🧠 Part 1: Theoretical Understanding (30%)

Q1: Define algorithmic bias and provide two examples.

Definition:
Algorithmic bias refers to systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one group over another.

Examples:

- **Hiring Algorithms**: An AI trained on historical hiring data might favor male candidates if the data reflects past gender discrimination.
- **Credit Scoring Models**: A model might give lower credit scores to minorities due to biased training data reflecting historical economic disparities.

Q2: Difference between transparency and explainability. Why are both important?

- **Transparency**: Refers to how open and accessible the design, data, and decision-making processes of an AI system are to stakeholders.
- **Explainability**: Refers to the ability to understand how and why an AI system reached a particular decision or prediction.

Importance:

- Transparency builds trust and accountability.
- Explainability helps users and regulators understand decisions, especially in high-stakes areas like healthcare or justice.

Q3: How does GDPR impact AI development in the EU?

- **Right to Explanation**: Users can demand explanations for automated decisions.

- **Consent Requirements**: AI systems must have explicit user consent for personal data processing.
- **Data Minimization & Fairness**: Developers must ensure data is relevant and bias-free.

Impact: GDPR enforces transparency, accountability, and user rights, shaping ethical AI development in the EU.

Ethical Principles Matching:

Principle	Definition
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.
A) Justice	Fair distribution of AI benefits and risks.

🕵️ Part 2: Case Study Analysis (40%)

👩‍💼 Case 1: Biased Hiring Tool (Amazon)

1. Source of Bias:

- Biased historical hiring data favoring male applicants.
- Use of proxy variables (e.g., resume wording) correlated with gender.
- Lack of fairness constraints during model development.

2. Three Fixes:

- Retrain model on **balanced, anonymized data**.
- Remove **gender-correlated features** (e.g., school attended if highly skewed).

- Apply **fairness-aware algorithms** (e.g., reweighting or adversarial debiasing).

3. Fairness Metrics:

- **Disparate Impact Ratio**
- **Equal Opportunity Difference**
- **Demographic Parity**

👮 Case 2: Facial Recognition in Policing

1. Ethical Risks:

- Wrongful arrests due to misidentification (esp. minorities).
- Erosion of privacy and consent.
- Disproportionate surveillance and systemic discrimination.

2. Policy Recommendations:

- Mandatory **third-party bias audits** before deployment.
- **Consent-based surveillance** in public spaces.
- Transparent **accountability framework** and **error disclosures**.
- Ban in high-risk settings unless accuracy across demographics is proven.

🧪 Part 3: Practical Audit (25%)

Task: Bias Audit of COMPAS Dataset Using AI Fairness 360

Steps:

1. Load the COMPAS dataset.
2. Define protected attribute: `race`.
3. Use fairness metrics:
    - False Positive Rate (FPR)
    - Equal Opportunity Difference
    - Disparate Impact Ratio
4. Visualize:
    - Bar plots for FPR across races.
    - Histograms of predicted vs. actual outcomes.

300-word Summary (Example):

> Our fairness audit of the COMPAS dataset revealed significant racial bias. The False Positive Rate (FPR) for African-American defendants was substantially higher than for Caucasians, suggesting overestimation of recidivism risk for Black individuals. Using AI Fairness 360, we also observed a disparate impact ratio below the recommended threshold (0.8), indicating unequal outcomes.
>
> To mitigate bias, we recommend:
>
> - Reweighing the dataset to balance racial groups.
> - Training models using fairness-constrained optimization.
> - Incorporating post-processing techniques to correct unfair predictions.
>
> These remediation strategies aim to build equitable risk assessment tools and avoid harm in judicial decisions.

🤔 Part 4: Ethical Reflection (5%)

Prompt Response (Example):

In a previous project on loan approval prediction, I realized the training data overrepresented wealthy urban applicants. Going forward, I will apply fairness metrics early in model development, ensure representative sampling, and implement transparent data practices. I aim to design systems that are inclusive and reduce socioeconomic bias.

🏥 Bonus Task (10%): Policy Proposal (Healthcare AI Ethics)

1-Page Guideline Summary (Key Points):

Title: Ethical AI Use in Healthcare – A Practical Guideline

1. **Patient Consent Protocols:**
    - Transparent data use disclosures.
    - Opt-in mechanisms with granular control.
    - Data anonymization policies.
2. **Bias Mitigation Strategies:**
    - Continuous bias audits across race, gender, age.
    - Diverse training datasets from multiple demographics.
    - Use of explainable and fair models (e.g., LIME + AI Fairness 360).
3. **Transparency Requirements:**
    - Explainable decision-making (e.g., treatment suggestions).
    - Model versioning and update logs.
    - Right to second opinion (human-in-the-loop approach).

📁 Submission Tips:

- **PDF Report**: Include all written answers, reflections, and case analyses.
- **GitHub Repo**:
    - `/notebooks/compas_fairness_audit.ipynb`
    - `/report/summary.md` (for audit)
    - README with group names and instructions.
- **Bonus Task**: Format as an article for posting on PLP Academy Community.
