AI Ethics and Governance in Practice Programme

Responsible Data Stewardship in Practice

Facilitator Workbook

Annotated to support facilitators in delivering the accompanying activities.

The Alan Turing Institute

Acknowledgements

This workbook was written by David Leslie, Cami Rincón, Morgan Briggs, Antonella Maia Perini, Smera Jayadeva, Ann Borda, SJ Bennett, Christopher Burr, Mhairi Aitken, Sabeehah Mahomed, Janis Wong, Youmna Hashem, and Claudia Fischer.

The creation of this workbook would not have been possible without the support and efforts of various partners and collaborators. As ever, all members of our brilliant team of researchers in the Ethics Theme of the Public Policy Programme at The Alan Turing Institute have been crucial and inimitable supports of this project from its inception several years ago, as have our Public Policy Programme Co-Directors, Helen Margetts and Cosmina Dorobantu. We are deeply thankful to Conor Rigby, who led the design of this workbook and provided extraordinary feedback across its iterations. We also want to acknowledge Johnny Lighthands, who created various illustrations for this document, and Alex Krook and John Gilbert, whose input and insights helped get the workbook over the finish line. Special thanks must be given to the Digital Office for Scottish Local Government and Sara Bernardini, Jennifer Ding, Isabel Fenton, and Robert Blackwell (The Alan Turing Institute) for helping us test the activities and review the content included in this workbook. Lastly, we want to thank Humairaa Mahomed (Tech Matters Organisation) for helping with the co-design of one of our use cases, and Bastian Greshake Tzovaras and Jonathan Gilbert (The Alan Turing Institute) for their timely feedback and meticulous peer review, which greatly enriched this document.

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

Cite this work as: Leslie, D., Rincón, C., Briggs, M., Perini, A., Jayadeva, S., Borda, A., Bennett, SJ., Burr, C., Aitken, M., Mahomed, S., Wong, J., Hashem, Y., and Fischer, C. (2024). *Responsible Data Stewardship in Practice.* The Alan Turing Institute.

Contents

About the Workbook Series

Key Concepts

Activities

Further Readings

About the AI Ethics and Governance in Practice Workbook Series

Who We Are

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

Origins of the Workbook Series

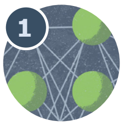
In 2019, The Alan Turing Institute's Public Policy Programme, in collaboration with the UK's Office for Artificial Intelligence and the Government Digital Service, published the UK Government's official Public Sector Guidance on AI Ethics and Safety. This document provides end-to-end guidance on how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

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

About the Workbooks

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

Programme Curriculum: AI Ethics and Governance in Practice Workbook Series

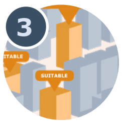


AI Ethics and Governance in Practice: An Introduction

Multiple Domains

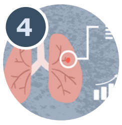
AI Sustainability in Practice Part One

AI in Urban Planning



AI Sustainability in Practice Part Two

AI in Urban Planning



AI Fairness in Practice

AI in Healthcare

Responsible Data Stewardship in Practice

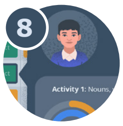
AI in Policing and Criminal Justice

AI Safety in Practice

AI in Transport

AI Explainability in Practice

AI in Social Care



AI Accountability in Practice

AI in Education

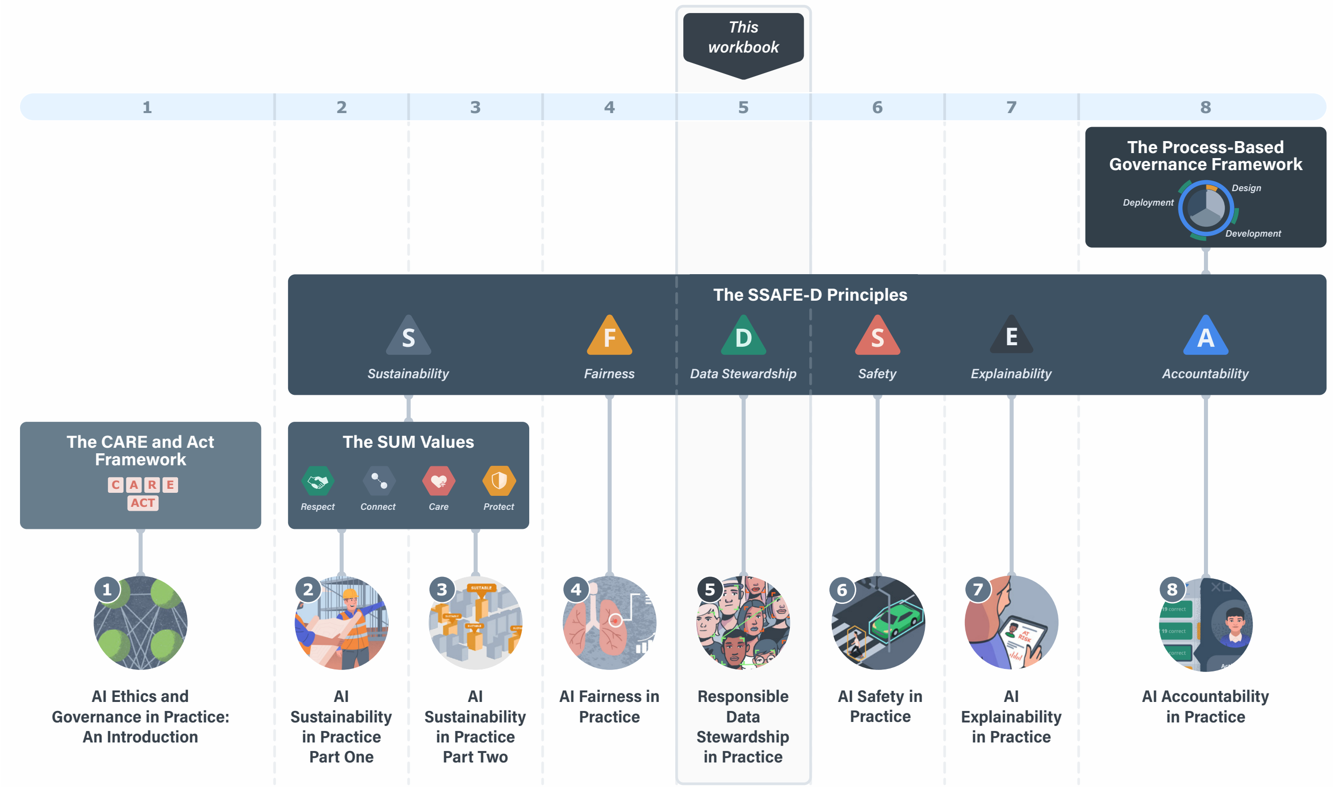
Explore the full curriculum and additional resources on the AI Ethics and Governance in Practice Platform at aiethics.turing.ac.uk.

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

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

Programme Roadmap

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


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

Intended Audience

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

There are two versions of each workbook:

- **Facilitator Workbooks** (such as this document) are annotated with additional guidance and resources for preparing and facilitating training workshops.

- **Participant Workbooks** are intended for workshop participants to engage with in preparation for, and during, workshops.

Introduction to This Workbook

This workbook aims to provide resources and training which help you and your team to ethically steward the data you access and utilise by proactively initiating and facilitating responsible data practices. You will learn how to use these tools and how they may be relevant at different stages of the project lifecycle. The tools, approaches, and policies introduced should be discussed with your core team and your stakeholders, and should be clearly documented.

Data is essential in developing AI models and systems, forming the core information on which they are trained, and, as such, shaping their knowledge base and epistemic (knowledge-contributing) capacity. For this reason, responsible data stewardship is crucial for developing ethical and responsible AI. This workbook is divided into two sections, Key Concepts and Activities:

Key Concepts Section

This section provides content for workshop participants and facilitators to engage with prior to attending each workshop. It provides definitions of key terms, an overview of the key components of responsible data stewardship, and introduces tools and practical guidance to ethically steward data in an AI project. Topics discussed include:

Part One: Introduction to Responsible Data Stewardship

1. Introduction to Responsible Data Stewardship
2. The Data Lifecycle
3. Key Components of Responsible Data Stewardship

Part Two: Putting Responsible Data Stewardship into Practice

1. Data Factsheet
2. Data Factsheets Across the Workflow

Activities Section

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

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

Exploring Data Use in Policing

Gain familiarity with the data lifecycle and overarching considerations for Responsible Data Stewardship.

Exploring Components of Responsible Data Stewardship

Consider the benefits and limitations of incorporating the different components of Responsible Data Stewardship as harm-mitigation mechanisms in AI projects.

Creating Data Factsheets

Groups will answer questions within a Data Factsheet pertaining to the case studies provided in the previous activity.

Note for Facilitators

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

Responsible Data Stewardship in Practice

Key Concepts

Part One: Introduction to Responsible Data Stewardship

As introduced in the AI Ethics and Governance in Practice: An Introduction workbook, data represents facets of phenomena seen in the world, as recorded through observation or measurement. In the context of AI, data is digitally recorded. Individual data points comprise part of a larger dataset, curated and stored together in service of a particular aim or domain focus. The type of data points included can range in type (e.g. numbers, words, images), and can either be stored in a particular format for a specified purpose (structured), or can be general, varied, and not predefined by a specific data format (unstructured).[1] [2]

Data plays a key role at every stage of the project lifecycle. It can be used to train, validate, and test an AI model, thereby shaping the type of AI system produced. In addition to utilising existing datasets, AI models may be used to generate **synthetic data** which may subsequently be used as training data. For these reasons, initiating and facilitating responsible data practices is essential to make sure that projects are ethically justifiable. Responsible Data Stewardship is equally guided by principles prioritising data integrity, data quality, and engagement, alongside data protection policies as highlighted in Explaining decisions made with AI, a co-badged guidance by the Information Commissioner's Office (ICO) and The Alan Turing Institute.

KEY CONCEPT

Synthetic Data

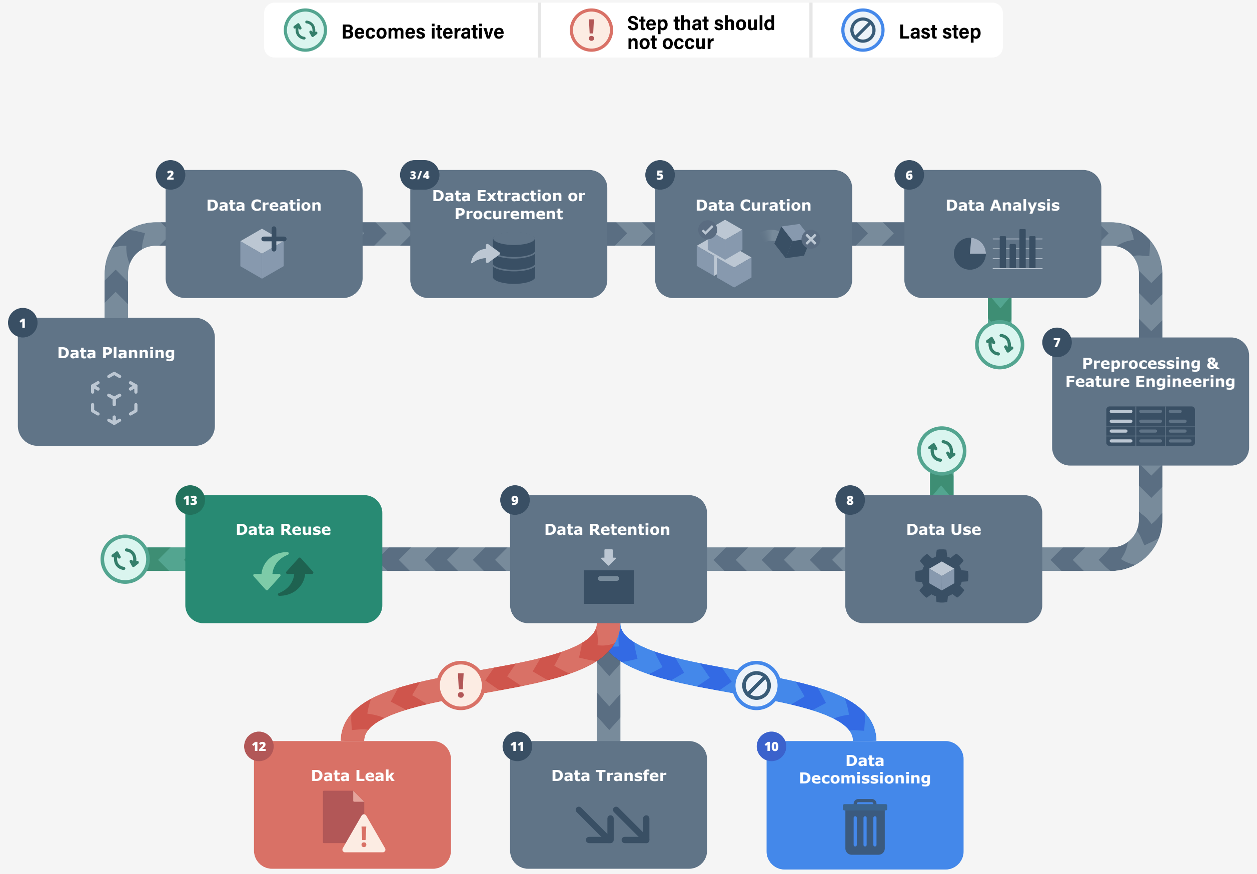
Artificial data that is generated from real data using an algorithmic model that reproduces the patterns and statistical properties of that data. Synthetic data can synthesise all variables of the original dataset (fully synthetic data), or selectively mimic certain variables (partially synthetic data).[3]

The Data Lifecycle

The data lifecycle refers to the numerous phases of data as it is collected, curated, analysed, used, decommissioned, and iteratively reviewed within the larger AI project lifecycle to ensure accurate, secure, and robust performance of an AI/ML system. Integral to every stage of the data lifecycle is responsible data stewardship.

In this workbook, teams can learn about the nuances of specific stages while understanding the relationship between stages. Importantly, teams should be aware of the iterative nature of the lifecycle wherein stages and processes within the lifecycle are revisited and refined over time through technical and non-technical means to improve the model and its outputs.[4] For instance, insights from data analysis may prompt an AI project team to further collect or refine data. Similarly, changes in regulatory requirements may necessitate the AI project team to revisit one or more stages of the data lifecycle.

Stages of the Data Lifecycle



1 **Data Planning**

Data planning refers to preliminary strategic activities that plan how data will be used to address the AI project’s aims, objectives, and processes. It involves assessing data sources, availability, and accessibility of the data needed to develop the AI system, and ensures that all aspects of data management are considered at the start of the project.[5]

2 **Data Creation**

Data creation refers to the purposeful identification of new or existing data and outlining how such data can be collected, extracted, or processed to create a useful set of data.

3 **Data Extraction**

Data extraction refers to the process of collecting different forms of data (i.e. structured, semi-structured, and unstructured) from a variety of sources. The level to which data is structured determines the need for the transformation and integration of data to enable exploration and analysis.

4 **Data Procurement**

Data procurement refers to the process of acquiring data from a variety of sources, including data from external sources, third-party vendors, and data extracted and/or collected by others. Procurement can be a complex and time-consuming process and requires coordination with other processes in the data lifecycle, such as data planning, extraction and curation, including the identification of specific datasets that are required for procurement. This process may involve legal agreements to obtain already existing datasets.

5 **Data Curation**

Data curation refers to the process of selecting, cleaning, and organising data to make it suitable for use in AI/ML applications. These measures ensure that data are of high quality and can be reused in the future.

6 **Data Analysis**

Data analysis involves iteratively exploring the makeup of the data through visualisation and summary statistics. Some questions at this stage may include: are there missing data (incomplete data)? Are there outliers (unexpected data), unbalanced classes (imbalanced data), or correlations?

7 **Preprocessing & Feature Engineering**

Preprocessing & feature engineering is the process of cleaning, normalising, and refactoring 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.

8 **Data Use**

Data is used to iteratively train and improve AI/ML models and systems to ensure that their outputs are as accurate as possible. This step spans multiple stages in the AI lifecycle; from Model Selection & Training to System Use & Monitoring.

9 Data Retention

Data retention refers to the process of retaining and maintaining data for current and future use. The process includes data that is being archived and managed as well as data from new data sources or model outputs that will be added to existing datasets.

Last step

10 Data Decommissioning

Data decommissioning refers to the process of permanently and securely removing data from a system, databases, and/or other data archival processes such that it is no longer part of the data lifecycle and will not be used in any of the data models or systems. Decisions on data decommissioning and retention are usually taken simultaneously.

11 Data Transfer

Data transfer refers to the exchange of data between systems or organisations. Data transfers should include formal socio-technical processes to ensure that only the necessary data is sent to the correct recipient. Encryption and specific data transfer tools may be used to support the safe transfer of data from an internal system to an external one.

Step that should not occur

12 Data Leak

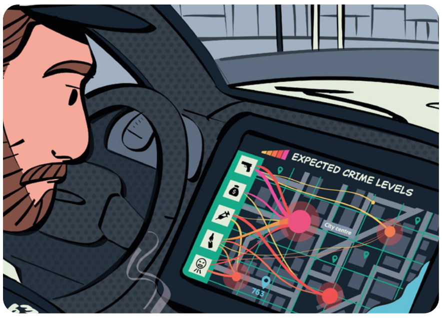
Data leak, or data breaches, refer to data that has been intentionally or unintentionally disclosed to an unauthorised third party. Where a data leak occurs steps should be taken immediately to identify the cause of the leak, what data has been leaked, and who might be affected. All data leaks must be documented and steps must be taken to ensure they do not happen again. If personal data is leaked this must be reported to the Information Commissioner's Office (ICO) within 72 hours of AI project teams becoming aware of it.

13 Data Reuse

Data reuse refers to the use of existing data (created, collected, or procured) for purposes beyond its original intent. AI project teams may decide to extract additional value of the data used in an AI project by repurposing it for a different project, application, or research.

Case Study

Example of Data Lifecycle: Predictive Risk Modelling[6]



An AI/ML system is used to support a local police force in adopting effective planning and intervention strategies which ensure public safety. This system predicts areas with high crime density (‘crime hotspots’), and generates a digital map that informs where and when police forces are deployed for patrol. The stages below show the unique journey of data used in this AI/ML system.

Stages of Predictive Risk Modelling Data Lifecycle

Data Creation

Data is extracted from numerous sources including:

- crime category, time, date, resolution, and location;
- geospatial image data created by satellites through time stamped photographs and video recordings of the earth; and
- government maintained records of public and private land, including information about land parcel use, coordinates, and ownership, creating land registry databases.

Data Procurement

In the development of this project, geospatial data was procured from open geospatial databases pertaining to the borough where the model would be deployed. Historic crime data was procured from the local police database.

Data Curation

The satellite imagery, land registry databases, and crime databases were combined to create a comprehensive geospatial (location) database, organised by features suitable for modelling.

Data Analysis

The collected data was checked for duplicates, gaps, or other inconsistencies.

Preprocessing & Feature Engineering

A spatial clustering technique was implemented to identify areas with historically high crime incidents.

Data Use

Once preprocessed through the spatial clustering technique, the geospatial database was then used to train, test, and validate a neural network model that predicts where and when crime will occur within the geographic area. The model illustrates outputs in a digital map highlighting areas where crime is predicted to occur.

Data Retention

The model's outputs are used to optimise the model. This is stored and retained alongside existing data.

Data Use

Additional data is created by the reporting of crimes following the deployment of the model. This data is used to further optimise the model.

KEY CONCEPT

Data Lineage

Data lineage refers to the tracking of specific data's unique journey from its origin through all the stages of the lifecycle, including how it is used, understood, updated, transformed, and shared. This process involves documenting where data comes from and where it is intended to be used in the future. The purpose of data lineage is to clearly map how data is used, or not used, for greater clarity on where and how potential errors may have occurred within the data provenance process.

This section draws on previous research and publications from the team.[7] [8] [9] [10]

A Closer Look at Data Stewardship

Data stewardship refers to the governance of the data lifecycle, from scoping out the role data plays in a project, to ensuring it is safely handled after the project is under way.

Data stewards (or data custodians) are individuals responsible and accountable for the management and care of assigned data holdings.[11] Data stewards may:

- facilitate collaboration to unlock the value of data;
- protect actors from harms caused by data sharing; and
- monitor users to ensure that their data use is appropriate and can generate data insights.[12]

In the last few years, we have witnessed the emergence of new data stewardship frameworks, including data trusts, data foundations, and data cooperatives. These have been designed to protect data subjects and facilitate the co-creation of data protection solutions through the involvement of data subjects and other stakeholders.[13] [14] [15]

Data stewardship allows for the societal value of data (i.e. the recognition of its intrinsic value and its potential to achieve societal objectives) to be harnessed in an inclusive, trustworthy, transparent, and responsible manner. It is key to protecting the data rights of individuals and communities, and aims to be participatory in aspects such as data collection, management, and use.[16] [17] Essentially, data stewardship raises the question: what data practices may enable the use of data for the public benefit or social good?

Data Governance

Data governance is a frequently used term to describe some of the processes outlined as part of data stewardship. Although there is no consensus on what data governance encompasses, the concept focuses on establishing policies surrounding data management and stewardship by taking a high-level, bird’s eye view of the data lifecycle. Data governance refers to everything designed to inform the extent of confidence in data management, data uses, and the technologies derived from these applications of data.[18] In contrast, data stewardship (and this workbook) focus on the practical implementation of specific tools and methodologies to optimally manage and steward the responsible use of data. Data governance may be understood as the policies and steps implemented to govern access, management, sharing, and uses of data, and provides the ability to better manage each stage of the data lifecycle.

Data Management

Another frequently used term is data management. This encapsulates the planning, development, implementation, and administration of systems for the acquisition, storage, security, retrieval, dissemination, archiving, and disposal of such data. In other words, it encompasses the daily procedures and technologies employed for the efficient and effective management of data. It includes tasks such as the development of data policy for metadata compilation.[19] More generally, data management refers to managing data where data is considered a resource of value.

Key Components of Responsible Data Stewardship

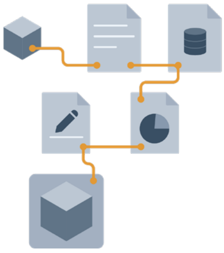
Responsible Data Stewardship is guided by principles which prioritise preservation of data integrity, data quality, and engagement, alongside data protection policies. It also requires consideration of data equity (see Data Equity on page 21).

Consideration for AI in Policing

Throughout this workbook, we will flag specific considerations related to responsible data stewardship for AI in policing. These considerations are included as call-out-boxes.

Principle 1

Data Integrity



The principle of Data Integrity refers to 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 the end of a project's lifecycle.[20] [21]

Characteristics of Data Integrity[24]

Attributable
Data should clearly demonstrate who observed and recorded it, when it was observed and recorded, and who it is about.

Consistent, Legible, and Accurate

Data should be easy to understand, recorded permanently, and original sources should be preserved. Data should be free from errors and conform with the data integrety protocol. Consistency includes ensuring data is chronological (e.g. has a date and time stamp that is in the expected sequence).

Complete

All recorded data requires an audit trail to show that nothing has been deleted or lost.[22]

Contemporaneous

Data should be recorded as it was observed, and at the time it was used.[23]

Data Traceability and Auditability

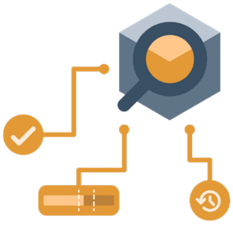
Any changes or revisions to the dataset (e.g. additions, augmentations, normalisation) that occur after the original collection should be clearly traceable and well-documented to support any auditing.

Consideration for AI in Policing

If geospatial data is being used, where, when, and how this data was gathered and placed within the dataset should be clearly documented in as much granular detail as possible. This is particularly important given how such details can affect policing cases.[25]

Principle 2

Data Quality[26]



The principle of Data Quality captures static properties of data, such as whether they 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.

Characteristics of Data Quality

Representativeness

The distribution of features that are included in the dataset, and the number of samples within each class, fit the underlying distribution that exists in the overall population (mentioned in Workbook 4: AI Fairness in Practice).

Fit-for Purpose and Sufficiency

Assessing whether the data available is sufficient for the intended purpose of the project, considering factors such as the use case, domain, and function of the system. Fitness-for-purpose and sufficiency should address sample size, representativeness, and availability of features relevant to problem (mentioned in Workbook 4: AI Fairness in Practice).

Source Integrity and Measurement Accuracy

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, with these biases replicated in the output. In addition to identifying potential for bias in data collection and processing, it is crucial to identify biases in system outputs, and to identify whether these cause harm, as well as ensuring that the data sample has optimal source integrity. This involves analysis of the contexts in which the data has been gathered, and how these might introduce inequitable ‘ground-truths’, and ensuring that the data gathering processes involved suitable, reliable, and impartial sources of measurement and sound methods of collection (mentioned in Workbook 4: AI Fairness in Practice).

Timeliness and Recency

If datasets include outdated data, then changes in the underlying data distribution may adversely affect the generalisability of the trained model (mentioned in Workbook 4: AI Fairness in Practice).

Relevance, Appropriateness and Domain Knowledge

The understanding and utilisation of the most appropriate sources and types of data are crucial for building a robust and unbiased AI system. Solid domain knowledge - of the underlying population distribution and of the goal of the project - is instrumental for choosing the optimal measurement inputs. Domain experts should collaborate closely with the technical team to determine of the optimally appropriate categories and sources of measurement (mentioned in Workbook 4: AI Fairness in Practice).

Consideration for AI in Policing

How Data Shapes AI Models

Data quality requires evaluating whether datasets adequately and accurately reflect experiences of different groups or communities - particularly communities impacted by policing decisions and/or crime. It is essential to identify the extent to which historic prejudices or biases have shaped the data collected, and how these biases might in turn affect decisions that are made from this data. Engaging with domain experts and impacted communities is important to address potential biases or omissions in the data, to identify additional/alternative sources of data that can strengthen the model, and to ensure that analysis and model development are informed by understanding the context and lived experience of impacted communities.

Data Equity[27]

Data reflects the contexts and circumstances in which it is collected. As has been described in the AI Fairness in Practice workbook, potential biases can enter into the datasets, and scaffold into the AI lifecycle. In addition, active decisions about data practices (e.g. what data to use and whose perspectives to include) may also impact how data on marginalised individuals, groups, or communities either addresses disparities, or perpetuates harms and injustices. These issues have direct bearing on your obligation to uphold the Public Sector Equality Duty.

Data equity centres on these concerns and provides a framework for your project team to:[28]

1. **Identify and critically reflect on the ways in which power, bias, and discrimination can enter the data lifecycle, through an equity and social justice lens.** For instance, the choice to acquire or use data for policing may impact the lives of historically marginalised people, raising questions such as: does a particular use or appropriation of data enable or disable inequity and discrimination? Does it preserve or combat harmful relations of power?

2. **Take actions to repair data practices that embed power imbalances and forms of inequity and discrimination.** For instance, in the context of data for policing, project teams may respond to historical injustices by acknowledging and accounting for the claims of groups or communities that have been historically marginalised by policing practices and structures.

Classes of Data Equity[29]

- **Representation equity** refers to ensuring that all groups in the target population are authentically reflected in the training of AI and in the final models. For instance, with police and crime data, previous case studies have highlighted that this data is shaped by historic biases and prejudices within policing, including discriminatory attitudes which affect rates of arrest as well as where/which crimes are reported and investigated.

- **Feature equity** seeks to make available the relevant attributes or variables to accurately represent individuals, groups, or communities in the data, and and to enable analysis that does not reinforce or amplify inequities. A data equity framework will, for instance, include variables of race and gender and conviction data to discover systemic biases and correct them.

- **Access equity** aims to ensure equitable access to data and models across domains, expertise, and roles. It also addresses issues of AI literacy disparities and the digital divide, which often disproportionally impact already marginalised groups or people within a given target population.
- **Outcome equity** aims to ensure that results of data collection and use are impartial and fair, and that no unanticipated consequences or harms outside the control of the system affect individuals, groups, or communities.

Putting Data Equity into Practice

Because data practices and choices about such practices occur at different stages of the data and AI lifecycle, data equity considerations are required from the start. As such, processes of data creation and data analysis must include reflection on the effects of historic biases within datasets, but also which measures are needed to address these biases. To incorporate data equity, AI project teams should:[30]

- Recognise the factors that affect individuals, groups, and communities' ability to thrive.
- Promote stakeholder engagement to understand the extent to which the data adequately and accurately represents lived experiences of impacted communities. In doing so, this may also highlight the need to collect additional data, or to supplement existing datasets with additional forms of knowledge (including community-reported data on crime and policing).
- Engage in collaborations with agencies and communities to design a shared data development agenda, which includes plans for ensuring data quality and access.
- Use language and actions that embed equity at the centre.
- Promote data practices where communities 'govern the collection, ownership, dissemination, and application of their own data' (p. 13).

KEY CONCEPT

Data Justice[31]

Understanding 'data justice' involves exploring the impacts of data practices and data-driven innovation through a social justice lens. Data justice is a guiding framework that responds to, and enables the transformation of, existing power asymmetries and inequitable or discriminatory social structures.

Principle 3

Data Protection and Privacy[33]

The principle of Data Protection and Privacy reflects ongoing developments and priorities set out in relevant data legislation and regulation, specifically as and when they pertain to fundamental rights and freedoms, democracy, and the rule of law. One such example is the right for data subjects to request inaccurate personal data to be rectified or erased. We describe below general characteristics of Data Protection and Privacy, but the reader is referred to guidance issued by the Information Commissioner's Office for more in depth explanations of compliance requirements for the Data Protection Law and Data Protection and Privacy best practices. We have provided links to helpful ICO resources at the end of the Characteristics of Data Protection and Privacy on the following page.

Consideration for AI in Policing

Given the sensitivity of data used for policing, it is important that data protection regulations and guidelines are strictly followed. If the data collected is inappropriately accessed, shared, or leaked, there could be grave consequences for those being investigated as part of the case. More generally, it may also result in loss of trust in the police force.

Characteristics of Data Protection and Privacy

Consent (or legitimate basis) for Processing

The Council of Europe published the Convention of Protection of Individuals with regard to Automatic Processing of Personal Data ('Convention 108') which represented the first legally binding international treaty for privacy and data protection. Convention 108 is aimed at both protecting individuals against abuses which may accompany the collection and processing of personal data, and seeking to regulate the cross-border flow of personal data. The Convention lays down the foundational concepts represented in subsequent data protection and privacy laws, such as the General Data Protection Regulation (Regulation (EU) 2016/679) (GDPR). The original Convention document has since been updated in 2018. A key passage relevant to this workbook is excerpted below from the 2018 publication:

> Each Party shall provide that data processing can be carried out on the basis of the free, specific, informed, and unambiguous consent of the data subject or of

> some other legitimate basis laid down by law. The data subject must be informed of risks that could arise in the absence of appropriate safeguards. Such consent must represent the free expression of an intentional choice, given either by a statement (which can be written, including by electronic means, or oral) or by a clear affirmative action and which clearly indicates in this specific context the acceptance of the proposed processing of personal data. Mere silence, inactivity or pre-validated forms or boxes should not, therefore, constitute consent. No undue influence or pressure (which can be of an economic or other nature) whether direct or indirect, may be exercised on the data subject and consent should not be regarded as freely given where the data subject has no genuine or free choice or is unable to refuse or withdraw consent without prejudice. The data subject has the right to withdraw the consent they gave at any time (which is to be distinguished from the separate right to object to processing). Full consideration should be given to the potential impact of any personal data processing on the fundamental rights and freedoms of the individuals involved.

Data Security

Data security requires the data controller and - where applicable - the data processor to take appropriate security measures against risks such as accidental or unauthorised access to, destruction, loss, use, modification, or disclosure of personal data. If the data controller becomes aware of data breaches which may seriously interfere with the rights and fundamental freedoms of data subjects, the component supervisory authority -like the ICO - must be notified without delay.

Data Minimisation

Any AI project team processing personal data must ensure that the data is adequate (sufficient to properly fulfil the stated purpose), relevant (has a rational link to that purpose), and limited to what is necessary (do not hold more data than needed for that purpose).[32]

Transparency

The transparency of AI systems can refer to several features, from transparency around their inner workings and behaviours, to transparency around the systems and processes that support them. An AI system is transparent when it is possible to determine how it was designed, developed, and deployed. This can include, among other things, a record of the data that was used to train the system, or the parameters of the system that transforms the input (e.g. an image) into an output (e.g. a description of the objects in the image). However, it can also refer to wider processes, such as transparency around whether there are legal barriers that prevent individuals from accessing information that may be necessary to fully understand fully how the system functions (e.g. intellectual property restrictions).

Proportionality

Proportionality in a broad sense encompasses both the necessity and the appropriateness of a measure — that is, the extent to which there is a logical link between the measure and the (legitimate) objective pursued. The term proportionality is used as an evaluative notion, such as in the case of a data protection principle that states only personal data that is necessary and appropriate for the purposes of the task should be collected.

Purpose Limitation

Personal data must adhere to the original purpose and limited to this use, unless a new purpose is either compatible with the original purpose, and additional consent is then received, or there is an obligation or function set out in law.

Accountability

Appropriate measures and records must be in place to demonstrate compliance and responsibility for how data has been processed in alignment with the other principles.

Lawfulness, Fairness, and Transparency

These three principles form part of the ‘lawful basis’ for the collection and use of personal data. Personal data must be used in a fair manner that is not unduly detrimental, unexpected, or misleading. Any processes in which data is used should not be in breach of any other laws, and teams must be clear, open, and honest with individuals about how their personal data is being used.

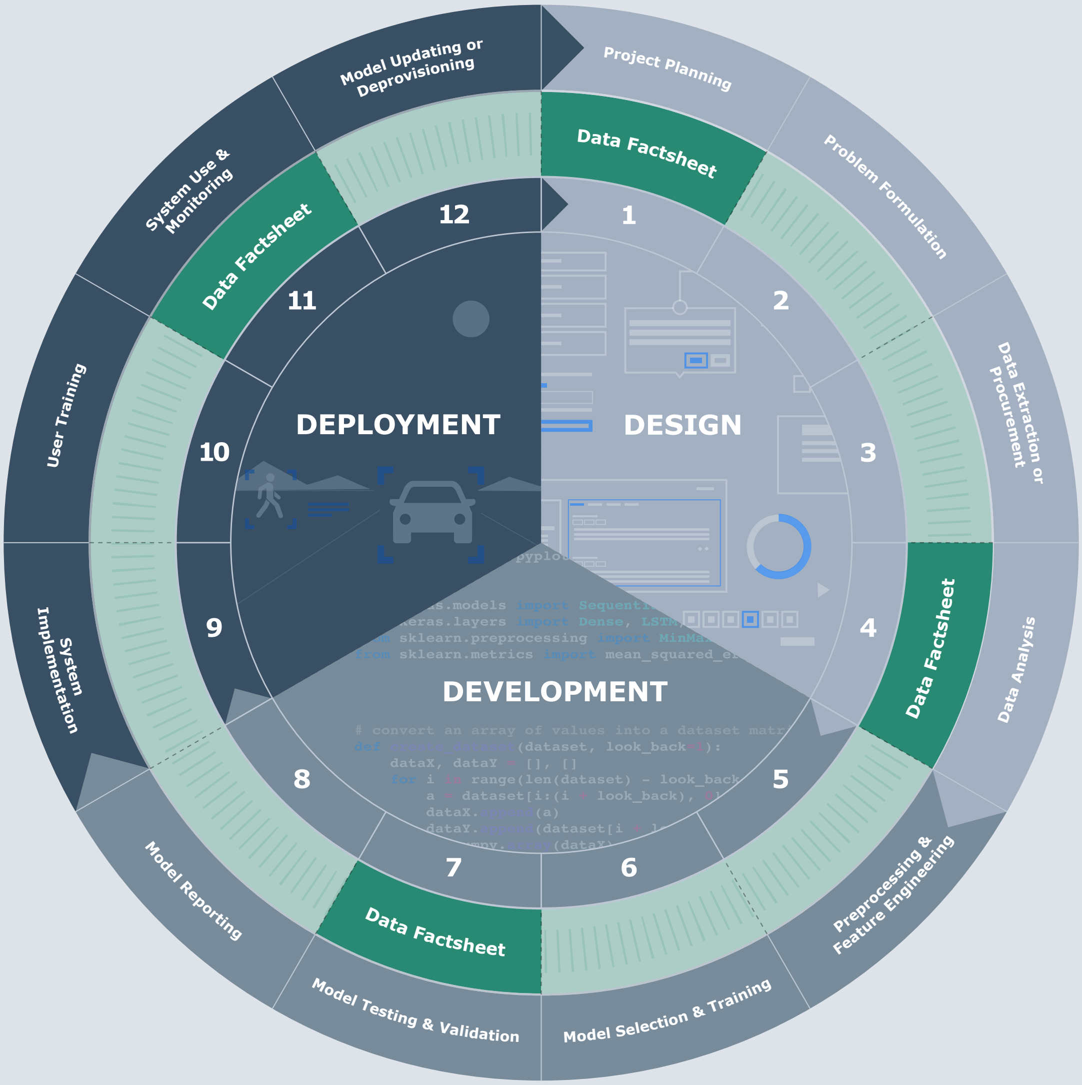
Respect for the Rights of Data Subjects

Respect for the rights of data subjects requires putting in place adequate mechanisms, or undertaking necessary actions so as to ensure that the rights of data subjects as defined under the Council of Europe’s Convention 108+ and the General Data Protection Regulation (GDPR) are upheld. Where necessary, this includes the responsible handling of sensitive data.

ICO Resources

- ICO guide to the data protection principles
- ICO training videos on data protection principles
- ICO guidance on Personal information - what is it?
- ICO training video on Personal information - what is it?
- ICO guidance on lawful basis
- ICO lawful basis interactive guidance tool

Part Two: Putting Responsible Data Stewardship into Practice



Data Factsheet

Facilitating Data Stewardship and Responsible Data Management involves considering your specific project use case to identify what and how data is to be collected, managed, analysed, and potentially shared. To ensure that this process is done with each of the components of Responsible Data Stewardship in mind, the data and datasets should be assessed through a Data Factsheet.

Data Factsheets facilitate the uptake of best practices for data integrity, quality, protection, equity, and privacy across the AI project workflow. Their purpose is to outline a comprehensive record of the data lineage of a project as well as qualitative input from team members about determinations made with regard to the different components of Responsible Data Stewardship (data integrity, quality, protection, and privacy). The Data Factsheet Template provided below helps establish data management plans across the project lifecycle. It facilitates the assessment of data management at each stage, given the importance of considering how data is to be used, applied, and shared.

This template also includes open-ended questions supporting the consideration of appropriate Data Risk Management (DRM) Frameworks. Existing DRM Frameworks are useful because they support the identification of potential issues in relation to the use of datasets as well as help mitigate potential harms. It is, however, important to note that once a framework is selected (if at all) its application should not be used in isolation but rather in conjunction with the process of maintaining a Data Factsheet. Based on these reflective, preliminary questions derived from the key themes in the DRM Frameworks, further considerations can be made regarding responsible data management and stewardship through the AI data lifecycle. While considering the data lifecycle, the DRM Framework section of the Data Factsheet can be referred back to as a reminder of the primary considerations in context of the data and dataset you and your team wish to steward and manage.

Data Factsheet Template for: Project Name

Assessing Data Stewardship Tools

a. Is the dataset dynamic or fixed?

b. Is the dataset proprietary or open/available to the public?

c. Does the dataset include identifiable, personal, and/or sensitive data?

d. Who is responsible for applying the Data Risk Management (DRM) framework?

e. Who is the intended audience of the DRM Framework?

f. Will the data or model output be publicly accessible? To what extent will the data, dataset, or model be shared in the future?

g. Is the application of a DRM Framework relevant to the project and problem identified, if relevant?

- If so, what risk management framework may me most suitable and why?

Project Planning and Management

1. Objectives

a. How will data minimisation and purpose limitation be ensured in this project?

2. Team Building and Management

a. Who will have direct access to the data and datasets? Is this the minimum number possible?

b. Will all team members undergo training in responsible data management prior to the start of the project?

c. What range of skillsets do the team have and are they representative of the objectives?

d. Are stakeholders who may be impacted part of the core team and/or will they be engaged/consulted at a later stage?

3. Data Sources

a. What data will be used to build this system?

b. What sorts of personal, and/or sensitive data is expected to be required?

c. Where will the data be sourced from?

d. Why were these datasets chosen?

e. What are the limitations of the datasets being used?

f. Will any data be excluded from the dataset?

g. How will data minimisation be ensured?

h. What potential biases might be present within the data?

i. What risks may be associated with potential biases in the data?

j. What actions will be taken to minimise/ address any biases present?

k. What measures will be taken to ensure that data are complete, up-to-date and accurate?

l. Will meta-data be stored alongside the datasets to aid and improve data analysis?

....

m. Will data subjects give consent for the data to be used for the intended purpose?

....

n. Where consent has not been given, will there be alternative governance arrangements in place that permit the use of the data for this purpose?

....

o. Will data subjects be able to opt-out?

....

p. Will there be an effective feedback loop in place between users of data and the collection of data, specifically if outsourced from third parties?

....

4. Data Transfer

a. What measures will be taken to ensure that data is transferred securely minimising risks of data leakage or security breaches?

....

b. Is the data transfer encrypted?

....

c. Will the data be transferred outside of the originating country and be subject to different regulations?

....

5. Data Analysis

a. What steps will be undertaken to minimise risks of identification of data subjects?

....

b. What steps will be undertaken to ensure data is accessed only in secure environments?

....

c. Will all analysts/researchers undergo adequate training in responsible data management?

....

d. What potential biases might affect or influence data analysis?

....

e. What steps have been taken to mitigate potential biases?

....

f. Will domain experts be consulted about the findings of exploratory data analysis to verify the relevance of the input variables?

....

g. Will the dataset be checked for: incomplete data, outdated data, duplicated data, errors in manual data entry, inconsistent formats, poor-quality metadata?

....

h. How will all errors been identified and logged?

....

i. How will data analysis be presented clearly in order to be understood by the intended audience as well as knowledge-specific stakeholders?

....

Design Phase (Problem Formulation)

1. Data Lineage

a. Will the data lineage for this project be documented, including:

- What data is collected?

  ....

- Where data is sourced?

  ....

- How data is used?

  ....

- How data is altered/processed/updated?

  ....

- Who has had access to data?

  ....

- If data is destroyed, and, if so, how?

  ....

- What outcomes have been produced from the data?

  ....

- If any new datasets are created?

  ....

b. If so, how will it be documented?

....

- What are the intended and possible unintended consequences of data use for this project?

  ....

2. Data Integrity

How are we planning to implement sufficient and transparently reported processes throughout the AI/project lifecycle to ensure that:

a. All data used in producing the system is attributable, consistent, complete, and contemporaneous?

....

b. Are all current and future data entries traceable to: i) an individual/source, ii) date and/or time of entry? Is this done manually or through a technical/automated means?

....

c. To ensure consistency, is the data chronological?

....

d. Have data entries been verified by a second person?

....

e. Have any data integrity weaknesses been identified?

....

f. If any, what corrective and preventative actions have been put in place? Is this in isolation or across all relevant activities and systems?

....

g. Has more than one copy of the data been stored securely and used to verify consistency throughout the project?

....

Considering answers to the previous questions, what actions will be taken to assure data integrity in the Problem Formulation, Data Extraction or Procurement, and Data Analysis phases of this project?

....

3. Data Quality

a. How are we planning to implement sufficient and transparently reported processes throughout the AI/ML project lifecycle to ensure that:

  - all data used in producing the system are accurate, reliable, relevant, appropriate, up-to-date, balanced, representative, and of adequate quantity and quality for the use case, domain, function, and purpose of the system?

b. Is there a process in place (e.g. an automated script) that automatically flags for human checking for any unexpected deviations, missing data, or unexpected formatting?

Considering answers to the previous questions, what actions will be taken to assure data quality in the Problem Formulation, Data Extraction or Procurement, and Data Analysis phases of this project?

4. Data Protection and Privacy

a. What measures will be taken to ensure that data is stored and accessed securely, minimising risks of data leakage or security breaches?

b. What steps will be taken to minimise risks of identification of data subjects?

c. Will any data be pseudonymised or anonymised?

d. Will data-protection-by-design principles be followed?

e. What technical measures will be implemented to ensure that data is protected (e.g. encryption and secure data transfer processes)?

f. How will data subjects find out more information about the data protection/privacy policy, their data protection rights, and how will these rights be exercised?

g. Will personal and/or sensitive data (including pseudonymised or anonymised data) been included/excluded?

- If included, what are the possible implications?
- If excluded, what are the possible implications?

h. [If public-facing] What user-centred design[34] will be included as part of the process?

i. [If public-facing] Will the use of dark patterns be excluded as part of the interface?

j. If the organisation has a data protection officer, have they been consulted?

k. Had a data protection impact assessment been created where necessary?

l. Has the purpose for data processing been clearly outlined and documented, and been made easily and clearly available (in plain language) to all individuals through privacy information?

m. Have more than one copy of the data been stored securely and used to verify consistency throughout the project?

n. Before data is copied or transferred outside of our secure environment, is there a multi-party approval system in place to minimise the risk of security leaks?

Considering answers to the previous questions, what actions will be taken to assure data protection and privacy in the Problem Formulation, Data Extraction or Procurement, and Data Analysis phases of this project?

Design Phase (Data Analysis)

1. Data Lineage

a. What data has been collected?

b. Where has data been sourced?

c. How has data been used?

d. How has data been altered/processed/updated?

e. Who has had access to data?

f. Has data been destroyed, and, if so, how?

g. What outcomes have been produced from the data?

h. Have any new datasets been created?

i. If a DRM Framework has been identified to be relevant, should this be completed alongside the questions and considerations listed within the lifecycle?

j. Has the design phase been transparent, particularly with regards to taking on user feedback and recommendations?

2. Data Integrity

a. Have any data/variables been transformed (e.g. changing text data to numeric data)?

b. How has data integrity been maintained and is this documented? (e.g. during data cleaning, data wrangling etc).

Considering answers to the previous questions, what actions will be taken to assure data integrity in the Development phase of this project?

3. Data Quality

a. Is the pre-processed dataset representative of the underlying population distribution?

....

b. Has the above been verified by domain experts and documented?

....

c. How are metrics and benchmarks documented?

....

Considering answers to the previous questions, what actions will be taken to assure data quality in the Development phase of this project?

....

4. Data Protection and Privacy

a. What measures have been taken to ensure that data is stored and accessed securely minimising risks of data leakage or security breaches?

....

b. What steps have been taken to minimise risks of identification of data subjects?

....

c. Has any data needed to be pseudonymised or anonymised been prepared?

....

d. Are data-protection-by-design[35] principles being followed?

....

e. What technical measures have been implemented to ensure that data is protected (e.g. encryption and secure data transfer processes)?

....

f. Is it clear how data subjects can find out more information about the data protection/privacy policy, their data protection rights, and how those rights can be exercised?

....

g. Has personal and/or sensitive data (including pseudonymized or anonymised data) been included/ excluded?

....

- If included, what are the possible implications?

....

- If excluded, what are the possible implications?

....

h. [If public-facing] What user-centred design has been included as part of the process?

....

i. [If public-facing] Have the use of dark patterns been excluded as part of the interface?

....

Considering answers to the previous questions, what actions will be taken to assure data protection and privacy in the Development phase of this project?

....

Development Phase → **Deployment Phase**

(Model Selection & Training → System Use & Monitoring)

1. Data Lineage

a. What data has been collected?

b. Where has data been sourced?

c. How has data been used?

d. How has data been altered/processed/updated?

e. Who has had access to data?

f. Has data been destroyed, and, if so, how?

g. What outcomes have been produced from the data?

h. Have any new datasets been created?

i. How has the data been inspected at this stage?

2. Data Integrity

a. Was the model selection process constrained by the requirement that any model must be interpretable to ensure individuals' right to be informed?

b. How has the 'testing' (unseen) and 'validation' datasets been selected to ensure data integrity?

c. Have training and testing data splits[36] been fully documented?

d. Is the number of variables included resulting in overfitting[37] or underfitting?

e. If so, what changes have been made to correct this?

f. How has the testing and validation process been conducted, and will this process be published?

....

g. Has the model been re-tested after such changes were made and performance and accuracy documented?

....

h. Has a (second) impact assessment been done?

....

i. Are ethical considerations of the model clearly identified, particularly if new ones have been discovered after the development of the model?

....

j. How does any visualisation of data and/or summary statistics demonstrate that the data is attributable, consistent, complete, and contemporaneous?

....

k. How often will regular performance tests and iterative impact assessments be conducted throughout the full duration of model deployment?

....

l. How has data been archived and has the associated metadata been stored alongside the relevant dataset, or is it otherwise securely traceable through documentation?

....

m. Is it possible to successfully retrieve data and datasets, with associated metadata, from the archives?

....

n. Is data stored in an interoperable and reusable format to promote replicability and transparency?

....

o. (Applicable after deployment) How is the system monitored to ensure it is serving the intended purpose and being used responsibly within that scope?

....

p. (Applicable after deployment) Have users of the model reported that the system is useful, reliable, and accurate amongst others?

....

q. To what extent have resources and practices been applied, reviewed and updated, to ensure the integrity of the data commensurate to the risk of a Data Integrity failure?

....

Considering answers to the previous questions, what actions will be taken to ensure data integrity in the following phase of this project?

....

3. Data Quality

a. Has the dataset been checked for: incomplete data, outdated data, duplicated data, errors in manual data entry, inconsistent formats, poor-quality metadata?

....

b. Are there any outliers (unexpected data) or imbalanced/unbalanced classes, or correlations?

....

c. Has external validation (and/or cross-validation) of the model been carried out prior to full deployment to verify whether training data were adequate stand-in for data encountered in novel settings?

....

d. Was the model selection process constrained by the requirement that any model must be interpretable to ensure individuals' right to be informed?

....

e. [If public-facing] Is the interface accessible? Considerations to address include accessibility for users with disabilities and those who are colour-blind, for example.

....

f. Has new data been collected and used to retrain/revalidate the model?

....

g. Are any processes in place (e.g. automated triggers) to frequently check whether the model is still representative of the original data generation process?

....

h. (Applicable after deployment) How have implementers of the system been trained to: understand the logic of the system, explain its decisions in plain language, and use independent and unbiased judgement to gauge the quality of the output?

....

i. (Applicable after deployment) Have new variables emerged since deployment that would enhance the overall data quality and performance of the model?

....

Considering answers to the previous questions, what actions will be taken to ensure data quality in the following phase of this project?

....

4. Data Protection and Privacy

a. What measures have been taken to ensure that data is stored securely, minimising risks of data leakage or security breaches?

....

b. Are there mechanisms in place that can allow for system testing without using real world data?

....

c. What steps have been taken to minimise risks of (re)identification of data subjects?

....

d. How long is data expected to be used and stored, and what is the process for which data will be deleted when no longer necessary?

....

e. How have implementers of the system been trained to use independent and unbiased judgement to ensure the rights of the data/decision subject are upheld?

....

Considering answers to the previous questions, what actions will be taken to ensure data protection and privacy in the following phase of this project?

....

Data Factsheets Across the Workflow

The Data Factsheet has most relevance within the design and development stages given the importance of establishing plans for how data is to be used, applied, and shared. However, the process of completing a Data Factsheet should be used as an exercise throughout the AI lifecycle to facilitate critical reflection on data management practices and as an exercise to inform – as well as document - actions taken to continually assure Responsible Data Stewardship. The Dataset Factsheet should be conceptualised at the Project Planning stage of the project to ensure that each component of Responsible Data Stewardship is considered from the onset of the project and throughout the data collection and procurement process. It is to be revisited at the end of the Data Analysis step, again at the end of Model Training, Testing, and Validation, and iteratively throughout System Use and Monitoring. By maintaining The Data Factsheet Template (page 27) during these critical stages, you and your team will be able to help to iteratively identify any data-related concerns and respond to these through activities that assure the achievement of each component of Responsible Data Stewardship at all points in the project.

Responsible Data Stewardship in Practice

Activities

Activities Overview

In the previous sections of this workbook, we have presented an introduction to the core concepts of Responsible Data Stewardship. In this section we provide concrete tools for applying these concepts in practice. Activities will help participants work towards understanding, establishing, and deploying ethical data governance processes in the context of AI through applications of existing risk management frameworks, responsible data principles, and data management methodologies.

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

Workshop resources for Responsible Data Stewardship in Practice

A Note on Activity Case Studies

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

Note for Facilitators

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

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

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

Exploring Data Use in Policing

Gain familiarity with the data lifecycle and overarching considerations for Responsible Data Stewardship.

Corresponding Sections

- → The Data Lifecycle (page 11)

Useful Sections to Revise

- → Technical Components of AI and Machine Learning (ML), Component 1: Data from AI Ethics and Governance in Practice: An Introduction

Exploring Components of Responsible Data Stewardship

Consider the benefits and limitations of incorporating the different components of Responsible Data Stewardship as harm-mitigation mechanisms in AI projects.

Corresponding Sections

- → Key Components of Responsible Data Stewardship (page 17)

Prerequisites

- → Activity: Exploring Data Use in Policing (page 47)

Creating Data Factsheets

Groups will answer questions within a Data Factsheet pertaining to the case studies provided in the previous activity.

Corresponding Sections

- → Part Two: Putting Responsible Data Stewardship into Practice (page 26)

Useful Sections to Revise

- → Stages of the AI/ML Project Lifecycle from AI Ethics and Governance in Practice: An Introduction

40 mins | Participant Instructions

Exploring Data Use in Policing

Objective

Gain familiarity with the data lifecycle and overarching considerations for Responsible Data Stewardship.

Team Instructions

1. In this activity, your team will be split into groups. In your groups, individually read over the case study for this activity. The facilitator will answer any questions you may have about the case study.

2. Have a group discussion considering the following questions:

   - How, if at all, might the use of data in this AI project support the public interest?

   - What stakeholders may be considered data subjects within this AI system?

   - Would all data subjects see this system as a public good?

   - What, if any, risks does this system pose to data subjects?

   - What might data subjects require from this project in order to trust it?

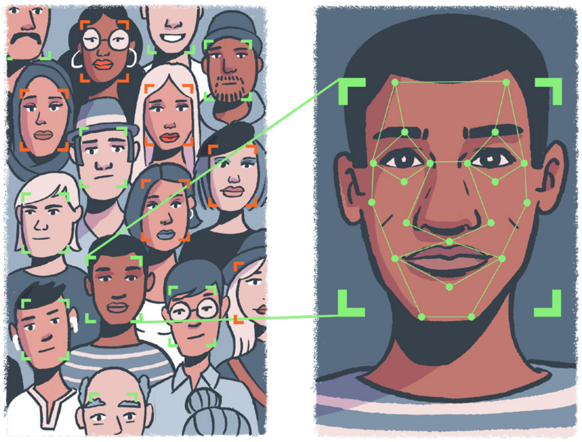
3. Have a volunteer within your group take notes about your discussion within your group's section of the board.

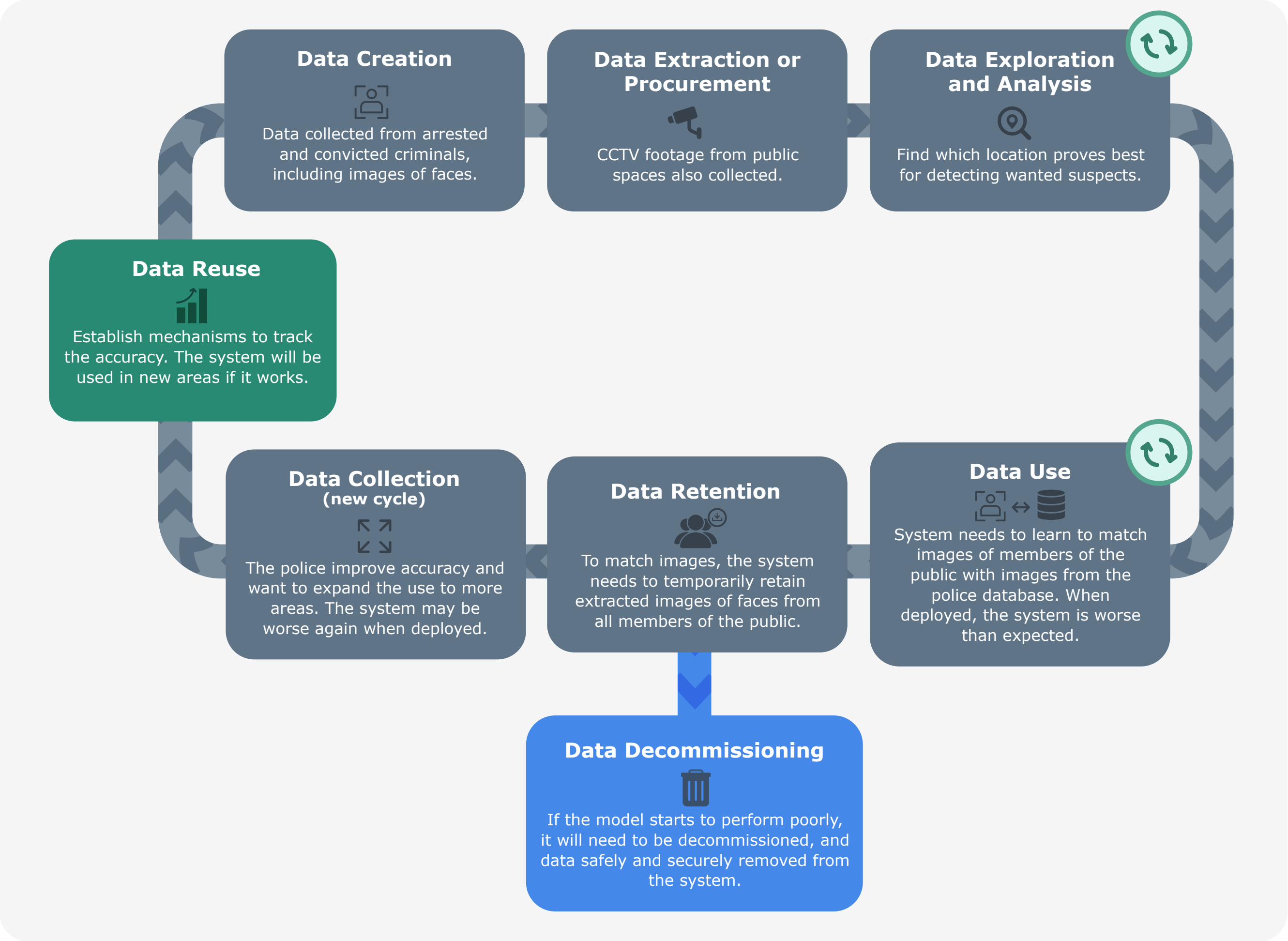
4. Reconvene as a team, having each group's note-taker share notes from their discussion.

5. Have a team-wide discussion about the possible benefits and risks posed by the AI system under consideration.

Case Study Data Lifecycle of an Automated Facial Recognition System

A local police force wants to improve their ability to identify wanted suspects and reduce crime. Currently, team members maintain that there are not enough officers to patrol public spaces and identify wanted suspects. A senior team member proposes developing an automated facial recognition system. The system will be trained on image data gathered from a surveillance system and has an objective of matching these images to a police database of wanted suspects and known criminals.





Data Creation

Data collected from arrested and convicted criminals, including images of faces.

Data Extraction or Procurement

CCTV footage from public spaces also collected.

Data Exploration and Analysis

Find which location proves best for detecting wanted suspects.

Data Use

System needs to learn to match images of members of the public with images from the police database. When deployed, the system is worse than expected.

Data Retention

To match images, the system needs to temporarily retain extracted images of faces from all members of the public.

Data Decommissioning

If the model starts to perform poorly, it will need to be decommissioned, and data safely and securely removed from the system.

Data Collection (new cycle)

The police improve accuracy and want to expand the use to more areas. The system may be worse again when deployed.

Data Reuse

Establish mechanisms to track the accuracy. The system will be used in new areas if it works.

Data Creation

To develop this system, the police force needs to have a **dataset with the prior records of known criminals or wanted suspects**, including images of their faces. This is because the facial recognition system cannot identify a wanted suspect or known criminal if it has no record of who they are.

The police force also needs **images of members of the public**. In developing this system, they decide to collect this data from CCTV cameras that are in use throughout public spaces, such as high streets or shopping centres.

Data Extraction or Procurement

The police force decides to **collect data from surveillance systems already installed in several public locations**: a shopping centre, a busy high street, and a park.

> Note: The collection of personal data for policing is rarely based on the consent of the individuals. Instead, police rely on a principle known as ‘legitimate interest’. That is, they can collect data if doing so serves a legitimate interest that allows them to fulfil their duty as a public authority. The police therefore collect images of members of the public from public areas for the purpose of identifying wanted suspects.

Data Analysis

Next, they analyse the data to determine which location proves best for detecting wanted suspects. By exploring and analysing the data, **the police force determine that a local shopping centre would be the best location for their system**.

The police’s facial recognition system needs to learn to match images of members of the public with images from the police database.

The images that the CCTV cameras record contain many faces at the same time. So, **the team decide to use AI to detect and extract the individual faces from the video and try to match these images with the dataset of wanted suspects**. This requires using a type of machine learning known as supervised learning.

Training of the System

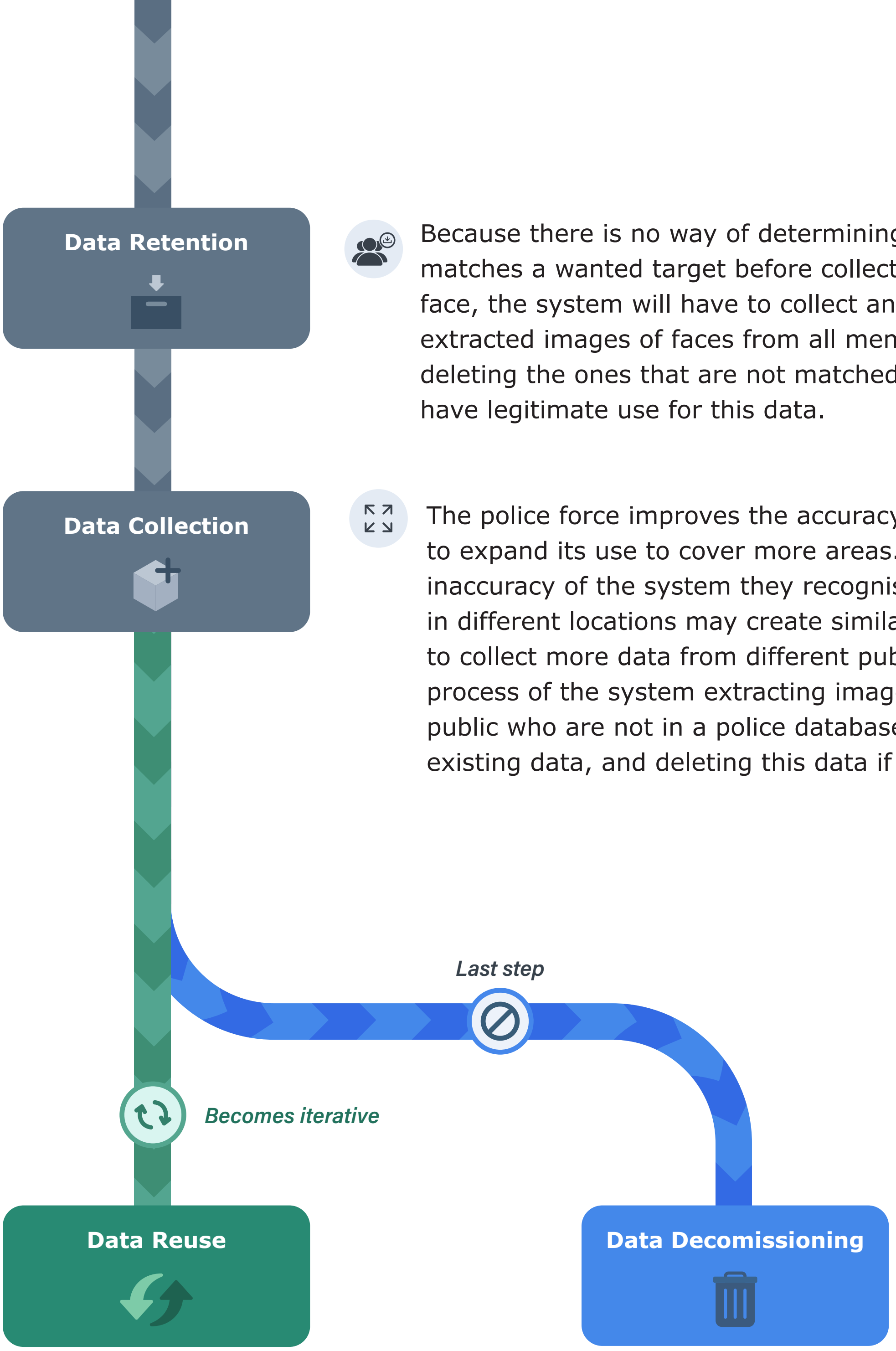
Each of the images used in the training of the system has a label. The label says whether the image belongs to a wanted suspect or an ordinary member of the public. The system takes an image from a training set and has to find a possible match in the police database. During training, the system cannot access the labels. But once the system has made its decision, the label reveals whether the decision was correct or incorrect.

The system uses this feedback to improve its algorithm over time. **The system’s goal is to get good enough at matching images from the labelled dataset so it can work in a real-world setting.**

When they deploy the system in the real world the system does not act as expected. **The accuracy of the system is a lot worse than it was during their testing**. They return to their Data Factsheet, and upon reviewing their data records, quickly realise that there are gaps in their original data.

The System in the Real World

When they use the system in the real world, **it encounters more diverse groups of people than are present in their dataset**. Specifically, it does not have many images of elderly people or people with darker skin colours. An overview of the **set of locations where data was collected reveals that these did not yield a dataset representative of the environment where the system was deployed**, causing the unexpected inaccuracy.



Data Retention

Because there is no way of determining whether an individual matches a wanted target before collecting visual data of their face, the system will have to collect and temporarily retain extracted images of faces from all members of the public, deleting the ones that are not matched as the police doesn’t have legitimate use for this data.

Data Collection

The police force improves the accuracy of the model and want to expand its use to cover more areas. Considering the previous inaccuracy of the system they recognise that the reuse of data in different locations may create similar issues. They decide to collect more data from different public areas, repeating the process of the system extracting images of members of the public who are not in a police database to supplement their existing data, and deleting this data if there is no match.

Last step

Becomes iterative

Data Reuse

The police establish mechanisms to track the accuracy of their system. If it continues to be good enough, they will redeploy it in new areas.

Data Decomissioning

If the model starts to perform poorly, it will need to be decommissioned, and data safely and securely removed from the system.

40 mins | Facilitator Instructions

Exploring Data use in Policing

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

2. Next, split the team into groups.

3. Let the groups know that they will have 15 minutes to read the case study and have their discussions.

   - **Facilitators** and **co-facilitators** are to join the groups and provide support as needed.

4. When 5 minutes have passed, let the groups know that they should transition onto their discussion if they haven't already.

5. Let the groups know when there are 3 minutes left for this section of the activity.

6. When the time has passed, ask the team to reconvene.

7. Give each volunteer note-taker 3 minutes to give feedback on their group's discussions.

8. When all groups have shared, lead a discussion about the possible benefits and risks posed by the AI system under consideration.

   - **Co-facilitator:** Use sticky notes to write out any possible benefits and harms identified. Place these notes in the Team Notes section of this activity within the workshop board.

40 mins | Participant Instructions

Exploring Components of Responsible Data Stewardship

Objective

Consider the benefits and limitations of incorporating the different components of Responsible Data Stewardship as harm-mitigation mechanisms in AI projects.

Team Instructions

1. This activity builds on the previous activity. Your team will be split in the same groups, each group being assigned a component of Responsible Data Stewardship.

2. In your groups, individually review the case study used for the previous activity, as well as the descriptions of your assigned component of Responsible Data Stewardship. As you read the case study, keep in mind how your group's assigned component relates to it.

3. Having reviewed the case study, have a group discussion about the role of your assigned component in the management of this project's data. Consider:

   - Why might it be important for your assigned component of Responsible Data Stewardship to be considered in this AI project?

   - What, if any, parts of this case study point to your assigned component of Responsible Data Stewardship being considered in the data lifecycle?

   - What, if any, elements point to your component not being considered?

   - What are the risks of this component not being considered?

4. Have a volunteer within your group take notes pertaining to your discussion objectives within your group's section of the board.

5. Reconvene as a team, having each group's note-taker feedback about their discussion.

6. Next, have a team-wide discussion about the extent to which the components of Responsible Data Stewardship may help mitigate possible risks posed by the AI system under consideration.

40 mins | Facilitator Instructions

Exploring Components of Responsible Data Stewardship

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

2. Next, split the team into groups, with each group focusing on a specific component of Responsible Data Stewardship. Assign the following components to the groups:

   - Group 1: Data Quality
   - Group 2: Data Integration

   Ensure that each group understands their designated component and answer any questions. Please note that Data Protection and Privacy will not be addressed in this particular task.

3. Let the groups know that they will have 15 minutes to read over the case study and have a discussion.

   - **Facilitators** and **co-facilitators** are to join the different groups throughout this time, providing support where needed.

4. When there are three minutes left for this section of the activity, let the groups know.

5. Ask the team to reconvene, giving the team a moment to read over the group's notes on the board.

6. Next, lead a discussion about the extent to which Responsible Data Stewardship may mitigate possible risks posed by the AI system under consideration. Ask the team to review the risks identified in the previous activity. Consider the questions on the following page.

7. Invite participants to share examples of how the components of Responsible Data Stewardship helped or could have helped mitigate risks posed by AI applied in their own fields.

Questions

a. How might the possible risks identified in the previous activity be mitigated by incorporating the components of Responsible Data Stewardship?

- **Co-facilitator:** Take notes about the group discussion in the Notes section of this activity on the board.

b. Which risks would be mitigated by the effective implementation of Responsible Data Stewardship?

- **Co-Facilitator:** As participants identify risks, move the sticky notes pertaining to these risks to the Mitigated Risks section of the Group Discussion part of the board.

c. What, if any, possible risks that fall outside of the scope of Responsible Data Stewardship?

- **Co-Facilitator:** As participants identify risks, move the sticky notes pertaining to these risks to the Risks Out of Scope section of the Group Discussion part of the board.

d. Would Risks out of scope be addressed by other principles in the guidance (such as Fairness, Sustainability, Explainability)?

- **Consider:** How might components of Responsible Data Stewardship works alongside other principles (such as Fairness, Sustainability, Explainability) to provide a comprehensive approach to risk mitigation in AI projects?
- **Co-facilitator:** Take notes about the group discussion in the Notes section of this activity on the board.

Facilitator Considerations Exploring Components of Responsible Data Stewardship

For the group discussion, it will be important to stress that Responsible Data Stewardship work alongside other principles (such as Fairness, Sustainability, Explainability) to provide a comprehensive approach to risk mitigation in AI projects. Some of the risks posed by this project, such as potential risks to Application Fairness, may be mitigated through activities associated with other principles, such as Bias Self-Assessments and Bias Risk Management, and Stakeholder Engagement Processes.

45 mins | Participant Instructions

Creating Data Factsheets

Objective

Practise answering questions within Data Factsheets as pertaining to specific AI projects.

Team Instructions

1. In this activity, you will return to the groups assigned in the previous activity.
2. Groups will be tasked with answering questions within a Data Factsheet created for the case study.
   - **Note:** the factsheet has been tailored to fit the scope of this activity. It only includes selected sections and questions that highlight significant components of the AI project.
3. In your groups, take a moment to individually read over the Data Factsheet Template provided.
4. As a group, discuss the unanswered questions within the template.
5. Have a volunteer write out your group's answers in sticky notes, placing them in their respective section of the template.
6. Having answered the questions, discuss how your group may present a summary of your discussion to the rest of the team. Consider:
   - What did you learn about this project when answering questions within its Data Factsheet?
7. Have a volunteer from your group write notes about your discussion.
8. Reconvene as a team, having volunteers from each group present their insights.

45 mins | Facilitator Instructions

Creating Data Factsheets

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

2. Next, split the team into the same groups as the previous activity. Groups will have 20 minutes to work on the case study.

3. Facilitators and co-facilitators are to join groups, providing support as needed.

4. Let the groups know when there are 10 minutes left for this section of the activity, asking them to transition onto preparing their summary if they haven’t already.

5. When 10 minutes have passed, ask the team to reconvene, giving a volunteer from each team a few minutes to present the insights of their group.

6. After each presentation, give the team some minutes to discuss. Consider:

    - What, if any, components of Responsible Data Stewardship seem most challenging to achieve in this project and why?

    - What, if any, are important actions that should be implemented to assure Responsible Data Stewardship within this project?

Appendix A: Generative AI and Data

The large-scale era of AI/ML technologies emerged in the mid-to-late 2010s with the radical leap in compute capacity, training dataset size, and model complexity that accompanied the birth of a new generation of industry-produced foundation models encompassing an ever-improving set of large language models (LLMs) and generative AI models.[38]

KEY CONCEPTS

Foundation Models (FMs)

Foundation models are AI technologies trained on very large, “broad” datasets that can be applied to a wide range of tasks and purposes.[39] These models are considered to form the “foundation” or base architecture of other systems. For example, a number of new applications, such as ChatGPT and the conversational features of the Microsoft Bing search engine, have been built on top of successive versions of OpenAI’s GPT foundation model, which has been designed for natural language processing (NLP) tasks.[40]

Generative AI

Deep-learning models that can generate high-quality text, images, code, and other content based on their training data.[41]

Large Language Models (LLMs)

The ‘large’ label refers to the number of values or parameters (often in the billions) that the model can change autonomously as it learns. LLMs use deep learning techniques, including a subset of neural networks called ‘transformers’. Large-sized pre-trained language models that are trained on vast amounts of data use transformers to perform natural language processing (NLP), for instance, and which have capabilities to generate natural language and other types of content to perform diverse tasks.[42] [43] [44] The Pathways Language Model (PaLM) is an example of a 540-billion parameter-transformer-based LLM developed by Google AI.

The use of self-supervised learning techniques that removed the need for manually labeled datasets enabled model training to include boundless and vast volumes of unannotated, internet-scale data.[45] [46] [47] [48] [49] However, this far outstripped the capacity of human project teams to manually check data quality and source integrity, let alone scrutinise sampling deficiencies and other problematic aspects that arose in data creation. The embrace of this 'effort to scale' at the expense of the 'attention to care'[50] necessary to safeguard responsible data stewardship has given rise to a broad range of risks related to massive and uncurated web-scraped training datasets,[51] [52] some of which are listed below:

Risks and harms associated with large scale and vast unannotated training data

Concerns about privacy risks. Potential and real harms can derive from:

- Data poisoning. The scaled and indiscriminate extraction of data from the internet has exponentially broadened the attack surface for the injection of adversarial noise into training datasets.[53] [54] [55] This kind of data poisoning can corrupt the parameters of trained foundation models, introducing unreliable, poor, and harmful performance for targeted inputs. These sorts of web-scale poisoning attacks can be launched inexpensively and with relative ease.[56]

- Privacy leakage and memorisation. The presence of personally identifiable information (e.g. email addresses and phone numbers) and sensitive documents (e.g. personal medical records) in massive pretraining corpora can yield privacy leaks during model prompting and data extraction attacks.[57] [58] [59] [60] Scholars have found that hazards of trained models emitting sensitive memorised data grow significantly as model capacity increases.[61]

Violations of data protection rights and infringements. In jurisdictions where data protection laws are in place, parties that are collecting and processing personally identifiable information must establish the legal basis for this activity. According to the General Data Protection Law (GDPR), this can occur either by gaining the consent of affected data subjects or by establishing a legitimate interest for using their data that is consonant with data protection principles such as fairness, transparency, and purpose limitation (ICO, 2022.). The establishment of a legal basis for secondary data use in foundation models and generative AI technologies requires considerations of "the context of the original processing and the subsequent purposes of use".[62] However, the unfathomability of the datasets used to train foundation models or generative AI systems categorically excludes the viability of careful and case-specific contextual considerations and thus impedes clear justification of purpose limitation and legitimate interest. Combined with the generalised failure to attain consent from those whose personally identifiable digital trace data have been scraped from the internet, the inability to establish a legitimate interest consistent with data protection principles calls into question the very lawfulness of the processing of personal data in the training of the foundation models or generative AI systems.

Copyright infringements and violations of intellectual property law. The impracticability of diligent human data curation and stewardship has led to a general failure both to obtain consent from copyright holders and to establish a legal basis for the legitimate use of copyrighted material (for instance, through data licensing regimes). Combined with the ability of generative AI systems to memorise and then replicate elements of this material that are embedded in their training data, such a failure to establish lawfulness has precipitated risks of outright 'digital forgery'[63] and AI-enabled content piracy or theft.[64] [65] [66] These risks of potential copyright violations have become an area of fierce debate amidst the rapid commercialisation of generative AI systems.

Risks of serious psychological, allocational, and identity-based harms. Potential and real harms can derive from:

- Discriminatory and toxic content embedded in web-scraped data. It is by now well-established that the large-scale, opaque datasets contain ingrained patterns of bias And discrimination,[67] [68] [69] [70] [71] [72] [73] [74] exclusionary norms,[75] toxic and abusive language,[76] [77] [78] microaggressions,[79] and stereotyping.[80] [81] [82] [83]

- Demographically skewed datasets. The data used to train foundation models also reflect inequities of internet access; inequalities that manifest in local, regional, and global digital and data divides; geographic biases in data collection (e.g. oversampling in the 'Global North'); overrepresentation of dominant languages (e.g. English), hegemonic cultural values (e.g. US views), and gender and age groups (e.g. males and younger people).[84] [85] [86] [87] [88] [89]

These data harbor representational imbalances that crystallise in foundation models and generative AI training datasets and that lead to the underrepresentation, invisibility, or erasure of historically marginalised and minoritised communities. Such imbalances can lead, in turn, to disproportionately poor model performance and deficient quality-of-service for these historically marginalised and minoritised demographic groups.[90] [91] [92] [93] [94]

Risks of discriminatory and psychological harm exacerbated by the additional negative impacts from unfair biases arising in technical mitigation measures.

Dodge et al (2021) show evidence, for instance, that blocklist filters comprised of banned words (which are used to clean web-crawled datasets) "disproportionately [remove] documents in dialects of English associated with minority identities (e.g. text in African American English, text discussing LGBTQ+ identities)" (p. 2). Bender et al. (2021) similarly point out that blocklists which contain generic words associated with obscenity and pornography can attenuate the presence and "influence of online spaces built by and for LGBTQ people" (p. 614).

Other researchers have demonstrated that annotator biases that surface in the construction of datasets used train classifiers for the detection of toxic language and hate-speech in foundation models datasets engender the embedding of racist attitudes, beliefs, and stereotypes that give rise to systemic prediction errors in these classifiers.[95] [96] [97]

In their analysis of the perpetuation of stereotypes by text-to-image generative AI systems, Bianchi et al. (2023) show that—even where systems like Dall-E contain supposed ‘guardrails’ set up to prevent the production of harmful content—noxious stereotyping behavior not only endures unmitigated but amplifies existing societal biases (p. 3).

For more reading on these topics, also refer to the AI Fairness in Practice workbook.

Endnotes

1 GPAI Data Governance Working Group. (2022). *A Framework Paper for GPAI's work on Data Governance.* Global Partnership on Artificial Intelligence (GPAI). GPAI Montréal Summit. https://gpai.ai/projects/data-governance/gpai-data-governance-work-framework-paper.pdf

2 Leslie, D., Katell, M., Aitken, M., Singh, J., Briggs, M., Powell, R., Rincón, C., Chengeta, T., Birhane, A., Perini, A., Jayadeva, S., & Mazumder, A. (2022). Advancing Data Justice Research and Practice: An Integrated Literature Review. The Alan Turing Institute in collaboration with The Global Partnership on AI. https://doi.org/10.5281/zenodo.6408304

3 Information Commissioner's Office (2022). *Chapter 5: Privacy-enhancing technologies (PETs)*. Draft anonymisation, pseudonymisation and privacy enhancing technologies guidance. Available at https://ico.org.uk/media/about-the-ico/consultations/4021464/chapter-5-anonymisation-pets.pdf

4 see British Standards Institution. (2023). Information technology – Artificial intelligence # Data life cycle framework. (BSI Standard No. ISO/IEC 8183). https://standardsdevelopment.bsigroup.com/projects/2022-00406#/section

5 British Standards Institution. (2023). Information technology – Artificial intelligence # Data life cycle framework. (BSI Standard No. ISO/IEC 8183). https://standardsdevelopment.bsigroup.com/projects/2022-00406#/section

6 Christopher Burr. (2021). A Citizen's Guide to Data: Ethical, Social, and Legal Issues. Zenodo. https://doi.org/10.5281/zenodo.5568861

7 Christopher Burr. (2021). *A Citizen's Guide to Data: Ethical, Social, and Legal Issues.* Zenodo. https://doi.org/10.5281/zenodo.5568861

8 GPAI Data Governance Working Group. (2022). *A Framework Paper for GPAI's work on Data Governance.* Global Partnership on Artificial Intelligence (GPAI). GPAI Montréal Summit. https://gpai.ai/projects/data-governance/gpai-data-governance-work-framework-paper.pdf

9 Leslie, D., Katell, M., Aitken, M., Singh, J., Briggs, M., Powell, R., Rincón, C., Chengeta, T., Birhane, A., Perini, A., Jayadeva, S., & Mazumder, A. (2022). Advancing Data Justice Research and Practice: An Integrated Literature Review. The Alan Turing Institute in collaboration with The Global Partnership on AI. https://doi.org/10.5281/zenodo.6408304

10 The Alan Turing Institute and ICO (2022, October). Explaining Decisions Made with AI. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/

11 Intra-governmental Group on Geographic Information (IGGI) (2005). The Principles of Good Data Management. The Office of the Deputy Prime Minister: London. Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/14867/Good_dataMan.pdf

12 Wong, J., Henderson, T., & Ball, K. (2022). Data protection for the common good: Developing a framework for a data protection-focused data commons. *Data & Policy, 4,* e3. doi:10.1017/dap.2021.400

13 Wong, J., Henderson, T., & Ball, K. (2022). Data protection for the common good: Developing a framework for a data protection-focused data commons. *Data & Policy, 4,* e3. doi:10.1017/dap.2021.400

14 Data Economy Lab (2021). Data trust. Available at https://tool.thedataeconomylab.com/data-models/10 (accessed 29 January 2024)

15 Pavel, V. (2021). Exploring legal mechanisms for data stewardship. Ada Lovelace Institute and UK AI Council)

16 Pavel, V. (2021). *Exploring legal mechanisms for data stewardship.* Ada Lovelace Institute and UK AI Council. Available at https://www.adalovelaceinstitute.org/wp-content/uploads/2021/03/Legal-mechanisms-for-data-stewardship_report_Ada_AI-Council-2.pdf

17 Lawrence, N., & Oh, N. (2021). Enabling data sharing for social benefit through data trusts. https://gpai.ai/projects/data-governance/data-trusts/enabling-data-sharing-for-social-benefit-through-data-trusts.pdf

18 British Academy and the Royal Society (2017) *Data Management and Use: Governance in the 21st Century*. London: The British Academy and the Royal Society.

19 Intra-governmental Group on Geographic Information (IGGI) (2005). The Principles of Good Data Management. The Office of the Deputy Prime Minister: London. Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/14867/Good_dataMan.pdf

20 Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data, 3*(1), 1-9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/pdf/sdata201618.pdf

21 Leslie, D., Burr, C., Aitken, M., Katell, M., Briggs, M., & Rincon, C. (2022). Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal. *arXiv preprint arXiv:2202.02776*

22 Emran, N. A. (2015). Data completeness measures. In Pattern Analysis, Intelligent Security and the Internet of Things (pp. 117-130). Springer International Publishing. https://doi.org/10.1007/978-3-319-17398-6_11

23 Courtney, R., & Ware, W. (1989). Some Informal Comments About Integrity and the Integrity Workshop.". In Proceedings of the Invitational Workshop on Data Integrity (Ruthberg, ZG and Polk, WT, editors), National Institute of Standards and Technology, Special Publication. p. 1-18

24 SL Controls. (n.d.). What is ALCOA+ and Why Is It Important to Validation and Data Integrity. https://slcontrols.com/en/what-is-alcoa-and-why-is-it-importantto-validation-and-data-integrity/

25 see College of Policing. (2023). Code of Practice on police information and records management. https://www.gov.uk/government/publications/police-information-and-records-management-code-of-practice

26 Leslie, D., Burr, C., Aitken, M., Katell, M., Briggs, M., & Rincon, C. (2022). Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal. *arXiv* preprint arXiv:2202.02776

[27] Public Sector Equality Duty. https://www.gov.uk/government/publications/public-sector-equality-duty

[28] Leslie, D., Katell, M., Aitken, M., Singh, J., Briggs, M., Powell, R., ... & Mazumder, A. (2022). Advancing data justice research and practice: An integrated literature review. https://advancingdatajustice.org/wp-content/uploads/2022/11/advancing-data-justice-research-and-practice-an-integrated-literature-review.pdf

[29] Jagadish, H., Stoyanovich, J., & Howe, B. (2022). The Many Facets of Data Equity. *ACM Journal of Data and Information Quality, 14*(4), 1-21.

[30] Adapted from data equity principles presented by CDC Foundation: https://www.cdcfoundation.org/data-equity-principles?inline

[31] Leslie, D., Katell, M., Aitken, M., Singh, J., Briggs, M., Powell, R., ... & Mazumder, A. (2022). Advancing data justice research and practice: An integrated literature review. https://advancingdatajustice.org/wp-content/uploads/2022/11/advancing-data-justice-research-and-practice-an-integrated-literature-review.pdf

[32] College of Policing (2014). Data protection. https://www.college.police.uk/app/information-management/data-protection

[33] Leslie, D., Burr, C., Aitken, M., Katell, M., Briggs, M., & Rincon, C. (2022). Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal. *arXiv preprint arXiv:2202.02776*

[34] User-centred design is an approach that allows end-users to influence the design of the AI system with the aim to increase usability. Dabbs, A. D. V., Myers, B. A., Mc Curry, K. R., Dunbar-Jacob, J., Hawkins, R. P., Begey, A., & Dew, M. A. (2009). User-centered design and interactive health technologies for patients. *CIN: Computers, Informatics, Nursing, 27*(3), 175-183.

[35] Data-protection-by-design is an approach that ensures that project teams consider privacy and data protection issues at the design phase of any AI system and then throughout the AI lifecycle. (For more information, see https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/guide-to-accountability-and-governance/accountability-and-governance/data-protection-by-design-and-default/)

[36] During Model Selection & Training stage of the AI/ML project lifecycle, preprocessed data is split into a training set, testing set, and validation set. Splitting the data into training and testing data ensures that the model can perform well for the original and newer data. See AI Ethics and Governance in Practice: An Introduction for more information.

[37] Overfitting occurs when the model's mapping function is matched too closely to the patterns arising in the training data. For more information, refer to the AI Safey in Practice Workbook.

[38] Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., … & Yang, M. H. (2023). Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys, 56*(4), 1-39.

39 Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Dem- szky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchan- dani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christo- pher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Kr- ishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. https://doi.org/10.48550/arXiv.2108.07258 Issue: *arXiv:2108.07258 arXiv: 2108.07258* [cs].

40 3 Domínguez Hernández, A., Krishna, S., Perini, A. M., Katell, M., Bennett, S. J., Borda, A., Hashem, Y., Hadjiloizou, S., Mahomed, S., Jayadeva, S., Aitken, M., & Leslie, D. (2024). Mapping the individual, social, and biospheric impacts of Foundation Models. *In The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24),* June 3–6, 2024, Rio de Janeiro, Brazil. ACM.

41 IBM (n.d.) What is generative AI? https://research.ibm.com/blog/what-is-generative-AI

42 Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology, 15*(3), 1-45.

43 Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223.*

44 IBM (n.d.) What are LLMs? https://www.ibm.com/topics/large-language-models

45 Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

46 Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

47 McIntosh, T. R., Susnjak, T., Liu, T., Watters, P., & Halgamuge, M. N. (2023). From google gemini to openai q*(q-star): A survey of reshaping the generative artificial intelligence (ai) research landscape. *arXiv preprint arXiv:2312.10868*.

48 Triguero, I., Molina, D., Poyatos, J., Del Ser, J., & Herrera, F. (2024). General Purpose Artificial Intelligence Systems (GPAIS): Properties, definition, taxonomy, societal implications and responsible governance. *Information Fusion, 103,* 102135.

49 Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., ... & Hendrycks, D. (2023). Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405.*

50 7 Seaver, N. (2021). Care and scale: decorrelative ethics in algorithmic recommendation. *Cultural Anthropology, 36*(3), 509-537.

51 Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623)

52 Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and applications of large language models. arXiv preprint *arXiv:2307.10169*.

53 Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

54 Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., ... & Tramèr, F. (2023a). Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149*.

55 Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., ... & Hadfield-Menell, D. (2024). Black-Box Access is Insufficient for Rigorous AI Audits. *arXiv preprint arXiv:2401.14446*.

56 Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2023b). Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, p. 2

57 Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)* (pp. 2633-2650).

58 Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

59 Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023, May). Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)* (pp. 346-363). IEEE.

60 Mozes, M., He, X., Kleinberg, B., & Griffin, L. D. (2023). Use of llms for illicit purposes: Threats, prevention measures, and vulnerabilities. *arXiv preprint arXiv:2308.12833*.

61 Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2023). Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, p.1

62 Garrido (2024). Data protection and generative AI: an unfinished answer. *Harvard Data Science Review*.

63 Somepalli, G., Singla, V., Goldblum, M., Geiping, J., & Goldstein, T. (2023). Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6048-6058).

64 Bird, C., Ungless, E., & Kasirzadeh, A. (2023, August). Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 396-410).

65 Piskopani, A. M., Chamberlain, A., & Ten Holter, C. (2023, July). Responsible AI and the Arts: The Ethical and Legal Implications of AI in the Arts and Creative Industries. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems* (pp. 1-5).

66 Sobel, B. (2024, February 16). Don't give AI free access to work denied to humans, argues a legal scholar. The Economist. Retrieved from https://www.economist.com/by-invitation/2024/02/16/dont-give-ai-free-access-to-work-denied-to-humans-argues-a-legal-scholar#

67 Abid, A., Farooqi, M., & Zou, J. (2021, July). Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 298-306).

68 Birhane, A., & Prabhu, V. U. (2021, January). Large image datasets: A pyrrhic win for computer vision?. In 2021 *IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1536-1546). IEEE.

69 Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623).

70 Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social biases in NLP models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*.

71 Jin, Y., Chandra, M., Verma, G., Hu, Y., De Choudhury, M., & Kumar, S. (2023). Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries. *arXiv e-prints, arXiv-2310*.

72 Kirk, H. R., Jun, Y., Volpin, F., Iqbal, H., Benussi, E., Dreyer, F., … & Asano, Y. (2021). Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems, 34,* 2611-2624.

73 Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., & Daneshjou, R. (2023). Large language models propagate race-based medicine. *NPJ Digital Medicine, 6*(1), 195.

74 Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2023). Auditing and mitigating cultural bias in llms. *arXiv preprint arXiv:2311.14096*.

75 Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., ... & Gabriel, I. (2022, June). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214-229).

76 Dinan, E., Abercrombie, G., Bergman, A. S., Spruit, S., Hovy, D., Boureau, Y. L., & Rieser, V. (2021). Anticipating safety issues in e2e conversational ai: Framework and tooling. arXiv preprint *arXiv:2107.03451*.

77 Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). Realtoxicityprompts: Evaluating neural toxic degeneration in language models. arXiv preprint *arXiv:2009.11462*.

78 Nozza, D., Bianchi, F., Lauscher, A., & Hovy, D. (2022). Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion.* Association for Computational Linguistics.

79 Breitfeller, L., Ahn, E., Jurgens, D., & Tsvetkov, Y. (2019, November). Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 1664-1674).

80 Barlas, P., Kyriakou, K., Guest, O., Kleanthous, S., & Otterbacher, J. (2021). To" see" is to stereotype: Image tagging algorithms, gender recognition, and the accuracy-fairness trade-off. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW3), 1-31.

81 Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., … & Caliskan, A. (2023, June). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1493-1504).

82 Kotek, H., Dockum, R., & Sun, D. (2023, November). Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference* (pp. 12-24).

83 Ma, W., Scheible, H., Wang, B. C., Veeramachaneni, G., Chowdhary, P., Sun, A., ... & Vosoughi, S. (2023, December). Deciphering stereotypes in pre-trained language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

84 Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623).

85 Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.

86 Johnson, R. L., Pistilli, G., Menédez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J., & Bertulfo, D. J. (2022). The Ghost in the Machine has an American accent: value conflict in GPT-3. *arXiv preprint arXiv:2203.07785*.

87 Parashar, S., Lin, Z., Liu, T., Dong, X., Li, Y., Ramanan, D., ... & Kong, S. (2024). The Neglected Tails of Vision-Language Models. *arXiv preprint arXiv:2401.12425*.

88 Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., ... & Vassilev, A. (2023). Evaluating the social impact of generative AI systems in systems and society. *arXiv preprint arXiv:2306.05949*.

89 Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2023). Auditing and mitigating cultural bias in llms. *arXiv preprint arXiv:2311.14096*.

90 Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623).

91 Dev, S., Monajatipoor, M., Ovalle, A., Subramonian, A., Phillips, J. M., & Chang, K. W. (2021). Harms of gender exclusivity and challenges in non-binary representation in language technologies. *arXiv preprint arXiv:2108.12084*.

92 Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Kolawole, T., Fagbohungbe, T., ... & Bashir, A. (2020). Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*.

93 Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., ... & Vassilev, A. (2023). Evaluating the social impact of generative AI systems in systems and society. *arXiv preprint arXiv:2306.05949.*

94 Talat, Z., Névéol, A., Biderman, S., Clinciu, M., Dey, M., Longpre, S., … & Van Der Wal, O. (2022, May). You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models* (pp. 26-41).

95 Davani, A. M., Atari, M., Kennedy, B., & Dehghani, M. (2023). Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics, 11,* 300-319.

96 Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., & Kamar, E. (2022). Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

97 Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., & Smith, N. A. (2021). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Bibliography and Further Readings

Data Management, Data Governance, and Data Stewardship

Digital Curation Center & University of Edinburgh. (2020). The Role of Data in AI: Report for the working group of the Global Partnership of AI. Retrieved from: https://gpai.ai/projects/data-governance/role-of-data-in-ai.pdf

Data Risk Management Frameworks Overview

Chmielinski, Kasia S., Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. 'The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence'. ArXiv:2201.03954 [Cs], 10 January 2022. http://arxiv.org/abs/2201.03954.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 'Datasheets for Datasets'. Communications of the ACM 64, no. 12 (December 2021): 86–92. https://doi.org/10.1145/3458723.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 'Model Cards for Model Reporting'. In Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–29. Atlanta GA USA: ACM, 2019. https://doi.org/10.1145/3287560.3287596.

Zelenka, N., & Di Cara, N. H. 'Data Hazards (Version 0.1)' [Computer software]. https://github.com/very-good-science/data-hazards

Data Nutrition Label Project. https://datanutrition.org/labels/

Data Hazards. https://datahazards.com/

ICO (2012). Anonymisation: managing data protection risk code of practice. https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf

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

aiethics.turing.ac.uk

Version 1.2

The Alan Turing Institute