ETHICAL PRINCIPLES & REAL-WORLD RISKS IN AI

AI has impact. Ethics gives it direction.

AI for Change
FOUNDATION

Why Ethics Must Come First

- AI systems shape real-world outcomes in hiring, credit, education, healthcare, and policing.
- Without ethical guardrails, algorithms can amplify bias, exacerbate inequality and conceal harmful logic.
- Ethics isn't idealism. It is infrastructure for trust and safety.

"84% of AI professionals report bias as a serious concern in model deployment" - IBM Global AI Adoption Index, 2023

AI for Change
FOUNDATION

The Core Ethical Principles in AI

Responsible AI development is built on four foundational principles:

- **Fairness:** Ensure AI systems do not reinforce or exacerbate societal inequities.
- **Transparency:** Ensure system logic and data sources are clearly understood by stakeholders.
- **Accountability:** Define who is responsible for system outcomes - legally, ethically, and operationally.
- **Human-Centeredness** - Prioritise human rights, dignity, and oversight throughout the AI lifecycle.

AI for Change
FOUNDATION

Bias in AI Systems - From Code to Consequence

Bias doesn't begin in the code - it begins in the world. Once in AI, it can automate and amplify real-world inequities.

Type of Bias	Real- World Example	Consequence
Data Bias	Amazon's AI hiring tool penalised women - trained on male- dominant resumes - USA (Reuters, 2018)	Gender discrimination in hiring practices
Labeling Bias	Danish AI welfare systems disproportionately targeted vulnerable groups including people with disabilities and those with foreign affiliations due to biased assumptions in data and annotations. (Amnesty International, 2024)	Disproportionate targeting and harm to low- income families
Feedback Loop Bias	Dutch "SyRI" system flagged migrant neighbourhoods for welfare fraud - Netherlands (SyRI ruling, 2020)	Systemic profiling and human rights violations

AI for Change
FOUNDATION

Mitigating AI Bias

Strategy	Description & Tools
• Pre-deployment Audits	Conduct audits on training data before model development to identify skewed patterns. 🛠️ *Tool: IBM AI Fairness 360*
• Diverse Annotation Teams	Use annotators from different demographics to reduce labeling bias. *Tip: Include inclusion KPIs in your data pipeline*
• Fairness-Enhancing Algorithms	Use techniques like: • Re-weighting • Adversarial debiasing • Fairness constraints 🛠️ *Tool: Microsoft Fairlearn*
• What-if & Sensitivity Testing	Test how model output changes across different demographics. 🛠️ *Tool: Google's What-If Tool*
• Post-deployment Monitoring	Continuously track system outputs for signs of emerging bias or drift. 🛠️ *Tool: Custom dashboards / Alerting systems*

AI for Change
FOUNDATION

Explainability ≠ Transparency

Why Both Matter

A system that's transparent but not explainable is like an open book in a language no one reads. We need both clarity of construction and clarity of outcomes.

Explainability	Transparency
How the AI *made* a decision	How the AI *was built*
Helps users, auditors, and regulators understand outcomes	Involves revealing data sources, logic and model design
Tools: SHAP, LIME, saliency maps	Tools: Model cards, data datasheets, documentation
Needed for fairness, safety and trust	Required for oversight and accountability

AI for Change
FOUNDATION

Case Study: Healthcare Risk Algorithm Bias

Content:

A widely used US hospital algorithm underestimated the risk of Black patients needing extra care.

Why?

The model used past healthcare spending as a proxy for need, but due to structural inequalities, Black patients had lower historical spending despite greater medical need.

Result:

Systematically under-prioritised Black patients.
Bias is hidden in proxy variables, not race itself.

Lesson:

Explainability is essential: models can be accurate and still discriminatory.

AI for Change
FOUNDATION

Build Trust with the Right Tools

Tool	Purpose
SHAP / LIME	Model explainability
Model Cards (Google)	Transparency around model usage
Datasheets for Datasets (Gebru et al.)	Data documentation
Fairlearn (Microsoft)	Bias mitigation + fairness dashboard
AI Fairness 360 (IBM)	Bias detection & mitigation toolkit
What-If-Tool (Google)	Interactive sensitivity testing

Ethics as Infrastructure - Not a Plug- In

"We can't fix unethical AI after deployment. Ethics must be part of the architecture, not the apology."
- *Marina, AI for Change Foundation*

AI for Change
FOUNDATION