PART 1 — THEORETICAL UNDERSTANDING (30%)

Q1: Define algorithmic bias and give two examples.

Algorithmic bias refers to systematic and unfair discrimination produced by an AI system. It occurs when data, model assumptions, or deployment processes lead to outcomes that disproportionately disadvantage certain groups.

Examples:

1. **Hiring Algorithms Favoring Men**
   Amazon's recruitment model downgraded CVs containing the word "women's," because it was trained on 10 years of male-dominated hiring data.
2. **Facial Recognition Falsely Identifying Minorities**
   Commercial facial recognition systems show higher false positive rates for Black and Asian individuals compared to white individuals, leading to wrongful police matches.

Q2: Transparency vs Explainability

Transparency

Refers to openness about an AI system's **data sources, design choices, training processes, algorithms, and limitations**.
It answers:
"What is inside the model?"

Explainability

Refers to the ability to **understand why the model made a specific decision.**
It answers:
"Why did the model produce this output?"

Why Both Are Important

- Enables **user trust and accountability**
- Helps detect **biases and errors**
- Required for **legal compliance** (GDPR, AI Act)
- Allows developers and auditors to **improve fairness**

Q3: How GDPR Impacts AI Development

GDPR influences AI development by enforcing strict regulations on:

- **Data consent**
  AI systems must collect data legally with explicit user permission.
- **Right to explanation**
  Users can request meaningful explanations of automated decisions (Article 22).
- **Data minimization**
  AI models may only collect and process data necessary for their function.
- **Privacy by design**
  Developers must integrate privacy protection into the model lifecycle.
- **Penalties**
  Non-compliance can result in multi-million-euro fines.

300-Word Bias Audit Report (COMPAS)

The COMPAS Recidivism dataset is widely known for exhibiting racial bias in predicting the likelihood of reoffending. This audit used IBM's AI Fairness 360 toolkit to evaluate disparities between privileged (Caucasian) and unprivileged (African-American) groups. Several fairness metrics were computed, including Disparate Impact, Statistical Parity Difference, False Positive Rate (FPR) difference, and Equal Opportunity Difference.

Initial dataset-level metrics show a **Disparate Impact ratio below 0.8**, indicating that the unprivileged racial group receives unfavorable predictions at a disproportionate rate. Statistical Parity Difference is significantly negative, confirming that African-American defendants are more likely to be labeled "high risk" regardless of actual recidivism outcomes.

Using Logistic Regression as a baseline model, a substantial **FPR difference** was observed. African-American individuals were incorrectly predicted as "high risk" far more often than Caucasian individuals. This raises major concerns because false positives in criminal justice contexts can directly lead to harsher bail decisions, longer sentences, or denial of parole. Equal Opportunity Difference also indicated unfairness, with true positive rates differing across groups, signaling unequal access to correct predictions.

These findings reflect structural inequalities inherited from historical policing and judicial data. The COMPAS model amplifies these disparities unless active mitigation strategies are applied.

To reduce bias, several techniques are recommended: (1) **Reweighing** to balance favorable/unfavorable label distribution across groups; (2) **Pre-processing debiasing**, such as Disparate Impact Remover; (3) **In-processing approaches**, including adversarial debiasing to enforce demographic parity; and (4) **Post-processing corrections**, such as Equalized Odds adjustment.

In conclusion, the COMPAS dataset contains substantial racial bias, and any model trained on it must undergo rigorous fairness interventions before deployment in real-world decisions involving individual liberty.

☑ PART 4 — ETHICAL REFLECTION (5%)

In my future AI projects, I will prioritize fairness, transparency, and accountability by integrating ethical principles throughout the development lifecycle. First, I will ensure that datasets are diverse and representative to avoid reinforcing societal biases. Before training any model, I will perform exploratory data analysis to detect outliers, imbalances, or discriminatory patterns.

I will adopt **privacy by design**, collecting only the data required and anonymizing sensitive attributes whenever possible. My models will include transparency mechanisms—such as interpretable architectures, feature importance evaluation, and clear documentation explaining how decisions are made.

Human oversight will be central. Automated decisions will not operate independently in high-risk contexts. Instead, I will implement human-in-the-loop review settings where experts can validate or override AI decisions. I will evaluate fairness using multiple metrics (statistical parity, equal opportunity, disparate impact) and apply mitigation techniques where needed.

Finally, I will maintain accountability by openly documenting limitations, updating systems regularly, and designing feedback mechanisms to detect ethical issues post-deployment. Ethical AI is not a one-time task but a continuous responsibility.

★ BONUS TASK — Healthcare AI Ethics Guideline (1 Page)

Guideline for Ethical AI in Healthcare

1. Patient Consent Protocols

- Obtain explicit, informed consent before collecting or using patient data.
- Provide clear explanations of how AI assists diagnosis or treatment.
- Allow patients to opt out without penalty.

2. Data Governance & Privacy

- Encrypt all medical data and store it securely.
- Use anonymization and pseudonymization.
- Limit data collection to medically relevant features.

3. Bias Mitigation Strategies

- Use balanced datasets reflecting gender, ethnicity, age, and socioeconomic groups.
- Perform regular fairness audits to check accuracy differences across demographic groups.
- Implement bias-resistant algorithms (reweighing, adversarial debiasing).

4. Transparency Requirements

- Provide clinicians with interpretable outputs and decision explanations.
- Maintain documentation detailing training data, limitations, and risk scenarios.
- Publish audit results to regulatory bodies.

5. Accountability & Safety

- Keep a human clinician responsible for final medical decisions.
- Report adverse outcomes and retrain models accordingly.
- Enforce strict validation before deployment.
