📄 **AI Ethics**

Theme: Designing Responsible and Fair AI Systems

✅ **Part 1: Theoretical Understanding**

Q1: What is Algorithmic Bias? Give Two Examples.

Algorithmic bias occurs when an AI system systematically produces unfair outcomes, often due to biased training data or flawed model design. It reflects historical or social inequalities present in the data.

Example 1:
An AI recruiting tool that penalizes resumes with women's names or female-coded language because the training data mostly featured successful male candidates.

Example 2:
A facial recognition system that misidentifies darker-skinned individuals more frequently due to underrepresentation in the training dataset.

Q2: Transparency vs Explainability in AI

- **Transparency** refers to how open and understandable the AI system's design and development process is — including datasets, algorithms, and decision-making processes.
- **Explainability** is the ability to explain how the AI arrived at a specific decision in a way humans can understand.

Importance:
Both are crucial for building trust. Transparency ensures accountability, while explainability allows users to challenge or understand outcomes, especially in high-stakes domains like healthcare or law.

Q3: GDPR and AI in the EU

The **General Data Protection Regulation (GDPR)** mandates:

- Right to explanation (Article 22)
- Data minimization
- Explicit consent
- Fair and lawful data processing

Impact:
AI systems must be explainable, non-discriminatory, and privacy-conscious. It influences how data is collected, stored, and how automated decisions are justified.

Ethical Principles Matching

Principle	Definition
A) Justice	Fair distribution of AI benefits and risks
B) Non-maleficence	Ensuring AI does not harm individuals or society
C) Autonomy	Respecting users' right to control their data and decisions
D) Sustainability	Designing AI to be environmentally friendly

✅ Part 2: Case Study Analysis

Case 1: Amazon's Biased Hiring Tool

Source of Bias:
The tool was trained on resumes submitted over a 10-year period—mostly from men—leading to biased weightings against female applicants and gendered terms.

Three Fixes:

1. **Use balanced training data:** Include resumes from diverse candidates.
2. **Remove gendered features:** Strip out words and proxies that signal gender.
3. **Human-AI collaboration:** Use the AI to assist, not replace, human recruiters.

Fairness Metrics:

- Disparate Impact Ratio
- Equal Opportunity Difference
- Demographic Parity

Case 2: Facial Recognition in Policing

Ethical Risks:

- **Wrongful arrests** due to misidentification
- **Privacy invasion** of innocent individuals
- **Disproportionate targeting** of minorities

Policies for Responsible Deployment:

- Ban real-time use in public spaces
- Require transparency, audits, and human oversight
- Implement strict use-cases and sunset clauses
- Train on balanced datasets

✅ Part 3: Audit Report Summary (COMPAS Dataset)

The **COMPAS Recidivism Dataset** was analyzed using IBM's **AI Fairness 360 toolkit** to investigate racial bias in risk assessments.

Key Findings:

- **Disparate False Positive Rate:** Black individuals were more likely to be incorrectly predicted as high-risk than white individuals.
- **Disparate Impact Ratio** and **Equal Opportunity Difference** values indicated clear imbalances in the predictions.

Visualization Results:
Bar charts showed higher false positive rates for Black individuals compared to White individuals, raising concerns about fairness and systemic bias.

Remediation Strategies:

- **Reweighing:** Adjust sample weights for fairness during training.
- **Preprocessing filters:** Remove or mask racial proxy variables.
- **Post-hoc explanation tools:** Provide context for risk scores.

This highlights the need for **ethical model development** and continuous monitoring to reduce harm and promote justice.

✅ Part 4: Ethical Reflection

In a personal project for **Edge AI food recognition**, I ensured ethical design by:

- Training on a diverse food dataset to avoid cultural bias
- Keeping all inference offline (privacy-first)
- Using explainable architecture (MobileNetV2 with clear classification outputs)

To improve fairness, I plan to:

- Expand the dataset to include underrepresented cuisines
- Conduct periodic audits for performance gaps
- Offer clear confidence scores for transparency

Ethical AI starts with intention and continues through responsible iteration.