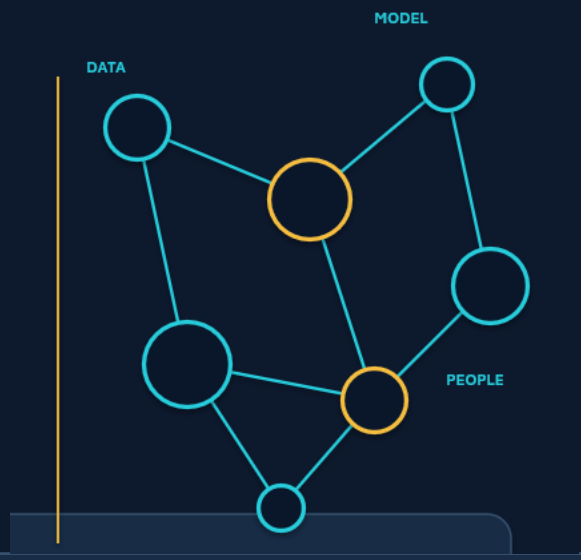
Designing Trustworthy AI

Understanding bias, human oversight and high-risk AI

Walpola Layantha Perera (Peri)

World Usability Day Leipzig 2026 · 12 November

Presentation · 10:50 -11:20 programme slot



EU AI ACT: FINANCIAL PENALTIES

- €35 million or 7% - prohibited AI practices.
- €15 million or 3% - breaches of other applicable AI Act obligations.
- €7.5 million or 1% - providing misleading or incorrect information to authorities.

The percentage is based on worldwide annual turnover. The maximum depends on the type of violation and the organisation's circumstances.
Key message: Responsible AI is not only an ethical responsibility; it is also a legal and financial responsibility.
Official references: EU AI Act Service Desk - Article 99: Penalties · European Commission - Navigating the AI Act.

THEMES

Fair data • Meaningful oversight • Accountable organisations

Fair data • Meaningful oversight • Accountable organisations

AI decisions can change real lives

OPPORTUNITY

AI can support analysis, consistency, accessibility and faster workflows.

RISK

Errors, bias or over-reliance can affect opportunity, safety and rights.

CORE QUESTION

Who is affected, who is accountable, and who can challenge the outcome?

Trustworthy AI is measured by what happens to people not only by model accuracy.

AI does not remove human responsibility

THE HUMAN ROLE IS CHANGING

Humans increasingly become:

- Orchestrators
- Supervisors
- Validators
- Strategic decision makers for AI systems

A RESPONSIBLE COLLABORATION MODEL

AI generates → Humans evaluate

AI automates → Humans govern

AI predicts → Humans decide what to do

Humans remain responsible for the decisions they own. This requires AI literacy, ethical awareness, organisational change management and a new professional mindset.

Bias can enter at every stage of the AI lifecycle

Bias is socio technical: it is not simply a flaw hidden inside an algorithm.

PROBLEM

Whose problem is defined? What counts as success?

DATA

Who is represented? Who is missing? What do labels encode?

MODEL

Which features, proxies, objectives and thresholds are selected?

USE

How do interfaces, people and workflows shape decisions?

FEEDBACK

Do outcomes reinforce patterns in future data?

Before training: prepare data responsibly

1 · PURPOSE & PROVENANCE

Define intended use and target population.

Record data source, collection context and permissions.

Check whether personal data use is lawful.

Document assumptions and limitations.

2 · QUALITY & REPRESENTATION

Check coverage and missing groups.

Validate labels and annotation rules.

Assess errors, duplicates and missingness.

Preserve relevant context and rare cases.

3 · PROTECTION & TRACEABILITY

Minimise personal data; protect access.

Separate train / validation / test sets.

Prevent leakage and record transformations.

Keep dataset versions and decision logs.

Ethical preparation means fit-for-purpose, lawful, documented, representative and protected data.

Benefits of getting the data right

BETTER PERFORMANCE

More reliable evaluation in the real target context; fewer surprises after deployment.

FAIRER OUTCOMES

Earlier visibility of group-level gaps and a better chance to mitigate avoidable harm.

TRUST & ACCOUNTABILITY

Traceable decisions, stronger evidence, easier audits and clearer explanations of limitations.

Good data governance does not guarantee fairness it makes risks visible and manageable.

A practical bias detection loop

1 · PROFILE

Inspect coverage, missingness, labels and provenance.

2 · MEASURE

Compare error rates and outcomes across relevant groups.

3 · INVESTIGATE

Find causes: sampling, proxies, workflow or measurement.

4 · MITIGATE

Improve data, design, thresholds or process; document trade offs.

5 · MONITOR

Track drift, complaints, appeals and changing impacts.

Choose metrics with domain experts: fairness is context dependent; no single metric settles it.

Meaningful oversight: notice, understand, intervene

NOTICE

Recognise uncertainty, anomalies, limitations and unexpected performance.

UNDERSTAND

Interpret outputs in context; know what the system can and cannot do.

INTERVENE

Disregard, override, reverse, escalate or stop the system safely.

A human reviewer needs competence, training, time, usable information and real authority not just a button to click.

How to make oversight real and avoid theatre

DO ✓

Assign a named role with authority.

Train people on limitations and automation bias.

Show confidence / uncertainty where meaningful.

Allow override, escalation and safe stop.

Record decisions and review incidents.

Match oversight to risk and context.

DON'T ✕

Don't use a human rubber stamp after the AI decides.

Don't overload reviewers with too many alerts.

Don't hide uncertainty or known failure modes.

Don't penalise staff for appropriate overrides.

Don't assume a human can catch every error.

Don't treat oversight as a substitute for system safety.

Oversight is effective only when the organisation enables people to act.

What makes an AI system high-risk?

Its intended purpose and use under the EU AI Act not simply the type of data it processes. Classification follows two routes: Annex I (regulated product/safety component) and Annex III (specified high-impact uses).

High-risk application Examples of data used

Recruitment: CVs, qualifications, work history, interview recordings.

Education: Grades, exam answers, learning activity, admission records.

Credit & essential services: Income, debts, credit history, eligibility records.

Biometrics & identity : Face images, voice, fingerprints, iris scans.

Traffic & infrastructure: Camera feeds, GPS, vehicle sensors, incident data.

Law enforcement: Case files, identity documents, travel and biometric records.

HOW MODELS LEARN

Training data teaches patterns; validation data tunes the model; test data evaluates performance. Live input data is used after deployment and is not automatically training data.

EU AI Act Article 10: High-risk AI training datasets must be governed for relevance, representativeness, quality and potential bias. Organisations must examine data origins, identify gaps and assess and mitigate harmful bias.

UX responsibility: Ask whose data is represented or missing, whether historical discrimination is reproduced, and whether people can understand, challenge or correct AI supported decisions.

Key message: Data quality shapes AI outcomes; human oversight and responsible design help protect people.,

High-risk is a legal classification, not a synonym for “AI that seems important.”

Core requirements: design, evidence and control

RISK MANAGEMENT

Identify and reduce foreseeable risks throughout lifecycle.

DATA GOVERNANCE

Assess data quality, suitability, representativeness and bias.

TECHNICAL FILE & LOGS

Document system, traceability and relevant events.

TRANSPARENCY

Give deployers clear instructions, capabilities and limits.

HUMAN OVERSIGHT

Enable trained people to monitor and intervene.

ACCURACY & SECURITY

Test robustness, accuracy and cybersecurity.

Provider and deployer obligations differ; identify your role in the AI value chain.

Who does what? Trustworthy AI is cross functional

Make AI literacy part of normal work.

LEADERSHIP

Accountability, resources, risk appetite and culture.

PRODUCT & UX

Purpose, affected users, usability and contestability.

DATA & ENGINEERING

Data quality, testing, security, documentation and monitoring.

LEGAL / COMPLIANCE

Applicable duties, impact assessment and evidence.

OPERATIONS

Training, human oversight, incident handling and feedback.

Make AI literacy, ethical awareness and change management part of normal work.

A six step organisational playbook

1 · CLASSIFY

Define use, users, impact and AI Act category.

2 · GOVERN

Assign owners, roles, resources and escalation.

3 · PREPARE

Document sources, quality, lawful use and bias controls.

4 · VALIDATE

Test performance, subgroup impacts and realistic scenarios.

5 · DEPLOY SAFELY

Train users; communicate limits; enable intervention.

6 · MONITOR

Track drift, incidents, complaints and corrective actions.

Keep evidence at every step: decisions, tests, owners, mitigations and reviews.

Innovation and responsibility must move together

PREPARE DATA ETHICALLY

Fit for purpose. Traceable. Protected. Tested for gaps and bias.

MAKE OVERSIGHT MEANINGFUL

Humans notice, understand, validate and intervene.

GOVERN THE WHOLE SYSTEM

Clear accountability, legal compliance and continuous monitoring.

The future of work is not only about using AI tools. It is about learning how humans and AI collaborate responsibly.

Why UX belongs in the AI conversation

- Recruitment and workplace tools
- Education and access to services
- Healthcare and wellbeing
- Mobility, transport and public infrastructure

The interface shapes how people understand, trust and act on AI outputs.

Design the human AI relationship

01 · EXPECTATIONS

Make it clear when AI is involved and what role it plays.

02 · UNCERTAINTY

When designing an AI system, users should be able to understand what the AI can do, what it cannot do, and when its answer may not be reliable.

03 · JUDGEMENT

Support informed human evaluation rather than automatic acceptance.

04 · AGENCY

Giving people meaningful control over AI-supported decisions. Users should not be forced to accept an AI decision without having a way to question it, correct mistakes or involve a human.

Design for calibrated trust not maximum trust.

Set expectations before the decision

DO

State when AI is involved and its role.

Explain the task and intended use.

Clarify whether output is advice, ranking or a decision.

Identify inputs that influence the result.

AVOID

Vague “AI powered” claims.

Human like certainty or anthropomorphic claims.

Hiding whether a person reviews the output.

Users should not have to guess what the AI is doing or how much authority its output has.

Make uncertainty and limitations visible

DESIGN FOR

Meaningful uncertainty cues.

Missing or low quality input warnings.

Out of scope and unsupported cases.

Relevant sources or context where appropriate.

AVOID

False precision.

Opaque scores with no interpretation.

Decorative explanations that do not help decisions.

Confidence percentages without context.

A clear “not enough information” can be safer than a confident looking answer.

Design against automation bias

RISKY PATTERNS

AI result is preselected by default.

Human review is rushed or hidden.

Alternative evidence is hard to find.

Overrides require excessive effort.

The interface implies certainty or authority.

BETTER PATTERNS

Enable independent review.

Show alternatives and relevant evidence.

Make challenge and override straightforward.

Support escalation and safe fallback.

Review workload and alert fatigue.

Thank you for your attention. Any questions or comments?