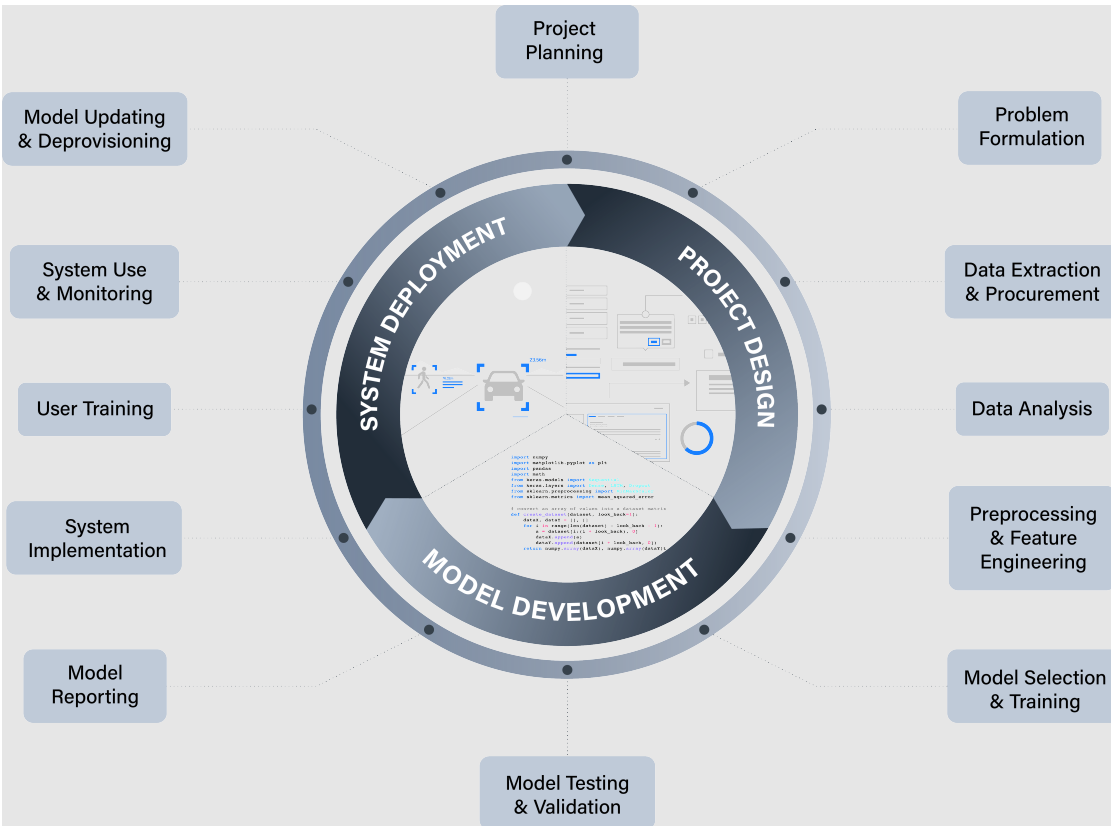


Figure 1 The Project Lifecycle Model

Bias Reflect-List

This section offers a breakdown of each bias under the social, statistical, and cognitive categories. Each bias is accompanied by a definition, illustrative example, significant stage(s) within the project lifecycle, lifecycle scope, and deliberative prompts.

Tip

When considering the biases, the team may choose to prioritise a subset of biases based on their perceived relevance for the project.



Social Biases

Below, we list nine social biases with descriptions and illustrative examples: historical, representation, label, annotation, chronological, selection, implementation, de-identification, and status quo bias.

Historical bias

Historical biases exist prior to the inception of any AI project, and they can exist even where data are responsibly sampled, collected, and processed.

A closer look at historical bias

Historical biases arise in AI innovation contexts when there is a gap or misalignment between the state of the world and the goals or objectives of the project and system being developed. Such a gap allows for historical patterns of inequity or discrimination to be reproduced, or even augmented, in the development and use of the system even when the system is functioning to a high standard of accuracy and reliability. For instance, even with perfect sampling and feature selection, a project will exhibit bias where it perpetuates (or exacerbates) socioeconomic inequalities through the outcomes it promotes, or the deployment of the system being developed.

Illustrative Example

Examples of historical bias include social dynamics that contribute to prejudicial arrest rates in policing, or social determinants of criminal behaviour and outcomes, such as poverty that can create higher risks of recidivism.

Significant Stages

- Project Planning
- Problem Formulation
- System Use and Monitoring

Lifecycle Scope


Pre-exists lifecycle

Deliberative prompts for historical bias

- Which internal or external stakeholders will be affected by the use of your model?
- Are there internal or external stakeholders that will be excluded from your model or experience barriers to using your system? If so, why?
- Is there a risk of worsening or perpetuating socioeconomic inequalities in the development and deployment of your model?

Representation bias

When a population is either inappropriately represented (e.g., not allowing sufficient self-representation in demographic variables) or a sub-group is under-represented in the dataset, the model may subsequently fail to generalise and under-perform for a sub-group (or sub-groups).

 **A closer look at representation bias**

Illustrative Example

An example of representation bias could be an app that allows citizens to report instances of “criminal behaviour” in their neighbourhood, but where some neighbourhoods are likely to report behaviours such as recreational drug use (e.g., smoking marijuana) whereas other neighbourhoods are likely to ignore such behaviours. Here, certain neighbourhoods may be over-represented for crime rates, when compared to others.

Significant Stages

- Problem Formulation
- Data Extraction & Procurement
- Pre-Processing and Feature Engineering

Lifecycle Scope

Project Planning → System Use and Monitoring

 **Deliberative prompts for representation bias**

- How have you measured and evaluated the representativeness of the dataset to ensure that the sample is adequate?
- Have you consulted the relevant stakeholder groups to verify that your dataset is representative?

Label bias

A label (or feature) used within an algorithmic model may not mean the same thing for all data subjects. There may be a discrepancy between what sense the designers are seeking to capture in a label or feature, or what they are trying to measure in it, and the way that affected individuals understand its meaning.

 **A closer look at label bias**

Where there is this kind of variation in meaning for different groups within a population, adverse consequences and discriminatory impact could follow. For example, designers of a predictive model in public health may choose “patient wellbeing” as their label, defining it in terms of disease prevalence and hospitalisation. However, subpopulations who suffer from health disparities and socioeconomic deprivation may understand wellbeing more in terms of basic functionings, the food security needed for health promotion, and the absence of the social environmental stressors that contribute to the development of chronic medical conditions. Were this predictive model to be used to develop public health policy, members of this latter group could suffer from a further entrenchment of poor health outcomes.

Illustrative Example

An example of label bias could be variation in the meaning of data categories over time (e.g., the geographic boundaries of court districts) or the meaning of social categories like race or gender if they have been self-reported.

Significant Stages

- Problem Formulation
- Pre-Processing and Feature Engineering

Lifecycle Scope

Project Planning → System Use and Monitoring



Deliberative prompts for label bias

- How have you identified problematic labels (or features), which may be imperfect proxies, within your dataset?
- Does your target variable have multiple meanings or interpretations?
- Are labels used across the project lifecycle and have they been clearly defined?

Annotation bias

Annotation bias occurs when annotators incorporate subjective perceptions or error into the work of annotating data.



A closer look at annotation bias

Data annotation often occurs under less than ideal scenarios, including contexts in which human error may be introduced due to fatigue or lack of focus, or from the insufficient training of annotators. Annotation bias can also result from positionality limitations that derive from demographic features, such as age, education, or first language, as well as other systemic cultural or societal biases that influence annotators.

Illustrative Example

An example of annotation bias is when police officers misidentify the race or ethnicity of a criminal suspect in an arrest report due to uncertainty or personal bias. Data sets produced in this context may misrepresent the prevalence of arrests amongst demographic subgroups, leading to erroneous conclusions about crime trends.

Significant Stages

- Problem Formulation
- Data Extraction or Procurement
- Pre-Processing and Feature Engineering

Lifecycle Scope

Project Planning → Pre-Processing and Feature Engineering



Deliberative prompts for annotation bias

- Who carried out the annotation of your dataset? What methods did they follow?
- Were there processes in place to ensure that multiple annotators followed the same standards (e.g., inter-rater reliability)?

Chronological bias

Chronological bias arises when individuals in the dataset are added at different times, and where this chronological difference results in individuals being subjected to different methods or criteria of data extraction based on the time their data were recorded.



A closer look at chronological bias

Illustrative Example

An example of chronological bias could be where a dataset used to build a predictive risk model in children's social care has data that spans over several years, in which large-scale care reforms, policy changes, adjustments in relevant statutes (such as changes to legal thresholds or definitions) have occurred. As such, there may also have been changes in data recording methods that could create major inconsistencies in the data points extracted from person to person.

Significant Stages

- Project Planning
- Data Extraction or Procurement

Lifecycle Scope

Project Planning → Data Analysis

Deliberative prompts for chronological bias

- Have you worked with domain experts to map the data journey and identify systematic variations between groups of data subjects or - Is there a wide variation in when your data were recorded?

Selection bias

Selection bias is a term used for a range of biases that affect the selection or inclusion of data points within a dataset.

A closer look at selection bias

In general, this bias arises when an association is present between the variables being studied and additional factors that make it more likely that some data will be present in a dataset when compared to other possible data points in the space. For instance, where individuals differ in their geographic or socioeconomic access to an activity or service that is the site of data collection, this variation may result in exclusions from the corresponding dataset based on those differences. Likewise, where certain socioeconomically deprived or marginalised social groups are disproportionately dependent on a social service to fulfil basic needs, members of those groups may be oversampled if data is collected from the provision of that service.

Illustrative Example

An example of selection bias is where pregnant women are routinely not selected for drug trials, due to increased risks. However, while safeguarding them during pregnancy, their lack of inclusion also leads to lower efficacy for their cohort (e.g. real-world lack of efficacy for certain pain killers).

Significant Stages

- Project Planning
- Data Extraction or Procurement

Lifecycle Scope

Project Planning → Data Analysis

Deliberative prompts for selection bias

- Have you examined the different stakeholders that are included or not included within the data and datasets are being considered?
- Are there stakeholder groups you can consult with to help minimize the likelihood of you and your team missing key stakeholder considerations?

Implementation bias

Implementation bias refers, generally, to any bias that arises when a system is implemented or used in ways that were not intended by the designers or developers but, nevertheless, made more likely due to affordances of the system or its deployment.

A closer look at implementation bias

Illustrative Example

Consider a biometric identification system that was initially designed by a public authority to assist in the detection of potential terrorist activity but is now repurposed to target and monitor non-violent activists or political opponents.

Significant Stages

- User Training
- System Use and Monitoring

Lifecycle Scope

Model Implementation → System Use and Monitoring



Deliberative prompts for implementation bias

- Has your system been repurposed from another project or team? If so, is the system fit-for-purpose?
- Does the use of the system now differ from how it was previously used?

Status quo bias

An affectively motivated preference for “the way things are currently”, which can prevent more innovative or effective processes or services being implemented.



A closer look at status quo bias

Illustrative Example

This bias can occur in cases where people are more critical of technological systems, even though they may outperform biased human decision-making. For example, a decision support system used to help school staff identify under-performing children, where the technological system has some biases, but these are less significant or impactful than the existing human biases that allow some students to fall “beneath the radar”.

Significant Stages

- Model Updating or Decommissioning
- Project Planning

Lifecycle Scope

Model Updating or Decommissioning → Project Planning



Deliberative prompts for status quo bias

- Have you assessed how your team members feel about the use or lack of use of technology in your project? Is this different to how things have usually been done within your team?
- Are you able to consult with something outside of your team to see if your project as well as the proposed problem and solution are appropriate?

De-agentification bias

De-agentification bias occurs when social structures and innovation practices systemically exclude minoritised, marginalised, vulnerable, historically discriminated against, or disadvantaged social groups from participating or providing input in AI innovation ecosystems.

A closer look at de-agentification bias

Protected groups may be prevented from having input into the development, use, and evaluation of models. They may lack the resources, education, or political influence to detect biases, protest, and force correction.

Illustrative Example

An example is the choice to design, develop, or deploy a system for monitoring historically marginalised communities, such as refugees and religious minorities. Such communities are often not represented in key decisions concerning the adoption and use of such systems though they may be significantly affected by them.

Significant Stages

- Project Planning
- Project Formulation

Lifecycle Scope

Project Planning → Model Updating or Decommissioning

Deliberative prompts for de-agentification bias

- Have you considered consulting, engaging, and working with protected and marginalized groups as part of your project? How have their perspectives and experiences been considered?



Statistical Biases

Below, we list seven statistical biases with descriptions and illustrative examples: missing data, measurement, wrong sample size, aggregation, evaluation, confounding, and training-serving skew bias.

Missing data bias

Missing data can cause a wide variety of issues within an AI project, and these data may be missing for a variety of reasons related to broader social factors.

***i* A closer look at missing data bias**

Missingness can lead to inaccurate inferences and affect the validity of the model where it is the result of non-random but statistically informative events. For instance, missing data bias may arise in predictive risk models used in social care where interview questions about socially stigmatised behaviours or traits like drug use or sexual orientation trigger fears of punishment, humiliation, or reproach and thus prompt non-responses.

Illustrative Example

Unhoused people are often under-counted or missing from health and benefits data sets because they are less likely to seek services or to seek them in the same location over time than people with stable housing. As a result, the needs and interests of this population may not be reflected in AI models trained on this data.

Significant Stages

- Data Analysis
- Model Selection and Training
- Model Testing and Validation

Lifecycle Scope

All Stages



Deliberative prompts for missing data bias

- How have you dealt with and recorded your handling of missing data (e.g., choice of imputation or augmentation method)?
- Have you consulted with domain experts to help you identify possible explanations for the missing data and whether they may be informative?

Measurement bias

This bias addresses the choice of how to measure the labels or features being used.



A closer look at measurement bias

It arises when the measurement scale being applied fails to capture data pertaining to the subjects in a fair and equitable manner.

Illustrative Example

A model used by police or courts of law to predict future criminality based on data detailing the prior arrests or criminal records of a person's relatives may produce measurement bias because patterns of arrest are not entirely objective. Arrest statistics can reflect discriminatory tendencies by police forces to focus on certain social groups or communities, or may reflect the personal biases of arresting officers.

Significant Stages

- Data Extraction or Procurement

Lifecycle Scope

Project Planning → Pre-processing and Feature Engineering



Deliberative prompts for measurement bias

- Are there multiple scales that could be used to measure your features? Is there reasonable disagreement about which of these scales is preferred? If so, how has this disagreement been addressed?

Wrong sample size bias

Using the wrong sample size for the study can lead to chance findings that fail to adequately represent the variability of the underlying data distribution, in the case of small samples, or findings that are statistically significant but not relevant or actionable, in the case of larger samples.



A closer look at wrong sample size bias

Wrong sample size bias may occur in cases where model designers have included too many features in a machine learning algorithm. This is often referred to as the “curse of dimensionality”, a mathematical phenomenon wherein increases in the number of features or “data dimensions” included in an algorithm means that exponentially more data points need to be sampled to enable good predictive or classificatory performance.

The ‘wrong’ sample size does not just have to refer to datasets that are too small (e.g. insufficient for generalisable inferences). A dataset can also have too many variables for the problem or use case and can introduce noise and unmanageable sparsity as well as going against the ‘data minimisation’ principle.

Illustrative Example

In time series data, this can also create a situation where the ‘fidelity’ of the dataset is too much for the computational resources of the organisation (e.g. weather forecasting over-sampling from monitors).

Significant Stages

- Data Extraction or Procurement
- Pre-processing or Feature Engineering

Lifecycle Scope

Project Planning → Model Testing and Validation



Deliberative prompts for wrong sample size bias

- Which methods or statistical indicators (e.g., p-values, confidence intervals) have been used and reported to help ensure that the findings did not arise by chance?
- Have you considered the likely use case for the results? How will this be reported (e.g., in ‘limitations’ section) to help readers assess the relevance of the results?

Aggregation bias

Aggregation bias arises when a “one-size-fits-all” approach is taken to the outputs of a trained algorithmic model (i.e. that model results apply evenly to all members of the impacted population) even where variations in subgroup characteristics mean that mapping functions from inputs to outputs are not consistent across these subgroups.



A closer look at aggregation bias

In other words, in a model where aggregation bias is present, even when combinations of features affect members of different subgroups differently, the output of the system disregards the relevant variations in condition distributions for the subgroups. This results in the loss of information, lowered performance, and, in cases where data from one subgroup is more prevalent than those of others, the development of a model that is more reliable for that sub-group.

Illustrative Example

Examples of aggregation bias include clinical decision-support systems in medicine, where clinically significant variations between patient cohorts (e.g. different sexes and ethnicities)—in terms of disease aetiology, expression, complications, and treatment—mean that systems which aggregate results by treating all data points similarly will not perform optimally for any subgroup.

Significant Stages

- Pre-processing or Feature Engineering

Lifecycle Scope

Pre-processing or Feature Engineering → System Use and Monitoring



Deliberative prompts for aggregation bias

- Which evaluation methods (e.g., model comparison) have you employed to help you identify aggregation bias and its impact on the various subgroups in your dataset?

Evaluation bias

Evaluation bias occurs during model iteration and evaluation and evaluation from the application of performance metrics that are insufficient given the intended use of the model and the composition of the dataset on which it is trained.



A closer look at evaluation bias

Illustrative Example

Evaluation bias may occur where performance metrics that measure only overall accuracy are applied to a trained computer vision system that performs differentially for subgroups that have different skin tones. Likewise, evaluation biases arise where the external benchmark datasets that are used to evaluate the performance of trained models are insufficiently representative of the populations to which they will be applied. In the case of computer vision, this may occur where established benchmarks overly represent a segment of the populations (such as adult light-skinned males) and thus reinforce the biased criteria for optimal performance.

Significant Stages

- Data Analysis
- Model Selection and Training
- Model Testing and Validation

Lifecycle Scope

Data Analysis → Model Updating or Decommissioning

Deliberative prompts for evaluation bias

- How will you divide your dataset into separate training and testing datasets?
- Will you validate the model against an external benchmark population? If not, have you taken steps to report these limitations?

Confounding

Confounding is a well-known causal concept in statistics, and commonly arises in observational studies.

A closer look at confounding

It refers to a distortion that arises when a (confounding) variable independently influences both the dependent and independent variables (e.g., exposure and outcome), leading to a spurious association and a skewed output.

Illustrative Example

Clear examples of confounding can be found in the use and analysis of electronic health records (EHRs). EHRs are observational data and often reflect not only the health status of patients, but also patients' interactions with the healthcare system. This can introduce confounders such as the frequency of inpatient medical testing reflecting the busyness or labour shortages of medical staff rather than the progression of a disease during hospitalisation, differences between onset of a disease and the date of diagnosis, and health conditions that are missing from the EHRs of a patient due to a non-random lack of testing. Contextual awareness and domain knowledge are crucial elements for identifying and redressing confounders.

Significant Stages

- Data Analysis

Lifecycle Scope

Data Analysis → Model Reporting

Deliberative prompts for confounding

- Are there methods you can use (e.g., propensity score matching, causal diagrams) that could help reduce bias that results from confounding (e.g., in the estimation of the average treatment effect)?
- Is the sample size sufficient (i.e., large enough) to minimise the impact of confounders?

Training-serving skew

Occurs when the model is deployed on individuals whose data are not similar to or representative of the individuals whose data were used to train, test, and validate the model.

A closer look at training-serving skew

This can occur, for instance, where a trained model is applied to a population in a different geographical area from that where the original data were collected or to the same population but at a time much later than that at which the training data were first collected. In both cases, the trained model may fail to generalise because the new, out-of-sample inputs are being drawn from populations with different underlying distributions.

Illustrative Example

Consider a model that predicts credit risk for loan applicants. The model is trained on a dataset that includes information about the loan applicants, such as their income, employment history, and credit score. But there is a disproportionate number from one demographic group in particular (e.g. elderly applicants).

As we train the model on this dataset, it may learn to associate certain characteristics with lower credit risk, which are not representative of the underlying relationship in the broader population.

When the model is deployed in production, therefore, and used to make predictions for loan applicants from other demographic groups (e.g., younger applicants), the model's performance will be biased in favour of the older applicants.

Significant Stages

- Model Selection and Training
- Model Testing and Validation
- System Use and Monitoring

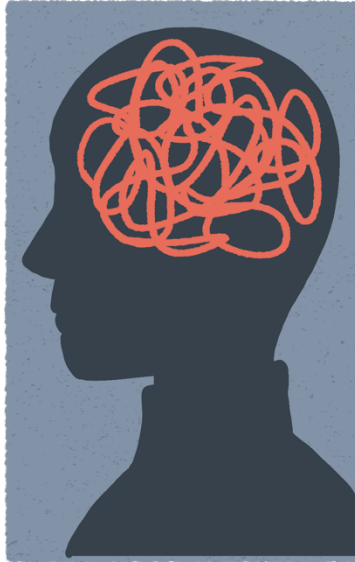
Lifecycle Scope

Data Extraction or Procurement → System Use and Monitoring



Deliberative prompts for training serving skew

- What steps have you taken to measure and evaluate the performance of your model within the intended domain (e.g., use of synthetic data, external validation on similar datasets)?
- Have you engaged domain experts to ensure these steps are adequate (e.g., sufficiently representative of the impacted users)?



Cognitive Biases

Below, we list nine cognitive biases with descriptions and illustrative examples: confirmation, self-assessment, availability, naïve realism, Law of the Instrument (Maslow's Hammer), optimism, decision-automation, automation-distrust, and semantic bias.

Confirmation bias

Confirmation biases arise from tendencies to search for, gather, or use information that confirms pre-existing ideas and beliefs, and to dismiss or downplay the significance of information that disconfirms one's favoured hypothesis.

 **A closer look at confirmation bias**

Confirmation bias can be the result of motivated reasoning or sub-conscious attitudes, which in turn may lead to prejudicial judgements that are not based on reasoned evidence. For example, confirmation biases could surface in the judgment of the user of an AI decision-support application, who believes in following common sense intuitions acquired through professional experience rather than the outputs of an algorithmic model and, for this reason, dismisses its recommendations regardless of their rational persuasiveness or veracity.

Illustrative Example

Consider a policymaker or minister who has strong attitudes on the economic impacts of immigration. If their pre-existing stance leads to them ignoring or downplaying models that serve as evidence against their views, and only considering models that support their existing attitudes, they are suffering from confirmation bias.

Significant Stages

- Problem Formulation
- Data Analysis
- System Use and Monitoring

Lifecycle Scope

Whole Lifecycle

Deliberative prompts for confirmation bias

- What mechanisms do you have in place within your team that can help ensure a diversity of viewpoints that may mitigate the effects of confirmation bias?

Self-Assessment bias

A tendency to evaluate one's abilities in more favourable terms than others, or to be more critical of others than oneself.

A closer look at Self-assessment bias

In the context of a project team, this could include the overly-positive assessment the group's abilities (e.g., through reinforcing groupthink).

Illustrative Example

Consider a project team that is carrying out an assessment about whether they have sufficient skills and resources to develop fair and explainable ML algorithms. Self-assessment bias could create a situation where the team are unlikely to acknowledge or notice gaps in their skills, which may significantly affect their ability to deliver a responsible product.

Significant Stages

- Project Planning

Lifecycle Scope

Whole Lifecycle



Deliberative prompts for self-assessment bias

- As part of your project planning, have you considered what may go wrong or have a negative impact?
- Are you able to be more flexible with your timeline to accommodate for identifying and addressing gaps of knowledge and skills within your team?
- Have you and your project team considered obtaining constructive criticism and suggestions from others?

Availability bias

The tendency to make judgements or decisions based on the information that is most readily available (e.g., more easily recalled).



A closer look at availability bias

When this information is recalled on multiple occasions, the bias can be reinforced through repetition—known as a ‘cascade’. This bias can cause issues for project teams throughout the project lifecycle where decisions are influenced by available or oft-repeated information (e.g., hypothesis testing during data analysis).

Illustrative Example

If a team uses a dataset that they already have access to, even though it is not actually the best data for their problem, this is a form of availability bias.

However, this is different from the form of availability bias that may affect people when recalling certain facts throughout a project’s lifecycle. Here, availability refers to the individuals ability to recall information, rather than to an ability to access data.

Significant Stages

- Data Analysis
- Model Selection and Training
- Model Testing and Validation

Lifecycle Scope

Whole Lifecycle



Deliberative prompts for availability bias

- Have you considered alternative sources, references, datasets, and methods that can help minimize gravitating towards readily available or memorable information?

Naïve realism bias

A disposition to perceive the world in objective terms that can inhibit recognition of socially constructed categories.



A closer look at naïve realism

Illustrative Example

An example of naïve realism would be treating ‘employability’ as something that is objectively measurable and, therefore, able to be predicted by a machine learning algorithm on the basis of objective factors (e.g., exam grades, educational attainment).

Significant Stages

- Problem Formulation

Lifecycle Scope

Project Planning → Pre-processing and Feature Engineering



Deliberative prompts for naïve realism

- Have you identified non-quantifiable or difficult-to-measure qualitative factors that may contribute to and affect your model or decision-making process? How are these documented and accounted for?

Law of the Instrument (Maslow’s Hammer)

This bias is best captured by the popular phrase ‘If all you have is a hammer, everything looks like a nail’.

A closer look at law of the instrument (Maslow’s Hammer)

The 'if all you have is a hammer, everything looks like a nail' phrase cautions against the cognitive bias of over-reliance on a particular tool or method, perhaps one that is familiar to members of the project team. For example, a project team that are experts in a specific ML technique, may over-use the technique and mis-apply it in a context where a different technique would be better suited. Or, in some cases, where it would be better not to use ML/AI technology at all.

Illustrative Example

If an organisation develops a system to parse natural language, and successfully deploys it for one task, but then uses it in a new project without considering whether it is the right tool, they are falling prey to this bias.

Significant Stages

- Project Planning
- Model Selection and Training
- Model Testing and Validation

Lifecycle Scope

Whole Lifecycle



Deliberative prompts for law of the instrument (Maslow's Hammer)

- Is the technology you're developing the best way forward for your project? Who has determined this?
- If you're repurposing an existing technology, is it fit-for-purpose for the task and project at hand?
- Does your team have the appropriate knowledge and skillset to adopt the current system, model or tool?

Optimism bias

Also known as the planning fallacy, optimism bias can lead project teams to underestimate the amount of time required to adequately implement a new system or plan.



A closer look at optimism bias

In the context of the project lifecycle, this bias may arise during project planning, but can create downstream issues when implementing a model during the 'model productionalisation' stage, due to a failure to recognise possible system engineering barriers.

Illustrative Example

During project scoping, a project management team incorrectly assume that it will only take 3 months to design, develop, and deploy a new algorithmic system, because a previous (and similar) project took this long. However, despite the success of the previous project, their assessment this time turns out to be an underestimate because they did not consult with their developers to fully understand important differences between the two projects.

Significant Stages

- Model Implementation
- System Use and Monitoring

Lifecycle Scope

Whole Lifecycle



Deliberative prompts for optimism bias

- Have you and your team been realistic with what can be achieved within the time allocated to the project?
- Are you able to be more flexible with your time and resources, particularly where stakeholder engagement is involved?

Decision-automation bias

Decision-automation bias arises when users of automated decision-support systems become hampered in their critical judgment, rational agency, and situational awareness as a result of their faith in the efficacy of the system.



A closer look at decision-automation bias

Decision-automation bias may lead to over-reliance or errors of omission, where implementers lose the capacity to identify and respond to the faults, errors, or deficiencies, which might arise over the course of the use of an automated system, because they become complacent and overly deferent to its directions and cues. Decision-automation bias may also lead to over-compliance or errors of commission where implementers defer to the perceived infallibility of the system and thereby become unable to detect problems emerging from its use for reason of a failure to hold the results against available information.

Illustrative Example

An immigration officer is using facial recognition software, which purportedly claims to detect instances of lying during asylum claim interviews. Over time, the officer stops relying on their own faculties, and leans too heavily on the predictions of this system, despite visual cues that contradict the facial recognition system's predictions.

Significant Stages

- User Training

Lifecycle Scope

User Training → System Use and Monitoring



Deliberative prompts for decision-automation bias

- Have you considered user requirements such as transparency or interpretability when designing your model?
- Does the intended use domain demand a greater need for interpretability, and how may this affect the model's accuracy (e.g., reducing model complexity)?
- Could long-term use of your model or system have a detrimental effect on the professional judgement of users (e.g., leading to deskilling)?

Automation-distrust bias

Automation-distrust bias arises when users of an automated decision-support system disregard its salient contributions to evidence-based reasoning either as a result of their distrust or scepticism about AI technologies in general or as a result of their over-prioritisation of the importance of prudence, common sense, and human expertise.



A closer look at automation-distrust bias

An aversion to the non-human and amoral character of automated systems may also influence decision subjects' hesitation to consult these technologies in high impact contexts such as healthcare, transportation, and law.

Illustrative Example

Members of a profession (e.g., judges, doctors) who rule out decision support systems based on (potentially unfounded) fears of these technologies, may be influenced by this bias. However, in such instances, their aversion to using technology in a constructive way, may prevent them from identifying and mitigating some of their own cognitive biases, or improving evidence-based decisions in their respective fields.

Significant Stages

- System Use and Monitoring

Lifecycle Scope

User Training → System Use and Monitoring



Deliberative prompts for automation-distrust bias

- Have you engaged intended users early on in project planning to identify barriers and co-design solutions that would increase the level of trust they have in your system?
- Is there information you could provide to help reduce any concerns users would have about how your model or system operates?

Semantic bias

Semantic bias occurs when discriminatory inferences are allowed to arise in the architecture of a trained model and to remain an element of the deployed system.



A closer look at semantic bias

When historical biases are baked into datasets in the form of discriminatory proxies or embedded prejudices (e.g., word embeddings that pick up on racial or gender biases), these biases can be semantically encoded in the model’s covariates and parameters. Semantic biases occur when model design and evaluation processes fail to detect and mitigate such discriminatory aspects.

Illustrative Example

An example of semantic bias is the use of discriminatory or value-laden terms to describe a person, behaviour, or phenomenon. Labelling a person who identifies as a woman as “aggressive” while labelling person who identifies as a man exhibiting similar behaviour as “assertive” may import subjective beliefs about gender-specific norms into a system.

Significant Stages

- Data Analysis
- System Use and Monitoring

Lifecycle Scope

Project Planning → Pre-processing and Feature Engineering



Deliberative prompts for semantic bias

- When codifying data have you and your team considered the semantic undertones of the words used?
- If you were placed in the shoes of the person, action, or phenomenon being labelled as a data point, would you use that word to describe it?

Bias Self-Assessment

Warning

This self-assessment should not be treated as a *checklist* that needs to be completed at the end of a project as a form of compliance. Doing so, reduces the value of the self-assessment to a simple tick-box exercise and minimises the scope of reflection and deliberation that is so vital to an effective self-assessment. This is why we refer to the list of biases as a *reflect-list*, rather than a *checklist*, and present the self-assessment as a high-level *procedure* that requires practical implementation within a project.

There are several ways this self-assessment could be implemented, but the following procedure provides a high-level overview that can serve as a starting point for your own implementation:

1. Carry out a preliminary assessment using the project lifecycle model and bias reflect-list
 - Go through each stage and identify those biases that could impact upon your project's goals and objectives
 - Evaluate the severity of impact from each bias to produce a list of the most significant biases
 - Identify any mitigation strategies that could be used to reduce the impact of each bias
2. Engage stakeholders and domain experts
 - Review the preliminary list of biases that could impact your project's goals and objectives—add or remove biases as necessary
 - Review the severity of the impact from each bias
 - Review the mitigation strategies that could be used to reduce the impact of each bias
3. Develop a draft bias mitigation plan
 - A list of the most significant biases that could impact your project's goals and objectives. Provide an explanation for why each bias is significant.
 - A list of mitigation strategies that will be used to reduce the impact of each bias. Provide an explanation for why each mitigation strategy is appropriate.
 - A list of any biases that were identified but require no action. Provide an explanation for why no action is required.
 - A list of any mechanisms that will be established to identify and address new biases that are identified during the project.
4. Review the plan at regular intervals

- Review the plan at regular intervals to ensure that the plan is still appropriate and that the project is on track to meet its goals and objectives.
 - Update the explanations from stage 3, to help document why initial mitigation strategies may have changed.
 - Revisit the plan with stakeholders and domain experts to ensure that the plan is still appropriate.
5. Publish (internally or externally) the bias mitigation plan as a record of the project’s decision-making process
- Review the plan within your team and with other relevant groups to help build best practices or identify any knowledge or capabilities gaps.

Table 6.1 presents avenues in which bias can be mitigated. It is a non-exhaustive list of potential bias mitigation techniques.

Table 6.1: Summary of Bias Mitigation Techniques (presented in alphabetical order)

Bias Mitigation Technique	Description
Additional Data Collection	Return to the data extraction (or procurement) stage to carry out additional data collection or reconsider methods of data extraction (e.g. revised experimental methods, more inclusive and accessible forms of engagement).
Data Augmentation	Augment your dataset using techniques appropriate to the objective (e.g. addressing sparsity), such as data linkage or mixing, synthetic data generation, imputation, adding noise, transformation.
Double Diamond Methodology	The Double Diamond methodology is a process for design that is well-suited to creative approaches to problem-solving and exploring multiple perspectives and possibilities. The method consists of four phases: 1. Discover: gain insight and identify the problem, understanding needs and challenges, and gather information in a highly exploratory manner. 2. Define: clarify the information from the previous stage to gain a narrower, well-defined area to focus on. 3. Develop: generate and test possible solutions, exploring the feasibility and desirability of the solutions, while also identifying areas that need additional work. 4. Deliver: deliver a final product or service that meets the original specification (e.g. minimum viable product), and which can be used to gather additional feedback.
Diversify Evaluation Metrics	Use additional evaluation metrics for your model to determine whether its performance applies equally for all individuals or sub-groups. Where relevant carry out intersectional analysis of multiple demographic or identity characteristics to identify biases

Bias Mitigation Technique	Description
Employ Model Interpretability Methods	that may not be apparent when considering a single characteristic.
External Validation	During data analysis, model testing and validation, and system use and monitoring, use appropriate model interpretability methods (e.g. local, model-agnostic, data visualisation) to ensure that your model is meeting the original objectives for your project.
Human-in-the-loop	Go beyond the <i>internal</i> validation of your model (i.e. training-testing split of data) and perform <i>external</i> validation with an entirely new dataset. You could engage with another team or organisation to help validate your study or model development in a new environment (e.g. different population of data subjects, novel geographical environment).
Identify Under-represented Groups	Agree on guidelines to ensure the use of data-driven technologies support human decision making by providing recommendations or automating routine tasks, while still allowing humans to make final decisions and have clear oversight.
Multiple Model Comparison	Analyse gaps in demographic data in consultation with community groups and domain experts. Develop appropriate methods to address gaps and limitations based on context-aware reflection.
Open Documentation	Train and test multiple models, both within the same class of models and also across classes to assess a broader range of possible performance values.
Participatory Design Workshops	Where possible, document the actions and decisions made throughout your project to support reproducibility and replicability efforts, assist users of your system, and promote best practices of transparency.
Peer Review	A form of stakeholder engagement that seeks to involve stakeholders within the design process to identify needs and preferences, co-create solutions, and ensure usability and acceptance.
Peer Review	Targeted review of work by an internal or external committee, red team, or other group to identify and evaluate any gaps or issues.

Bias Mitigation Technique	Description
Quality Control Procedures	Conduct regular assessments of your model or system against established quality control procedures (e.g. analytical quality assurance) to ensure that issues are identified early on (e.g. clerical errors in data input that may arise from time-pressured human inputters or annotators).
Regular Auditing	Work with another team, committee, or organisation to perform regular audits of your project, focusing on key areas such as transparency and explainability, data quality, model performance, user satisfaction, and equitable impact.
Skills and Training	Organise and facilitate skills and training events, such as webinars, workshops, self-directed learning, to upskill project team members or users (e.g. understanding and communicating uncertainty of predictive models, interactions with system interface).
Stakeholder Engagement	Carry out meaningful forms of engagement to consult or partner with wider stakeholders. This could include hosting community fora, conducting online surveys or interviews, or even running a citizen jury or assembly.



Iterative documentation and reporting

As you can see from the above procedure, a bias mitigation plan is a living document, which also serves as input into other project deliverables. Therefore, you may wish to approach the iterative development of bias mitigation using tools and services that are familiar to your team (e.g., version control and tracking software). Similarly, you may find it useful to identify several key stages where an interim report could be documented (e.g., summary of bias reflection and mitigation activities that have already been carried out). For instance, you could choose to publish two reports at the end of the *project design* and *model development* stages that document which biases have been mitigated (and how), which biases require action at a future stage, and which biases require no action.

Furthermore, if you wished to build constraints into the project's decision-making and governance you could use a traffic light system such as the following:

- Green: biases that have been satisfactorily mitigated
- Amber: biases you have determined pose minimal risk and have not actioned
- Red: biases where significant risk remains

Doing so would require you to develop some sort of calculus for determining the severity of risk that each bias poses, and then agreeing on whether a project should continue while, say, an amber bias remains unmitigated. Such decisions are highly contextual and prescriptive though, so they should be discussed within your team and with other relevant groups and stakeholders during the design and discover of a project.

Selected Bibliography

Baker, T., Smith, L., & Anissa, N. (2019). Educ-AI-tion Rebooted? Exploring the future of artificial intelligence in schools and colleges. 56.

https://media.nesta.org.uk/documents/Future_of_AI_and_education_v5_WEB.pdf

Bennett, S., & Cutler, N. (2019, October 28). Lab Long Read: Policy Consultations - Part

2: A role for data science?. Policy Lab and Department for Transport (DfT).

<https://openpolicy.blog.gov.uk/2019/10/28/lab-long-read-policy-consultations-part-2-role-of-data-science/>

Berryhill, J., Heang, K. K., Clogher, R., & McBride, K. (2019). Hello, World: Artificial intelligence and its use in the public sector. OECD Working Papers on Public Governance (36), OECD Publishing. https://www.oecd-ilibrary.org/governance/hello-world_726fd39d-en

Burke, A. (2020). Robust artificial intelligence for active cyber defence. Alan Turing Institute: Defence and Security Programme. Retrieved from:

https://www.turing.ac.uk/sites/default/files/2020-08/public_ai_acd_techreport_final.pdf

Cambridge Consultants. (2019). Use of AI in Online Content Moderation: 2019 report produced on behalf of Ofcom.

https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf

Central Digital and Data Office. (2020). Using chatbots and webchat tools: How to use chatbots and webchat tools to improve your users' experience of your service.

<https://www.gov.uk/guidance/using-chatbots-and-webchat-tools>

Dencik, L., Hintz, A., Redden, J. and Warne, H. (2018) Data Scores as Governance: Investigating uses of citizen scoring in public services. Research Report. Cardiff University.

<https://datajusticelab.org/wp-content/uploads/2018/12/data-scores-as-governance-project-report2.pdf>

Department for Digital, Culture, Media, and Sport & Home Office. (2020). Online Harms White Paper: Full government response to the consultation.

<https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response>

Geburu, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2020). Datasheets for Datasets. arXiv:1803.09010 [Cs].

<http://arxiv.org/abs/1803.09010>

Griffiths, H. (2016). IoT Adoption Among Cities in the UK (27). IoTUK.

https://iotuk.org.uk/wp-content/uploads/2016/08/IoT_Adoption_Security_Report.pdf

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. arXiv:1805.03677 [Cs]. <http://arxiv.org/abs/1805.03677>

Home Office. (2018). Biometrics Strategy Better public services Maintaining public trust. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/720850/Home_Office_Biometrics_Strategy_-_2018-06-28.pdf

Home Office. (2019). Fourth report on statistics being collected under the exit checks programme. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/826381/fourth-report-on-statistics-being-collected-under-the-exit-checks.pdf

Griffiths, H. (2016). IoT Adoption Among Cities in the UK (27). IoTUK. https://iotuk.org.uk/wp-content/uploads/2016/08/IoT_Adoption_Security_Report.pdf

P. Hacker, A. Engel, and M. Mauer “Regulating ChatGPT and other Large Generative AI Models”, In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, USA, 2023. <https://doi.org/10.1145/3593013.3594067>

Leslie, D., Holmes, L., Hitrova, C., & Ott, E. (2020). Ethics of machine learning in children’s social care. Zenodo. <https://zenodo.org/record/3676569>

Leslie, D., Burr, C., Aitken, M., Cowls, J., Katell, M., & Briggs, M. (2021). Artificial intelligence, human rights, democracy, and the rule of law: A primer. SSRN Electronic Journal. <https://doi.org/10.2139/ssrn.3817999>

Leslie, D., Katell, M., Aitken, M., Singh, J., Briggs, M., Powell, R., Rincón, C., Perini, A., Jayadeva, S., & Burr, C. (2022). Data Justice in Practice: A Guide for Developers. Arxiv. <https://doi.org/10.48550/ARXIV.2205.01037>

Local Government Association. (n.d.) Behavioural insights: resources and best practice. <https://www.local.gov.uk/our-support/behavioural-insights/behavioural-insights-resources-and-best-practice>

Ministry of Defence (2021). Defence in a competitive age (CP411). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/974661/CP411_-_Defence_Command_Plan.pdf

Oswald, M., Grace, J., Urwin, S., & Barnes, G. C. (2018). Algorithmic risk assessment policing models: lessons from the Durham HART model and ‘Experimental’ proportionality. Information & communications technology law, 27(2), 223-250. <http://shura.shu.ac.uk/17462/>

Privacy International. (2021). Digital stop and search: how the UK police can secretly download everything from your mobile phone. <https://privacyinternational.org/sites/default/files/2018-03/Digital%20Stop%20and%20Search%20Report.pdf>

Rhodes, A. (2020). Digitalisation of energy: An Energy Futures Lab briefing paper. Energy Futures Lab. <https://spiral.imperial.ac.uk/handle/10044/1/78885>

Rigano, C. (2019). Using artificial intelligence to address criminal justice needs. National Institute of Justice Journal, 280(1-10), 17. <https://nij.ojp.gov/topics/articles/using-artificial-intelligence-address-criminal-justice-needs>

Symons, T. (2016). Datavores of Local Government. Nesta. <https://www.nesta.org.uk/report/datavores-of-local-government/>

Soomro, S., Miraz, M. H., Prasanth, A., & Abdullah, M. (2018). Artificial intelligence enabled IoT: traffic congestion reduction in smart cities. IET Conference Proceedings, IET Digital Library. <https://digital-library.theiet.org/content/conferences/10.1049/cp.2018.1381>

Tahayori, B., Chini-Foroush, N., & Akhlaghi, H. (2021). Advanced natural language processing technique to predict patient disposition based on emergency triage notes. Emergency Medicine Australasia, 33(3), 480-484. <https://onlinelibrary.wiley.com/doi/10.1111/1742-6723.13656>

Ubaldi, B., Le Fevre, E. M., Petrucci, E., Marchionni, P., Biancalana, C., Hiltunen, N., Intravaia, D. M., & Yang, C. (2019). State of the art in the use of emerging technologies in the public sector (31). OECD Publishing. https://www.oecd-ilibrary.org/governance/state-of-the-art-in-the-use-of-emerging-technologies-in-the-public-sector_932780bc-en

End Notes

¹ This method is adapted from the Turing publication: Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. <https://doi.org/10.5281/zenodo.3240529>

² The first four *Public Sector Guidance* workbooks were launched on November 2nd, 2023. The remaining four workbooks will be published in early 2024.

³ See our glossary for more descriptive information about the capabilities of AI. Also, [https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641547/EPRS_STU\(2020\)641547_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641547/EPRS_STU(2020)641547_EN.pdf)

⁴ The SSAFE-D Principles were developed independently but are similar to AI principles adopted elsewhere, including the OECD's four principles on artificial intelligence. See: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>. The OECD principles are echoed in the OAI's 2023 white paper on AI regulation: <https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper>

⁵ Note that for the purposes of this guidance, we refer to the process of either building or acquiring AI technologies as a *project*, to encompass different scenarios for adopting and implementing AI solutions. By labelling each AI implementation as a project, we hope to signal that AI solutions should be considered and evaluated individually to ensure it is the subject of sufficient ethical deliberation.

⁶ Leslie et al., 2022.

⁷ Leslie, 2019

⁸ OECD, 2019

⁹ Goyal, et al., 2023.

¹⁰ Burke, A., 2020.

¹¹ Leslie, D., 2020

¹² Tahayori, 2021.

¹³ Baker et al., 2019.

¹⁴ Dencik et al., 2018.

¹⁵ Symons, T. (2016).

¹⁶ Symons, T. (2016).

¹⁷ Griffiths, H. (2016).

¹⁸ Local Government Association. (n.d.)

¹⁹ Urban Intelligence, 2021.

²⁰ Rhodes, A. (2020).

²¹ Symons, T. (2016).

²² Soomro et al. (2018).

²³ Ubaldi et al. (2019).

²⁴ Berryhill et al. (2019).

²⁵ Ubaldi et al. (2019).

²⁶ Ubaldi et al. (2019)

²⁷ Burke, A. (2020).

²⁸ Ministry of Defence (2021).

²⁹ Oswald et al. (2018)

³⁰ Rigano, C. (2019).

³¹ Privacy International. (2021).

³² Burke, A. (2020).

³³ Home Office. (2018).

³⁴ Home Office. (2019).

³⁵ Cambridge Consultants. (2019).

³⁶ Department for Digital, Culture, Media, and Sport & Home Office. (2020).

³⁷ Central Digital and Data Office. (2020).

³⁸ Ubaldi et al. (2019)

³⁹ Bennett, S., & Cutler, N. (2019, October 28).

⁴⁰ See [Appendix E: Procurement Ethics](#) for a support in decision-making concerning the procurement of third party models, services, or system components.

⁴¹ (Gebu et al., 2020)

⁴² (Holland et al., 2018)

⁴³ The precautionary principle is a concept adapted from environmental management that has since been adapted to science, technology, and other areas of risky human activity. A definition employed by UNESCO's World Commission on the Ethics of Scientific Knowledge and Technology reads: 'When human activities may lead to morally unacceptable harm that is scientifically plausible but uncertain, actions shall be taken to avoid or diminish that harm.' For additional discussion of the precautionary principle in policy guidance, see [https://www.europarl.europa.eu/RegData/etudes/IDAN/2015/573876/EPRS_IDA\(2015\)573876_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/IDAN/2015/573876/EPRS_IDA(2015)573876_EN.pdf)

⁴⁴ Note: at the time of writing, the CDDO is in the process of developing updated and comprehensive guidance regarding GenAI systems. Current guidance can be found here: <https://www.gov.uk/government/publications/guidance-to-civil-servants-on-use-of-generative-ai/guidance-to-civil-servants-on-use-of-generative-ai>

⁴⁵ Hacker et al., 2023.

⁴⁶ For a taxonomy of risks associated with large language models such as ChatGPT, see reference [8].

⁴⁷ Leslie, 2019

⁴⁸ OECD, 2019

⁴⁹ Goyal et al. (2023).

⁵⁰ <https://publications.parliament.uk/pa/cm5803/cmselect/cmcumeds/1643/report.html>

⁵¹ See: <https://www.dataprotection.ie/en/organisations/know-your-obligations/data-protection-impact-assessments#what-is-a-data-protection-impact-assessment>

⁵² If an AI system or service or a data set is supplied in-part or entirely by a vendor, of the system involves transmitting data to third-party systems, use of the Procurement Ethics tool is recommended.