The Alan Turing Institute

AI Ethics and Governance

Day 4: Fairness, bias, accountability and governance

24/11/2022

Professor David Leslie
Director of Ethics and Responsible Innovation

OVERVIEW

Introduction to Practical Ethics

01

AI Harms & Values

02

AI Sustainability

03

Fairness, Bias Mitigation & Accountability

04

Explainability, CARE & Act principles

05

OVERVIEW

Introduction to Practical Ethics	AI Harms & Values	AI Sustainability	**Fairness, Bias Mitigation & Accountability**	Explainability, CARE & Act principles
01	02	03	**04**	05

CONTENTS

Day 3 recap

01 Fairness

Q&A

02 Bias mitigation

Q&A

Activity 1: Bias report design

Lunch break

03 Accountability and governance

Q&A

Activity 2: Establishing proportionate governance actions

Q&A

Recap from Day 3

AI Sustainability

Recap from Day 3

AI Sustainability

Anticipatory reflection

Recap from Day 3

AI Sustainability

- Anticipatory reflection
- The AI project lifecycle

Recap from Day 3

AI Sustainability

- Anticipatory reflection
- The AI project lifecycle
- Stakeholder engagement processs

01

AI FAIRNESS

AI fairness: The big picture

AI fairness: The big picture

- Understanding how concepts of fairness are used in the field of AI ethics and governance helps us understanding where and how unfair biases arise across AI project lifecycles

- Relevant notions of fairness operate both as ethical and legal criteria based upon which biases can be identified across the AI project workflow and as normative yardsticks against which they can be measured and then appropriately mitigated.

AI fairness: The big picture

- AI technologies, no matter how neutral they may seem, are designed and produced by human beings, who are bound by biases that can arise both in their cognitive processes and in their social environments

- Pre-existing or historical configurations of discrimination and social injustice—as well as the prejudices and biased attitudes that are shaped by such configurations—can be drawn into the AI innovation lifecycle and create unfair biases at any point in the project workflow.

Issues with defining the concept

- Many entry points for bias across AI lifecycle
- Many different sociotechnical contexts in which AI is used
- Different views on what fairness means

We will take a context-based and society-centred approach to understanding AI fairness based on two pillars

Defining fairness

On one hand, there is a plurality of views on what the meaning of fairness is

Defining fairness

On one hand, there is a plurality of views on what the meaning of fairness is

Nonetheless, there is considerable convergence around the importance of equality and non discrimination in fairness

General definition of fairness

Core belief in the equal moral status of all human beings and in the right of all human beings to equal respect, concern, protection, and regard before the law

Two key concepts

Equality

Non-discrimination

Three kinds of discrimination

Discriminatory Harassment

Unwanted or abusive behaviour linked to a protected characteristic violates someone's dignity, degrades their identity, or creates an offensive environment for them.

Three kinds of discrimination

- **Discriminatory Harassment**
- Unwanted or abusive behaviour linked to a protected characteristic violates someone's dignity, degrades their identity, or creates an offensive environment for them.

- **Direct Discrimination** (dispa
treatment)
- Individuals are treated adversely based on the membership in some protected class

Three kinds of discrimination

- **Discriminatory Harassment**
- Unwanted or abusive behaviour linked to a protected characteristic violates someone's dignity, degrades their identity, or creates an offensive environment for them.

- **Direct Discrimination** (disp
treatment)
- Individuals are treated adversely based on the membership in some protected class

- **Indirect Discrimination**
(disparate impact)
- Existing provisions, criteria, policies, arrangements, or practices—which could appear to be neutral—disparately harm or unfairly disadvantage members of some protected class.

Back to AI fairness...

AI fairness

- When it comes to AI systems: **‘Do no discriminatory harm’**

AI fairness

- When it comes to AI systems: **‘Do no discriminatory harm’**
- AI fairness considerations should ask three questions:

AI fairness

- When it comes to AI systems: **‘Do no discriminatory harm’**
- AI fairness considerations should ask three questions:

1/ How could the design, development, and use of the AI system lead to the discriminatory harassment of impacted individuals?

AI fairness

- When it comes to AI systems: **‘Do no discriminatory harm’**
- AI fairness considerations should ask three questions:

1/ How could the design, development, and use of the AI system lead to the discriminatory harassment of impacted individuals?

2/ How could the design, development, and use of the AI system lead to the disproportionate adverse treatment of impacted individuals from protected groups?

AI fairness

- When it comes to AI systems: **‘Do no discriminatory harm’**
- AI fairness considerations should ask three questions:

1/ How could the design, development, and use of the AI system lead to the discriminatory harassment of impacted individuals?

2/ How could the design, development, and use of the AI system lead to the disproportionate adverse treatment of impacted individuals from protected groups?

3/ How could the design, development, and use of the AI system lead lead to indirect discrimination against impacted individuals from protected groups?

A multi-pronged approach integrates:

- **Formal approaches to equality and non-discrimination** which focus primarily on consistent and impartial application of rules and equal treatment before the law

- **Substantive and transformative approaches** which focus on equalizing distribution of opportunities and outcomes and on addressing the material pre-conditions and structural changes needed for the universal realisation of equitable social arrangements

Principle of discriminatory non-harm

Principle of discriminatory non-harm

- Prioritise the identification and mitigation of biases and discriminatory influences

Principle of discriminatory non-harm

- Prioritise the identification and mitigation of biases and discriminatory influences
- Address discriminatory patterns that may originate in the data used to train, test, and validate the system and in the model architectures

Principle of discriminatory non-harm

- Prioritise the identification and mitigation of biases and discriminatory influences
- Address discriminatory patterns that may originate in the data used to train, test, and validate the system and in the model architectures
- Ensure that research, innovation, and implementation practices are undertaken in an optimally responsible and ethical manner

AI fairness - different contexts

AI fairness - different contexts

Social world context

AI fairness - different contexts

Social world context

Data context

AI fairness - different contexts

Social world context

Data context

Design, development and deployment context

AI fairness - different contexts

Social world context

Data context

Design, development and deployment context

Ecosystem context

Subcategories

AI Fairness

Subcategories

AI Fairness

Data fairness

Subcategories

AI Fairness

Data fairness

Application fairness

Subcategories

AI Fairness

Data fairness

Application fairness

Model design and development fairness

Subcategories

AI Fairness

Data fairness

Metric-based fairness

Application fairness

Model design and development fairness

Subcategories

AI Fairness

Data fairness

Metric-based fairness

Application fairness

System implementation fairness

Model design and development fairness

Subcategories

AI Fairness

- Data fairness
- Metric-based fairness
- Application fairness
- System implementation fairness
- Model design and development fairness
- Ecosystem fairness

Subcategories

Data fairness

- Representativeness
- Fit-for-purpose and sufficiency
- Source Integrity and Measurement Accuracy
- Timeliness and Recency
- Relevance, Appropriateness and Domain Knowledge

Subcategories

Application fairness

1/ Equity considerations surrounding real-world context of the policy issue to be solved

2/ Equity considerations surrounding the group targeted by the AI innovation intervention

3/ Equity considerations surrounding the way that the model's output shapes the range of possible decision-outcomes

Subcategories

Model design and development fairness

- Design
  - Problem formulation
- Development
  - Data pre-processing
  - Model selection
  - Model training, testing, and validation
  - Estimating and validating model structures

Subcategories

Metric-based fairness

- Involves the mathematical mechanisms that can be incorporated into an AI model to allocate the distribution of outcomes and error rates for relevant subpopulations (e.g., groups with protected or sensitive characteristics)

- There is a great diversity of beliefs in this area as to what makes the consequences of an algorithmically supported decision allocatively equitable, fair, and just.

- Regardless, project teams must ensure that the choices made regarding formal fairness metrics are lawful and conform to governing equality, non-discrimination, and human rights laws. They must also be transparent about the rationale behind their choices.

Subcategories

System implementation fairness

Implementation plan should include preparation for unbiased and responsible deployment of AI system by its on-the-ground users

- Decision-Automation Bias (over-reliance)
- Decision-Automation Bias (over-compliance)
- Automation-Distrust Bias

Subcategories

System implementation fairness

Taking account of the context of impacted individuals in system implementation

- Fair implementation processes should include considerations of the specific context of each impacted individual

Subcategories

System implementation fairness

Taking account of the context of impacted individuals in system implementation

- Fair implementation processes should include considerations of the specific context of each impacted individual

Subcategories

System implementation fairness

Taking account of the context of impacted individuals in system implementation

- Fair implementation processes should include considerations of the specific context of each impacted individual

- Equitable treatment of decision subjects entails that their unique life circumstances and individual contexts be taken into account in decision-making processes that are supported by AI-enabled statistical generalisations.

Subcategories

Ecosystem fairness

- AI system is embedded in a wider social system of economic, legal, cultural, and political structures or institutions
- Inequities and biases at this ecosystem level can steer or shape AI systems
- They can occur across a wide range of AI research and innovation contexts
- Ecosystem fairness highlights the importance of mitigating the range of inequity-generating path dependencies that originate at the ecosystem level which are often omitted from analyses of AI project lifecycles

Questions?

02 BIAS MITIGATION

WAITING ROOM
CHECKING

How does bias enter AI systems?

AI fairness - different contexts

Social world context

Data context

Design, development and deployment context

Ecosystem context

World

The social and historical reality that is antecedent to the design, development, and deployment of AI applications where pre-existing societal patterns of discrimination and social injustice arise

Data

The humanly collected measurements of social, demographic, and biophysical information and the historically generated and socially constructed products of human action and interaction

The wider social system of economic, legal, cultural, and political structures or institutions—and the policies, norms, and procedures through which these structures and institution influence human action.

Ecosystem

The sociotechnical design, development, and deployment lifecycle of AI systems, which is viewed as a complex and interrelated set of social and technological processes

Design

Real-world patterns of health inequity and discrimination

Real-world patterns of health inequity and discrimination
Historical inequity & social determinants of health
Structural biases in research agendas, medical knowledge, & clinical understanding
Underrepresentation in clinical trials and genetic studies
Unequal access & resource allocation
Implicit biases in the care environment
Discriminatory institutional processes
Inequities in admission and treatment
Biased medical training
Discriminatory data
Sampling biases & lack of representative datasets
Patterns of bias & discrimination baked into data distributions
Measurement errors, questionable source integrity, & poor data quality
Inconsistency, incompleteness, and bias in coding
Data invisibility & missingness lead to exclusion
WORLD
DATA

Real-world Patterns of Health Inequity & Discrimination

Historical inequities and the social determinants of health

- Differentials in social conditions (power, resources, exposure to health hazards, stressors, etc.) create distinct mechanisms or pathways that lead to inequities in health.

Structural biases in research agendas, medical knowledge, & clinical understanding

- Entrenched misconceptions of biological difference & discriminatory categorisations systemically bias medical knowledge & lead to knowledge gaps & neglected medical research on minoritised, marginalised & disadvantaged populations

Underrepresentation in clinical trials and genetic studies

- Research cohorts underrepresent marginalised & disadvantaged populations, preventing comprehensive understanding of health & disease & institutionalising a narrow, dominating demographic group as the standard.

Unequal access & resource allocation

- Underserved, disadvantaged, & marginalised groups, are disproportionately affected by barriers to accessing healthcare & tend to visit under-resourced clinics.

Discriminatory institutional processes

- Healthcare system policies & institutionalised clinical norms shaped by historical and structural biases, discriminate against marginalised & disadvantaged groups

Inequities in admission and treatment

- Control of resources & decision-making by dominant groups lead to disparities in admissions & treatment based on patients' sensitive or protected characteristics

Biased medical training

- Training curricula in medical education sustain discriminatory understandings of health determinants, disregard structural inequities, & focus training materials on examples that exclusively represent members of privileged & dominant groups

Implicit biases in the care environment

- Implicit biases against minoritised, marginalised & disadvantaged populations shape physician behaviour & produce differences in healthcare access and medical treatment.

Discriminatory Data

Data invisibility and missingness

- Minoritised, marginalised, & disadvantaged populations (& the categories, processes, & activities relevant to their health and wellbeing) are disregarded, neglected, or excluded from data collection & procurement processes rendering them "invisible" or "hidden"
- Observational data that are collected in distressed, under-resourced, or historically biased medical settings reflect unobservable confounders such as patients' interactions with discriminatory institutional processes, biased clinical judgements about treatment pathways, and implicit biases in the care environment

Influencing factors:

Sampling biases & lack of representative datasets

- Individuals who have inequitable access to healthcare systems, distrust clinical & research environments for reason of systemic discrimination, or have limited access to digital platforms or devices (on which data are collected) are not accurately represented in EHRs & datasets
- Minoritised, marginalised, & disadvantaged groups may be overlooked in data collection processes, which favour the interests of privileged & dominant groups, leading to issues of sampling & selection biases

Influencing factors:

Measurement errors, questionable source integrity, & poor data quality

- Under-resourced clinical environments can lack the personnel, expertise, equipment, training capacity, and digital maturity to consistently collect high quality & complete data
- Source integrity can be prejudiced by implicitly biased clinical judgments reflected in clinical notes, screenings, tests and medications ordered, and treatment decisions
- Inequitable tools & hardware designed for dominant groups can mismeasure minority groups

Influencing factors:

Inconsistency, incompleteness, and bias in coding

- Misrepresentation, misrecording, or erasure of demographic characteristics such as gender or ethnicity by biased coding systems & practices can obscure patient needs, adversely impact patient's access to appropriate screenings, diagnosis, & treatments, & prejudice datasets

Influencing factors:

Patterns of bias & discrimination baked into data distributions

- Historical inequities and structural biases can be baked into datasets in the form of discriminatory proxies or embedded prejudices that become semantically encoded in an AI model's covariates and parameters

Influencing factors:

Real-world patterns of health inequity and discrimination
Historical inequity & social determinants of health
Structural biases in research agendas, medical knowledge, & clinical understanding
Underrepresentation in clinical trials and genetic studies
Unequal access & resource allocation
Implicit biases in the care environment
Discriminatory institutional processes
Inequities in admission and treatment
Biased medical training
Discriminatory data
Sampling biases & lack of representative datasets
Patterns of bias & discrimination baked into data distributions
Measurement errors, questionable source integrity, & poor data quality
Inconsistency, incompleteness, and bias in coding
Data invisibility & missingness lead to exclusion
WORLD
DATA
DESIGN
Biased label choices & feature selection introduce discriminatory proxies
Unchecked model inferences reproduce patterns of inequity encoded in the data
System outputs are misapplied to subgroups that are under-represented in training data
Power imbalances shape research agendas & problem selection
Model developers fail to evaluate differential performance leading to disparate impacts
Biased design, development, and deployment practices

Real-world patterns of health inequity and discrimination
Historical inequity & social determinants of health
Structural biases in research agendas, medical knowledge, & clinical understanding
Underrepresentation in clinical trials and genetic studies
Unequal access & resource allocation
Implicit biases in the care environment
Discriminatory institutional processes
Inequities in admission and treatment
Biased medical training
Discriminatory data
Sampling biases & lack of representative datasets
Patterns of bias & discrimination baked into data distributions
Measurement errors, questionable source integrity, & poor data quality
Inconsistency, incompleteness, and bias in coding
Data invisibility & missingness lead to exclusion
WORLD
DATA
ECOSYSTEM
DESIGN
Enduring lack of diversity conditions the medical AI innovation ecosystem at all levels
Priority of financial gain steers AI innovation agendas away from priorities of health equity
Lack of data infrastructure & organisational incentive acts as a barrier to mobilising equity-centred AI in healthcare
Deepening digital divides entrench barriers to healthcare access
Inequalities are exacerbated by health informatics tech designed for privileged groups
Ecosystem-level inequities
Biased label choices & feature selection introduce discriminatory proxies
Unchecked model inferences reproduce patterns of inequity encoded in the data
System outputs are misapplied to subgroups that are under-represented in training data
Power imbalances shape research agendas & problem selection
Model developers fail to evaluate differential performance leading to disparate impacts
Biased design, development, and deployment practices

A classification of biases

WAITING ROOM
CHECKING

World

WORLD
Historical bias
Structural Racism
Institutional Bias

WORLD
Historical bias
Structural Racism
Institutional Bias

WORLD
DATA
Historical bias
Structural Racism
Institutional Bias

WORLD
Historical bias
Structural Racism
Institutional Bias
DATA
Admission Rate Bias
Diagnostic Access Bias
Chronological Bias
Prevalence-Incidence Bias
Representation Bias
Selection Bias
Missing Data Bias
Wrong Sample Size Bias
Data Coding Bias
Diagnostic Decision Bias

WORLD
Historical bias
Structural Racism
Institutional Bias
DATA
Admission Rate Bias
Diagnostic Access Bias
Chronological Bias
Prevalence-Incidence Bias
Representation Bias
Selection Bias
Missing Data Bias
Wrong Sample Size Bias
Data Coding Bias
Diagnostic Decision Bias

WORLD
Historical bias
Structural Racism
Institutional Bias
DATA
Admission Rate Bias
Diagnostic Access Bias
Chronological Bias
Prevalence-Incidence Bias
Representation Bias
Selection Bias
Missing Data Bias
Wrong Sample Size Bias
Data Coding Bias
Diagnostic Decision Bias
DESIGN

WORLD
Historical bias
Structural Racism
Institutional Bias
DATA
Admission Rate Bias
Diagnostic Access Bias
Chronological Bias
Prevalence-Incidence Bias
Representation Bias
Selection Bias
Missing Data Bias
Wrong Sample Size Bias
Data Coding Bias
Diagnostic Decision Bias
DESIGN
Hardware Bias
Annotation Bias
Label Choice Bias
Measurement Bias
Cohort Bias
Model Selection Bias
Evaluation Bias
Cause-Effect Bias
Implementation Bias
Automation Distrust Bias
Decision Automation Bias
Population Bias
Aggregation Bias
Reporting Bias
Race Correction Bias
Confounding
Semantic Bias
Social Determinant Blindness

World
Historical bias
Structural Racism
Institutional Bias
Data
Admission Rate Bias
Diagnostic Access Bias
Chronological Bias
Prevalence-Incidence Bias
Representation Bias
Selection Bias
Missing Data Bias
Wrong Sample Size Bias
Data Coding Bias
Diagnostic Decision Bias
Design
Hardware Bias
Annotation Bias
Label Choice Bias
Measurement Bias
Cohort Bias
Model Selection Bias
Evaluation Bias
Cause-Effect Bias
Implementation Bias
Automation Distrust Bias
Decision Automation Bias
Population Bias
Aggregation Bias
Reporting Bias
Race Correction Bias
Confounding
Semantic Bias
Social Determinant Blindness

World
Historical bias
Structural Racism
Institutional Bias
Data
Admission Rate Bias
Diagnostic Access Bias
Chronological Bias
Prevalence-Incidence Bias
Representation Bias
Selection Bias
Missing Data Bias
Wrong Sample Size Bias
Data Coding Bias
Diagnostic Decision Bias
Ecosystem
Design
Hardware Bias
Annotation Bias
Label Choice Bias
Measurement Bias
Cohort Bias
Model Selection Bias
Evaluation Bias
Cause-Effect Bias
Implementation Bias
Automation Distrust Bias
Decision Automation Bias
Population Bias
Aggregation Bias
Reporting Bias
Race Correction Bias
Confounding
Semantic Bias
Social Determinant Blindness

World
Historical bias
Structural Racism
Institutional Bias
Data
Admission Rate Bias
Diagnostic Access Bias
Chronological Bias
Prevalence-Incidence Bias
Representation Bias
Selection Bias
Missing Data Bias
Wrong Sample Size Bias
Data Coding Bias
Diagnostic Decision Bias
Ecosystem
Ecosystem Bias
Privilege Bias
Research Bias
Informed Mistrust
De-agentification Bias
Allocation Discrepancy
Law of the Instrument
McNamara Fallacy
Optimism Bias
Positive Results Bias
Biases of Rhetoric
Design
Hardware Bias
Annotation Bias
Label Choice Bias
Measurement Bias
Cohort Bias
Model Selection Bias
Evaluation Bias
Cause-Effect Bias
Implementation Bias
Automation Distrust Bias
Decision Automation Bias
Population Bias
Aggregation Bias
Reporting Bias
Race Correction Bias
Confounding
Semantic Bias
Social Determinant Blindness

World
Historical bias
Structural Racism
Institutional Bias
Data
Admission Rate Bias
Diagnostic Access Bias
Chronological Bias
Prevalence-Incidence Bias
Representation Bias
Selection Bias
Missing Data Bias
Wrong Sample Size Bias
Data Coding Bias
Diagnostic Decision Bias
Ecosystem
Ecosystem Bias
Privilege Bias
Research Bias
Informed Mistrust
De-agentification Bias
Allocation Discrepancy
Law of the Instrument
McNamara Fallacy
Optimism Bias
Positive Results Bias
Biases of Rhetoric
Design
Hardware Bias
Annotation Bias
Label Choice Bias
Measurement Bias
Cohort Bias
Model Selection Bias
Evaluation Bias
Cause-Effect Bias
Implementation Bias
Automation Distrust Bias
Decision Automation Bias
Population Bias
Aggregation Bias
Reporting Bias
Race Correction Bias
Confounding
Semantic Bias
Social Determinant Blindness

World
Historical bias
Structural Racism
Institutional Bias
Data
Admission Rate Bias
Diagnostic Access Bias
Chronological Bias
Prevalence-Incidence Bias
Representation Bias
Selection Bias
Missing Data Bias
Wrong Sample Size Bias
Data Coding Bias
Diagnostic Decision Bias
Cognition
Ecosystem
Ecosystem Bias
Privilege Bias
Research Bias
Informed Mistrust
De-agentification Bias
Allocation Discrepancy
Law of the Instrument
McNamara Fallacy
Optimism Bias
Positive Results Bias
Biases of Rhetoric
Design
Hardware Bias
Annotation Bias
Label Choice Bias
Measurement Bias
Cohort Bias
Model Selection Bias
Evaluation Bias
Cause-Effect Bias
Implementation Bias
Automation Distrust Bias
Decision Automation Bias
Population Bias
Aggregation Bias
Reporting Bias
Race Correction Bias
Confounding
Semantic Bias
Social Determinant Blindness

World
Historical bias
Structural Racism
Institutional Bias
Data
Admission Rate Bias
Diagnostic Access Bias
Chronological Bias
Prevalence-Incidence Bias
Representation Bias
Selection Bias
Missing Data Bias
Wrong Sample Size Bias
Data Coding Bias
Diagnostic Decision Bias
Cognition
Self-Assessment Bias
Confirmation Bias
Availability Bias
Naïve Realism
Ecosystem
Ecosystem Bias
Privilege Bias
Research Bias
Informed Mistrust
De-agentification Bias
Allocation Discrepancy
Law of the Instrument
McNamara Fallacy
Optimism Bias
Positive Results Bias
Biases of Rhetoric
Design
Hardware Bias
Annotation Bias
Label Choice Bias
Measurement Bias
Cohort Bias
Model Selection Bias
Evaluation Bias
Cause-Effect Bias
Implementation Bias
Automation Distrust Bias
Decision Automation Bias
Population Bias
Aggregation Bias
Reporting Bias
Race Correction Bias
Confounding
Semantic Bias
Social Determinant Blindness

Questions?

Activity 1: Bias self-assessment

Go to link [insert link for MIRO Board]

ACCOUNTABILITY AND GOVERNANCE

03

ACCOUNTABILITY

Accountability

- End-to-end governing principles across the AI lifecycle
- Humans are **answerable** for the parts they play across the entire AI system
- Results of this work are **traceable** from start to finish

Two relevant challenges for accountability

Two relevant challenges for accountability

Accountability gap

Two relevant challenges for accountability

Accountability gap

Complexity of AI production processes

Subcomponents of accountability

Answerability

- **Human creators** of AI systems should be required to **justify AI-supported decisions**
- Requires a **continuous chain of human responsibility** across the whole AI project delivery workflow
- **Explanations and justifications** must be offered by **competent human authorities in plain, understandable, and coherent language**

Subcomponents of accountability

Audibility

- Responds to **how** the designers and implementers of AI systems are to be held accountable
- Demonstrates both **the responsibility** of design, development, and deployment practices and the **justifiability of outcomes** of the AI system
- Also has to do with **traceability** - process through which all stages of AI lifecycle are documented in a way that is **accesible** and **easy to understand**

What does successful audibility look like?

Keep records and make available information that enables monitoring of the innovation processes that produced the systems

Keep track of the accountable parties within the project team and others involved in the supply chain

Keep track of the governance actions taken across the entire AI innovation workflow

Keep records and make accessible information that enables monitoring of data through all stages of the AI lifecycle (Data Factsheet)

AI systems must be designed to facilitate end-to-end answerability and auditability.

This requires both responsible humans-in-the-loop across the entire AI lifecycle as well as activity monitoring protocols that enable end-to-end oversight and review.

Types of accountability

Anticipatory accountability

Remedial accountability

Types of accountability

Anticipatory accountability

Ensuring accountability at the design and development stages of AI project - prior to the deployment of the system.

Bolsters soundness of design and implementation processes.

Ex ante accountability should be prioritised over remedial or ex-post accountability.

Types of accountability

Accountability related to the outcomes of an AI system's deployment (ex-post)

Not be a first-resort but a fallback. Strong regimes of remedial accountability however, are still crucial.

Comprehensive audibility regimes as part of overall accountability framework.

Providing explanations to affected stakeholders for the AI system's decision is part of remedial accountability.

Remedial accountability

GOVERNANCE

Putting accountability into practice

End-to-end operability of good governance practices should guide your strategy to embed accountability across the project workflow

Putting accountability into practice

End-to-end operability of good governance practices should guide your strategy to embed accountability across the project workflow

1/ Maintaining strong regimes of professional and institutional transparency

Putting accountability into practice

End-to-end operability of good governance practices should guide your strategy to embed accountability across the project workflow

1/ Maintaining strong regimes of professional and institutional transparency

2/ Establishing and maintaining a clear and accessible Process-Based Governance Framework (PBG Framework)

What is a process-based governance (PBG) framework?

Process-based governance (PBG) Framework

Template used to document governance actions established to integrate norms, values and principles within the design, development, and deployment of AI systems.

Principles and Related Governance Actions

SUSTAINABILITY

design, develop, and deploy AI systems assuring continuous sensitivity to their real-world impacts

SEP (Stakeholder Engagement Process)

Process facilitating a contextually informed understanding of the social environment and human factors that may be impacted by, or may impact, individual AI projects, and the uptake of proportionate stakeholder engagement and input throughout the AI lifecycle.

SIA (Stakeholder Impact assessment)

Process facilitating the iterative evaluation of the social impact and sustainability of individual AI projects, as well as the corroboration of these potential impacts in dialogue with stakeholders, when appropriate.

FAIRNESS

design, develop, and deploy AI systems assuring a minimum threshold of discriminatory non-harm

BSA & RM (Bias Self-Assessment & Risk Management)

Process facilitating the evaluation of how AI projects align with the principle of fairness through the iterative identification and documentation of risks of bias across the lifecycle, and assurance actions implemented to address these.

FPS (Fairness Position Statement)

Document establishing the metric-based fairness criteria for individual AI projects, providing an explanation in plain and nontechnical language.

TRANSPARENCY & EXPLAINABILITY

design, develop, and deploy AI systems assuring the ability to explain and justify their process and outcomes

Transparency & EAM (Explainability Assurance Management)

Iterative process aimed to facilitate the implementation and evaluation of transparency and explainability assurance activities across the project lifecycle and assist in providing clarification of AI system outputs to a range of impacted stakeholders.

Principles and Related Governance Actions

SAFETY

design, develop, and deploy AI systems assuring their technical accuracy, reliability, security, and robustness

SSA & RA (Safety Self-Assessment and Risk Management)

Process facilitating the evaluation of how AI projects align with safety objectives through the iterative identification and documentation of risks of potential safety risks across the lifecycle, and assurance actions implemented to address these.

RESPONSIBLE DATA MANAGEMENT

design, develop, and deploy AI systems assuring data quality, integrity, protection, and privacy

Data Factsheet

Live document facilitating the uptake of best practices for Responsible Data Management across the AI project workflow by facilitating the documentation of a comprehensive record of the data lineage iterative assessments of data integrity, quality, protection, and privacy.

ACCOUNTABILITY

design, develop, and deploy AI systems assuring end-to-end answerability and auditability

PBG Framework

Live document outlining governance actions, relevant team members and roles involved in each action, timeframes for follow-up actions, and logging protocols, for individual AI projects.

All Possible Governance Actions Across the AI Project Workflow

Establishing Proportionate Governance Actions

Just as with the determination of proportionate stakeholder involvement, the establishment of proportionate governance protocols should involve an assessment of the potential risks and hazards of the model or tool under consideration.

PBG Framework

A constructed PBG framework provides a big picture of:

- Relevant steps in the workflow in which specific actions have been established to meet governance goals

PBG Framework

A constructed PBG framework provides a big picture of:

- Relevant steps in the workflow in which specific actions have been established to meet governance goals
- Relevant team members and roles involved in each governance action

PBG Framework

A constructed PBG framework provides a big picture of:

- Relevant steps in the workflow in which specific actions have been established to meet governance goals
- Relevant team members and roles involved in each governance action
- Explicit timeframes for follow-up actions, re-assessments, and continual monitoring

PBG Framework

A constructed PBG framework provides a big picture of:

- Relevant steps in the workflow in which specific actions have been established to meet governance goals
- Relevant team members and roles involved in each governance action
- Explicit timeframes for follow-up actions, re-assessments, and continual monitoring
- Clear and well-defined protocols for logging activity and instituting mechanisms for end-to-end audibility

PBG Framework Template

- Sustainability
- Accountability
- Fairness
- Safety
- Responsible Data Management
- Transparency

1 Establishing Proportional Governance Actions

DESIGN	Project Planning	SEP	SIA	SSA & RM	BSA & RM		Data Factsheet	PBG Framework	Trans-parency & EAM
	Problem Formulation								
	Data Extraction or Procurement								
	Data Analysis						Data Factsheet		
DEVELOPMENT	Preprocessing and Feature Engineering								
	Model Selection								Trans-parency & EAM
	Model Training, Testing, and Validation		SIA	SSA & RM	BSA & RM	FPS	Data Factsheet	PBG Framework	
	Model Reporting								Trans-parency & EAM
DEPLOYMENT	Model Productionalisation								
	User Training								
	System Use and Monitoring		SIA	SSA & RM	BSA & RM		Data Factsheet	PBG Framework	Trans-parency & EAM
	Model Updating or Deprovisioning								

2 Determining Roles Involved and Timeframes

Action	Roles Involved			Timeframe
	Design	Development	Deployment	Previous activity:
SEP (Stakeholder Engagement Process)				Timeframe for revisitation:
	Design	Development	Deployment	Previous activity:
SIA (Stakeholder Impact assessment)				Timeframe for revisitation:
	Design	Development	Deployment	Previous activity:
SSA & RM (Safety Self-Assessment and Risk Management)				Timeframe for revisitation:
	Design	Development	Deployment	Previous activity:
BSS & RM (Bias Self-Assessment & Risk Management)				Timeframe for revisitation:
	Design	Development	Deployment	Previous activity:
FPS (Fairness Position Statement)				Timeframe for revisitation:
	Design	Development	Deployment	Previous activity:
Data Factsheet				Timeframe for revisitation:
	Design	Development	Deployment	Previous activity:
PBG Framework				Timeframe for revisitation:
	Design	Development	Deployment	Previous activity:
Transparency & EAM (Explainability Assurance Management)				Timeframe for revisitation:

Follow-up actions, re-assessments, and continual monitoring

3 Outlining Logging Protocols

a. How will governance actions be documented?

.....

b. How will artefacts yielded by governance activities be centralised and organised?

.....

c. Who will be responsible for logging updates to activities and artefacts?

.....

Establishing audibility with a process log

Any questions?

Activity 2: Establishing proportionate governance actions

Thank you!

See you tomorrow for Day 5: Transparency, explainability and the CARE & Act principles