Designing Responsible and Fair AI Systems”

🧠 Part 1: Theoretical Understanding (30%)

1. Short Answer Questions

Q1: Algorithmic Bias

- **Definition:** Algorithmic bias occurs when an AI system reflects or amplifies existing societal biases due to skewed data, flawed assumptions, or design flaws.
- **Examples:**
    1. A hiring algorithm trained on past male-dominated resumes downgrading female applicants.
    2. Loan approval systems offering higher credit limits to white applicants over minorities with similar financial profiles.

Q2: Transparency vs Explainability

- **Transparency** refers to understanding the inner workings of an AI system (architecture, data flows).
- **Explainability** refers to how well the system's decisions can be understood by humans.
- **Importance:** Transparency helps developers monitor systems; explainability builds user trust and supports accountability.

Q3: GDPR’s Impact on AI

- **GDPR** enforces data protection and privacy rights in the EU.
- **Key Impacts:**
    - “Right to explanation” for algorithmic decisions.
    - Requires **data minimization** and **explicit consent**.
    - Limits use of personal data in profiling and automated decision-making.

---

2. 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 (40%)

Case 1: Amazon’s Biased Hiring Tool

Source of Bias:

- Training data was historical resumes, mostly from male applicants → the model learned male-dominant patterns.

Three Fixes:

1. **Rebalance the training data** to include diverse and gender-neutral examples.
2. **Apply fairness-aware pre-processing** (e.g., reweighting) or in-processing methods (fair classifiers).
3. **Exclude gender proxies** (e.g., women’s colleges, pronouns) from features.

Fairness Metrics:

- **Disparate Impact Ratio**
- **Equal Opportunity Difference**
- **False Negative/Positive Rate Gap by Gender**

Case 2: Facial Recognition in Policing

Ethical Risks:

- **Wrongful arrests** due to misidentification of minorities.
- **Surveillance creep** and **loss of privacy**.
- **Disproportionate harm** to marginalized communities.

Policies for Responsible Deployment:

1. Ban in high-risk contexts (e.g., real-time surveillance) unless accuracy > 99% across all groups.
2. Mandatory **bias audits** before deployment.
3. Public **transparency reports** and independent oversight.

🧪 Part 3: Practical Audit – COMPAS Dataset (25%)

Steps:

1. **Load dataset** using `pandas`.

2. **Use AI Fairness 360 toolkit** (especially the `BinaryLabelDatasetMetric`, `ClassificationMetric`).
3. Analyze metrics:
    - Disparate impact
    - Statistical parity difference
    - False Positive Rate by race
4. **Visualize** using `matplotlib`: bar charts or disparity plots.

300-Word Report (Example Template)

We audited the COMPAS Recidivism dataset using AI Fairness 360. Our focus was on racial bias in predicting re-offending risks.

We found significant disparities in **False Positive Rates (FPR)**: African-American defendants had an FPR of 45%, while Caucasian defendants had 23%. This implies Black individuals are nearly twice as likely to be incorrectly labeled as "high risk."

The **Disparate Impact Ratio** for African-American defendants was 0.62 (ideal = 1), indicating unfair outcomes under the "four-fifths rule."

Remediation Steps:

- Use reweighing during preprocessing.
- Apply `AdversarialDebiasing` in-processing model.
- Include fairness constraints during training.

Future audits should involve community stakeholders and periodic evaluations.

🤔 Part 4: Ethical Reflection (5%)

Prompt Answer Example (200–300 words):

In a past project, I developed a resume screening tool using NLP. At the time, I didn't assess for gender or racial bias. If I revisit this, I would:

1. **Audit training data** for representation.
2. **Implement explainability** tools like SHAP to understand decision paths.
3. **Include a fairness module** using AI Fairness 360.
4. **Seek user feedback** from diverse groups.

My future work will prioritize **transparency**, **user autonomy**, and **harm prevention**, aligning with EU AI ethics guidelines.

🩺 Bonus Task (Extra 10%) – Ethical AI in Healthcare

1-Page Policy Proposal Highlights:

Title: Ethical AI Guidelines for Healthcare

- **Patient Consent:**
  - Informed consent before AI usage.
  - Right to opt-out and understand risks.
- **Bias Mitigation:**
  - Mandatory audits across race, gender, age.
  - Fairness-aware models (e.g., reweighing, adversarial debiasing).
- **Transparency:**
  - Explainable AI for diagnoses and risk scores.
  - Regular public reports and third-party audits.