Module 2

Bias, Fairness & Inclusion in AI Systems

Ethical Engineer

Co-funded by the European Union

www.ethicalengineer.eu

CONTENTS

01 Diagnose Bias in the System

02 Test Fairness in Context

03 Design Fair & Inclusive Controls

04 Make & Defend an Engineering Decision

Co-funded by the European Union

This project has been funded with support from the European Commission. The author is solely responsible for this publication (communication) and the Commission accepts no responsibility for any use may be made of the information contained therein. In compliance of the new GDPR framework, please note that the Partnership will only process your personal data in the sole interest and purpose of the project and without any prejudice to your rights.

CC BY SA

Ethical Engineer

Ethical Engineer Course at a Glance

Core content and learner output for each module

Module 01	Module 02	Module 03	Module 04	Module 05	Module 06
Ethics & Transparency	**Bias, Fairness & Inclusion**	**Privacy & Data Protection**	**Responsibility & Safety**	**Values-based Design**	**Sustainability & Impact**
Ethical principles and EU values Transparency and explainability AI and robotics case analysis OUTPUT: ethical issue map	Bias across the AI lifecycle Fairness goals and evidence Inclusive engineering cases OUTPUT: bias audit	Privacy by design and GDPR Data governance and PETs Assessment and contestability OUTPUT: privacy design	Roles and accountability Failure, resilience and handover Logging and incident learning OUTPUT: traceability record	Ethical design requirements Data, model and interface choices Trade offs and agile checkpoints OUTPUT: decision matrix	Energy, water and hardware Societal and lifecycle impacts Metrics, governance and redesign OUTPUT: sustainability plan

CAPSTONE: combine the six outputs in an ethical engineering dossier using one WP2 case or engineering project

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Module Orientation

Purpose

Equip engineering students to identify, measure and reduce unfair outcomes in AI systems while embedding inclusive team and stakeholder practices.

Audience

Higher education students in engineering, computer science, robotics and related disciplines.

Builds On

Module 1 ethical principles, transparency and engineering decision records.

[Accessibility]
Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Module Orientation

Learning Sequence

Diagnose bias, select evidence, design controls and defend an engineering decision.

Learner Outputs

Bias pathway map, fairness test plan, inclusion-by-design checklist, bias audit record and engineering decision brief.

[Accessibility]
Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Learning Outcome

By the end of the module, students will be able to:

1. Explain how problem framing, data, labels, features, models, feedback loops and team assumptions can create bias.
2. Distinguish fairness goals and select evidence suited to an engineering context.
3. Evaluate performance and impacts across relevant and intersectional groups.
4. Apply diversity and inclusion principles to team collaboration, stakeholder participation and system design.
5. Propose technical, organisational, participatory and human-review controls.
6. Document trade-offs, residual risks, monitoring and routes to challenge a decision.

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Bias becomes an Engineering Issue when it Changes Outcomes

- AI systems classify, rank, predict and allocate. These functions influence access to work, healthcare, education, finance and public services. Unfair outcomes can arise even when a model is technically accurate overall.
- The engineering question is specific: Which choice in the system creates a difference in treatment or outcome, for whom, under which conditions, and with what evidence?

A defensible response connects affected people to requirements, data, performance tests, human review, remedy and monitoring.

[Accessibility]
Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

European Frameworks Connect Fairness with Engineering Evidence

The EU Ethics Guidelines identify diversity, non-discrimination and fairness as requirements for trustworthy AI.

The AI Act and EU Charter provide the strongest legal and rights anchors: risk management, data governance, human oversight, dignity, non-discrimination and equality.

ISO 30415 is supporting guidance for inclusive organisational culture and decision-making. DigComp supports critical evaluation of digital information, data and technology.

[Accessibility]
Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

01

Diagnose Bias in the System

Trace how objectives, data, labels, models and feedback loops can shape outcomes.

Bias is a System Property

Bias is a systematic distortion that produces misleading or unfair results. It can arise from unbalanced data, historical inequality, narrow team perspectives, missing stakeholder representation or operational choices.

Data may carry historical inequalities, yet bias can also enter through the objective, target variable, label rules, feature design, threshold, interface or deployment process.

[Accessibility]
Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Bias Can Enter at Every Stage of the AI Lifecycle

Choose one engineering decision. At each stage, identify the source of bias, the affected groups and the evidence required.

01	02	03	04	05	06
Problem Framing	**Data**	**Labels & Proxies**	**Model & Thresholds**	**Interface & Human Action**	**Deployment & Feedback**
What outcome is optimised, and who defined success?	Who and which operating conditions are represented?	What is treated as success, need or risk?	Which errors and uncertainties matter, and to whom?	How is the output presented, interpreted and acted upon?	How can drift, appeals and feedback loops change outcomes?

CAPSTONE: A bias pathway map linking each stage to evidence, affected groups, diversity and inclusion considerations, and a practical control.

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Completed Bias Pathway: Self-driving Taxi Deployment

Case study:
Ethical Engineer self-driving taxis.

Decision:
Approve an autonomous taxi for a mixed urban district. Each stage records bias, affected people, evidence and control.

01 Problem Framing

- **Bias:** Fleet efficiency defines success; rare hazards receive less weight.
- **Affected:** pedestrians, cyclists, disabled passengers and emergency crews.
- **Evidence/Control:** co-define safety and accessibility; assign incident accountability.

02 Data

- **Bias:** Data favour clear weather, mapped roads and common traffic patterns.
- **Affected:** people in roadworks, low visibility and poorly mapped districts.
- **Evidence/Control:** audit coverage, add edge cases and restrict the operating domain.

Completed Bias Pathway: Self-driving Taxi Deployment

Case study:
Ethical Engineer self-driving taxis.

Decision:
Approve an autonomous taxi for a mixed urban district. Each stage records bias, affected people, evidence and control.

03 Labels & Proxies

- **Bias:** Near misses and human interventions are labelled inconsistently.
- **Affected:** road users whose behaviour or mobility is poorly detected.
- **Evidence/Control:** agree safety labels; test annotations and proxy measures.

04 Model & Thresholds

- **Bias:** Detection and stopping thresholds trade comfort against collision risk.
- **Affected:** pedestrians, cyclists and passengers in rare conditions.
- **Evidence/Control:** compare scenario errors; apply safety margins and fail-safe limits.

Completed Bias Pathway: Self-driving Taxi Deployment

Case study:
Ethical Engineer self-driving taxis.

Decision:
Approve an autonomous taxi for a mixed urban district. Each stage records bias, affected people, evidence and control.

05 Interface and human action

- **Bias:** Confidence, handover and emergency options are unclear.
- **Affected:** disabled passengers, remote operators and emergency responders.
- **Evidence/Control:** test accessible alerts; provide intervention and escalation routes.

06 Deployment and feedback

- **Bias:** Well-covered routes generate more data and appear safer.
- **Affected:** underrepresented districts, their residents and displaced drivers.
- **Evidence/Control:** monitor incidents by area; audit drift, complaints and transition impacts.

Completed Bias Pathway: Self-driving Taxi Deployment

Case study:

Ethical Engineer self-driving taxis.

Decision

Approve only within the validated operating domain, with accessible controls, remote intervention, incident accountability and area-level monitoring.

Pause operation when safety or coverage conditions are breached.

[Educator Guidance]

- Give learners two minutes to compare this completed map with their own map on the previous slide.
- Ask which control changes the design, which limits deployment and which provides remedy after harm.
- Emphasise that the final decision is conditional. Evidence defines where the system may operate and when it must stop.

[Accessibility]

- Read the six stages in order and provide the linked case study in an accessible digital format.

Sampling, Measurement & Label Choices Create Different Risks

Sampling bias occurs when relevant people or operating conditions are missing. Historical bias appears when past inequalities are treated as neutral evidence.

Measurement bias occurs when a variable records groups differently. Label bias occurs when the target reflects inconsistent judgement or an unequal process.

[Accessibility]
Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Proxy Variables Can Hide Value Choices

- A proxy is used when the desired concept cannot be measured directly. Cost may stand for health need. Previous hiring may stand for future performance. Arrest may stand for offending.
- A proxy can be predictive while creating unfair outcomes because social conditions influence it. Engineers should ask what the proxy represents for each group, which causal factors shape it and whether the intended decision can be justified.

Document the proxy, its limits and an alternative measure.

[Accessibility]
Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Feedback Loops can turn an Initial Bias into a System Pattern

- AI outputs often change the data collected next. Extra policing creates more recorded incidents. More inspections reveal more faults. Rejected applicants produce no later performance data.
- A feedback loop can make the model appear increasingly certain while narrowing what the system can observe.

Engineering controls include exploration samples, independent data sources, counterfactual review, monitoring of selection rates and a route for people to supply missing evidence.

[Accessibility]
Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Inclusion Improves the Problem Definition

- Inclusion means that affected people can influence requirements, evaluation and remedies. Representation in a team matters, while participation in decisions is the stronger engineering control.
- Diversity management means the planned way an organisation recognises human differences and gives varied perspectives influence in work and decisions. Systemic diversity management makes this visible in governance: who is consulted, who can question assumptions, who can request evidence and who can challenge a harmful objective.
- Identify direct users, people represented in the data, people subject to decisions, operators, maintainers and communities affected indirectly. Ask whose experience is missing and who can challenge the proposed objective.
- A psychologically safe team makes it possible to report uncertainty, disagreement and potential harm early.

[Accessibility]
Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Clinical Practice Example: Non-Inclusive Cardiac Pump Design

Case Study: Ethical Engineer: pulsatile heart pump.

Initial Task: : Define the group of patients who may benefit from implantation of a VAD with a fixed stroke volume of 65 mL. Propose solutions for patients of different body sizes who fall outside the safe operating range of this pump.

01 Initial Assignment

Confirm that the standard device is suitable for the individual patient and meets the principles of beneficence and non-maleficence.

02 Interdisciplinary Team

surgeon, anesthesiologist, perfusionist, biomedical engineer, ethics consultant

03 What the Non-Inclusive Pump Design Revealed

Body proportions differ among patients of different sizes. A fixed stroke volume may lead to arterial hypertension in smaller patients, whereas in larger patients it may cause hypotension and inadequate organ perfusion.

Clinical Practice Example: Non-Inclusive Cardiac Pump Design

Case Study: Ethical Engineer: pulsatile heart pump.

Initial Task: Define the group of patients who may benefit from implantation of a VAD with a fixed stroke volume of 65 mL. Propose solutions for patients of different body sizes who fall outside the safe operating range of this pump.

04 Reframing the Problem

Define the safe operating range of the pump: identify which patients may benefit from mechanical circulatory support (65 mL VAD) and for which patients it is unsuitable.

05 Technical Requirements

For patients within the safe operating range, use the VAD with the standard recommended settings. In patients who develop hypertension, adjust the device settings to achieve optimal arterial pressure.

06 Decision & Review

Use the 65 mL VAD with standard settings only if arterial pressure monitoring confirms values within the normotensive range. If arterial hypertension is detected, define the subsequent management and the physician's responsibilities, including adjustment of the VAD settings, continued patient monitoring, and assessment of residual risk.

Ethical Engineer

Clinical Practice Example: Non-Inclusive Cardiac Pump Design

Case Study: Ethical Engineer: pulsatile heart pump.

Inclusive approach changed the pump settinngs

A standard pump implantation evolved into complex patient-safety decision-making that required deviation from the recommended pump settings in order to uphold the principles of beneficence and non-maleficence and save the patient's life.

[Educator Guidance]

- Ask learners to identify the exact point at which inclusion changes the engineering brief.
- Invite one group to argue for the standard device, one for adaptation and one for postponement. Each group must state the evidence and residual risk.
- Emphasise that participation produces requirements and decision criteria. It is an engineering control, not a consultation exercise added at the end.

[Accessibility]

- Read each row aloud and provide the linked heart-pump case in an accessible digital format.

02

Test Fairness in Context

Select fairness goals, measures, evidence and acceptance criteria for the engineering context.

Fairness Links System Choices to Unequal Effects

Fairness concerns how an AI system distributes benefits, risks, errors, delays and opportunities across affected people. An overall accuracy score can hide serious differences between groups or operating conditions.

Fairness is therefore both a technical design objective and a human-rights commitment linked to dignity, non-discrimination and equality. Unfairness can enter through the objective, data, labels, proxies, thresholds, interface, human action or deployment process. The engineering question is whether a system choice creates an unjustified disadvantage.

[Accessibility]
Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Fairness Links System Choices to Unequal Effects

01

Fairness Goal

State which outcome, opportunity or error must be distributed fairly, and explain why it matters in this decision.

02

Evidence

Test performance and impact across relevant groups and conditions. Report group counts, uncertainty and data limitations.

03

Engineering Response

Set acceptance criteria, technical and organisational controls, human review, monitoring and a route to remedy.

Fairness goals can conflict. Engineers must justify the selected goal in relation to the decision, affected people and consequences of error. Common goals include equal access, equal opportunity, comparable error rates, calibration, individual consistency and a fair process.

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

A Fairness Metric is Evidence, not the Decision

Report absolute results, uncertainty and group counts. Examine intersectional groups where evidence is sufficient. Combine quantitative tests with qualitative evidence about process, dignity, inclusion, accessibility and access to remedy.

Begin with the Decision & the Affected Population

The evaluation plan should reflect realistic conditions, including missing data, sensor quality, language, disability, environmental variation and whether affected stakeholders were represented in requirements gathering.

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Disaggregate Performance & Show Uncertainty

- Overall accuracy can hide weak performance for a smaller group. Report confusion-matrix measures, calibration or task-specific quality by relevant group and operating condition.
- Include sample size, confidence interval or another uncertainty statement. Where evidence is too limited, record the gap and restrict deployment rather than presenting unstable results as proof of fairness.

Repeat tests after model, data or threshold changes and during operation.

[Accessibility]
Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Intersectionality Changes the Test Design

- People experience systems through several characteristics and circumstances at once. A model may perform adequately for women overall and for a racial group overall, while performing poorly for women within that group.

- Intersectional analysis needs sufficient data and careful privacy protection. Combine targeted quantitative testing with stakeholder consultation and scenario analysis.

Record which intersections were evaluated, which remain uncertain and how monitoring will detect harm after deployment.

[Accessibility]
Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Thresholds Allocate Errors & Benefits

- A score does not determine an outcome until a threshold and action rule are chosen. Moving the threshold changes false positives, false negatives, workload and access.
- Compare candidate thresholds using affected-group results and the real consequence of each error. Consider a review band where uncertain cases receive human assessment.

The decision record should state the selected threshold, acceptance criteria, review route, owner and trigger for change.

[Accessibility]
Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Project Case: AI in Hiring Processes

Engineering focus: historical data, target definition, proxy features, subgroup testing, explanation, contestability and diversity-aware validation of job requirements.

Analyse the Hiring Case as a System

Test whether historical hiring is a valid label for capability. Examine school, career-gap and language features as possible proxies. Compare selection and error rates across relevant groups. Review whether recruiters can override the score, whether diverse stakeholders helped define the job requirement and whether candidates can challenge an error.

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Project Case: AI in Facial Recognition

Open the Ethical Engineer facial recognition case study and define the relevant error for the chosen use.

- The project case examines facial recognition used to improve security and streamline processes such as airport boarding. The benefits depend on reliable identity matching, while privacy, data security and demographic performance create significant risks.
- **Engineering focus**: use context, image quality, one-to-one or one-to-many matching, error consequences, demographic tests and human review.

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

NIST Evidence Changes the Facial Recognition Decision

- NIST tested large numbers of algorithms and found demographic differentials in the majority of those evaluated, with results depending on algorithm, application and data. This means that a general claim about facial recognition is weak evidence.

- Specify the algorithm version, operating mode, image conditions and population. Test the error linked to the harm. A false match in surveillance differs from a failed verification at a gate.

Add human review, quality checks, appeal and incident monitoring

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Applied Task: Write a Fairness Test Plan

- Select the hiring or facial recognition case. Produce a plan that another engineering team could execute.
- Define the decision, population, groups, conditions and error consequences. Select two quantitative measures and one qualitative source. State data-quality checks, minimum sample requirements, uncertainty reporting and acceptance criteria.
- Add a decision rule for insufficient evidence, an owner, a review date and a route for affected people to challenge an outcome.
- Add a 5-point Inclusion-by-Design checklist covering diverse data, stakeholder input, accessibility, documentation, testing and deployment.
- **Output:** one-page test plan, checklist and two-minute justification.

[Accessibility] Offer the task in written, spoken or visual form. Read the scenario aloud, explain acronyms and allow accessible digital completion.

03

Design Fair & Inclusive Controls

Translate findings into technical, organisational, participatory and monitoring controls.

Start Mitigation Before Model Training

Before modelling, draft an Inclusion-by-Design checklist. Document data origin, purpose, consent or legal basis, known gaps, intended use and representation limits. Include stakeholder review where the target, proxy or intended use may encode social disadvantage.

Use sampling and weighting only when they address the identified mechanism.

[Accessibility]
- Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Training & Post-processing Controls Need a Stated Fairness Goal

- Model constraints, reweighting, resampling, representation learning and threshold adjustment can change group outcomes. Their suitability depends on the selected fairness goal and risk context.

- Compare performance, fairness, uncertainty and robustness after each change. Check whether gains for one group create unacceptable losses for another or weaken safety.

Record the trade-off and keep a baseline so that the effect of the intervention remains visible.

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Human Review must be Designed as a Real Control

Design the review trigger, evidence display, decision options, escalation route and record. Test reviewer consistency and workload. Oversight teams should have relevant technical, domain and lived-experience perspectives. Give affected people clear notice and an accessible route to correction or appeal.

Diverse Teams Need Decision Rights & Evidence

Team diversity can reveal blind spots, yet representation alone does not guarantee influence. ISO 30415 frames inclusion as culture, equity, respect and accountability; engineering practice must translate this into decision rights. Define who can question the objective, pause deployment, request subgroup evidence and approve residual risk.

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

External Case: Rite Aid Facial Recognition

- The United States Federal Trade Commission alleged that Rite Aid used facial recognition without reasonable procedures to prevent consumer harm. The complaint described false-positive alerts and failures in testing, image quality, staff training, monitoring and risk assessment.
- **Engineering lesson:** subgroup performance is one part of a control system. Image quality, database rules, operator action, escalation and incident evidence also shape harm.

Ask which control should have prevented deployment and which evidence should have triggered a stop.

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Turn the Rite Aid Lessons into Controls

A fairness problem rarely has a single technical fix.

The Rite Aid case shows how harm can result from several connected failures - poor input quality, weak matching rules, operator behaviour, inadequate training and missing governance. Effective controls therefore need to operate across the whole system.

Image Quality

Reject inputs that fall below a verified quality threshold.

Match Evidence

Show match confidence, image-quality limits and alternative candidates where relevant.

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Turn the Rite Aid Lessons into Controls

Operator Action

Require independent review before any adverse action. An alert alone must never determine the outcome

Training

Verify staff competence through realistic scenarios and refresher checks.

Monitoring

Track false alerts, affected groups, complaints and emerging patterns of harm.

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Turn the Rite Aid Lessons into Controls

Governance

Require approval before expanding use and define clear suspension triggers.

Engineering Principle

Every control needs a measurable acceptance criterion, an accountable owner and evidence that it is working

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Dilemma: Fair Allocation Under Missing Data

Decision

GO, REVISE or STOP. Define a fair rule for uncertain cases, minimum evidence, human review, accessible appeal, monitoring, accountable owner and how mature, international or interrupted-study students are represented in the review process.

[Accessibility] Offer the task in written, spoken or visual form. Read the scenario aloud, explain acronyms and allow accessible digital completion.

04

Make & Defend an Engineering Decision

Complete the bias audit, record residual risk and defend a decision using evidence.

Project Case A: Bias in Data

- The Ethical Engineer case shows how unrepresentative or historically biased data can produce inaccurate predictions and unequal outcomes. Use it to identify the specific data-generating process rather than treating bias as a generic property.

- Define who is missing, which variable is distorted and how that affects the decision. Propose one data control, one system control and one monitoring measure.

Open the Bias In Data Case Study and record the residual uncertainty.

[Accessibility] - Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Project Case B: AI in Historic Photo Assessment

- This case considers an AI system that predicts which historic images may attract attention or revenue. Internet-derived data can reward popularity while stripping sensitive images of historical context.

- The engineering dilemma concerns the objective, training signal, representation of communities and the interface used by curators.

Open the Historic Photo Assessment Case. Decide how cultural context, stakeholder review and limits on automated ranking should change the design.

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Compare the Cases Before Choosing a Mitigation

Match the control to the bias mechanism, not to the label “bias”.

- **Bias in Data begins with evidence quality:** who is missing, which variable is distorted and how that affects the decision.
- **Historic photo assessment adds value conflict:** popularity, revenue and cultural meaning may point in different directions.

Ask four questions before selecting a control:

- What system choice creates the disadvantage?
- What evidence would show whether the harm is real?
- Which control changes the mechanism: target, data, threshold, interface, review or deployment limit?
- What residual risk remains after the control?

[Accessibility] - Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Team Activity: Complete a Bias Audit

Choose one Ethical Engineer case. Assign systems, data, testing, stakeholder and recorder roles.

1. Define intended use and the critical decision.
2. Map the bias pathway.
3. Choose a fairness goal and evidence.
4. Propose technical and organisational controls.
5. State missing evidence, residual risk and remedy.
6. Decide GO, REVISE or STOP.

Output

Completed audit record and a two-minute evidence-based decision brief.

[Accessibility] - Offer the task in written, spoken or visual form. Read the scenario aloud, explain acronyms and allow accessible digital completion.

From Analysis to a Defensible Engineering Decision

A strong bias audit turns evidence into an accountable argument

01	02	03	04	05
Claim	**Evidence**	**Controls**	**Residual Risk**	**Accountability**
What are you proposing: GO, REVISE or STOP?	Which subgroup tests, stakeholder evidence, operational evidence and uncertainty support the claim?	What changes the design, limits deployment, improves review or provides remedy?	What uncertainty or potential harm remains after controls?	Who owns the decision, what is monitored, what triggers review and how can affected people challenge it?

[Accessibility] - Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Worksheet: Bias And Fairness Decision Record

System & Intended Use: .....

Critical Decision & Affected People: .....

Bias Mechanism / Pathway: .....

Fairness Goal & Why It Fits: .....

Evidence, Groups, Conditions & Uncertainty: .....

Inclusion / Stakeholder Evidence: .....

[Accessibility] Offer the task in written, spoken or visual form. Read the scenario aloud, explain acronyms and allow accessible digital completion.

Continues On Next Slide

Technical Control & Acceptance Test: .....

Governance / Inclusion Control & Owner: .....

Decision: GO REVISE STOP .....

Residual Risk & Monitoring Trigger: .....

Remedy / Appeal Route: .....

Dissent Or Alternative View: .....

[Accessibility] Offer the task in written, spoken or visual form. Read the scenario aloud, explain acronyms and allow accessible digital completion.

Role Play: Procurement Review Panel

- A city plans to buy an AI system that prioritises housing inspections. The vendor provides overall accuracy but no subgroup or neighbourhood results. Historical complaint data are sparse in areas with language barriers and low trust.
- **Roles:** procurement lead, housing engineer, vendor, community representative, data protection officer and safety lead.

Agree the evidence and contract conditions required before a pilot. State one condition for REVISE and one for STOP. Record who can suspend use and how residents can challenge an omission.

[Accessibility] Offer the task in written, spoken or visual form. Read the scenario aloud, explain acronyms and allow accessible digital completion.

Individual Reflection & Transfer

Write a short engineering argument.

1. Which system choice created the strongest fairness or inclusion concern?
2. Which fairness goal did you select, and what does it leave unresolved?
3. Whose perspective or experience was missing from the original evidence?
4. Which new evidence would cause you to change your decision?
5. Which technical, organisational and inclusion controls belong in the requirements?
6. What residual risk would you accept, and what would make you stop deployment?
7. How would your reasoning change in another engineering discipline?

Output:

Connect affected people, evidence, controls and a justified decision.

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Quiz

1. Why can high overall accuracy coexist with unfair outcomes?
2. What is the difference between sampling bias and label bias?
3. Why must a fairness metric be linked to a use context?
4. What should an engineer report when subgroup evidence is limited?
5. Name two conditions required for meaningful human review.
6. Which control addresses an invalid proxy variable most directly?
7. What belongs in a fairness decision record?
8. How can diversity and inclusion improve the engineering process?

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Quiz Answers & Teaching Rationale

1. Overall results can hide group- and condition-specific errors.
2. Sampling bias concerns representation; label bias concerns the target or judgement recorded.
3. Metrics express different fairness goals and error consequences.
4. Report the gap and uncertainty, then restrict or revise use.
5. Authority, time, information, competence and a record are relevant.
6. Reconsider the target, feature or process that created the proxy.
7. Decision, affected people, goal, evidence, controls, owner, residual risk, monitoring and remedy.
8. By widening evidence, revealing assumptions, strengthening review and giving affected people routes to challenge decisions.

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Learning Outcome Check

Before finishing, confirm that you can:

1. Trace bias through an AI system and its decision context.
2. Select and justify a fairness goal.
3. Evaluate performance across relevant and intersectional groups, including uncertainty.
4. Explain how diversity, participation and inclusion change requirements and evidence.
5. Recommend technical, organisational, participatory and human-review controls.
6. Defend a GO / REVISE / STOP decision with residual risk, monitoring and remedy

Evidence Produced:

- Bias Pathway Map
- Fairness Test Plan
- Inclusion-by-design Checklist
- Bias Audit
- Engineering Decision Brief.

[Accessibility] Use accessible reading order, describe relevant visual branding on request and provide the slide text in an accessible digital format.

Funding & Educational Use Notice

This open educational resource supports teaching and discussion. It does not constitute legal, regulatory, safety or professional engineering advice. Use current law, standards, institutional policies and domain-specific guidance when applying its content.

Co-funded by the Erasmus+ Programme of the European Union.

Ethical Engineer

Thank You

You have completed **Module 2**

www.ethicalengineer.eu

Co-funded by the European Union