Part 1: Theoretical Understanding (30%)

Q1: Define algorithmic bias and provide two examples of how it manifests in AI systems.

Algorithmic bias refers to systematic and repeatable errors in an AI system that result in unfair outcomes, such as privileging or disadvantaging certain groups.

Example 1: **Hiring Tools** — AI trained on historical hiring data may favor male candidates if the past data reflects gender bias in recruitment.

Example 2: **Credit Scoring** — Loan approval algorithms may score applicants from certain neighborhoods lower due to socioeconomic bias present in the training data.

Q2: Explain the difference between transparency and explainability in AI. Why are both important?

- **Transparency** is the openness about how an AI system is built and functions (e.g., access to model architecture, data sources, and development processes).
- **Explainability** refers to how easily humans can understand why an AI made a specific decision or prediction.

Importance:

- Transparency helps **regulators and developers** audit AI systems.
- Explainability helps **users and stakeholders** trust and challenge AI outcomes, ensuring ethical and legal compliance.

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

GDPR enforces data protection and privacy rights, directly affecting AI systems that process personal data. Key impacts include:

- **Right to explanation**: Users can demand an explanation of automated decisions.
- **Data minimization**: AI systems must limit data use to only what is necessary.
- **Consent**: Explicit user consent is required for data collection and processing.
- **Accountability**: Developers must ensure fairness, transparency, and data protection by design.

---

2. 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 AI)

- **Source of Bias**:
  - The model was trained on 10 years of resumes from predominantly male applicants, embedding historical gender bias into its predictions.
- **Three Fixes**:
  1. **Diversify training data**: Use balanced datasets with equal representation of gender and other demographic factors.
  2. **Debias features**: Remove or neutralize gender-correlated terms and features (e.g., gendered language or names).
  3. **Audit regularly**: Implement continuous fairness testing and human-in-the-loop oversight.
- **Fairness Evaluation Metrics**:
  - **Demographic parity**: Equal selection rates across genders.
  - **Equal opportunity**: Equal true positive rates for all groups.
  - **Disparate impact ratio**: Ratio of positive outcomes between groups close to 1 (ideal is ≥ 0.8).

Case 2: Facial Recognition in Policing

- **Ethical Risks**:
  - **Wrongful arrests** due to misidentification of minorities.
  - **Privacy violations** from constant surveillance without consent.
  - **Discrimination and distrust** in marginalized communities.
- **Recommended Policies**:
  1. **Bias audits** before deployment and after updates.
  2. **Restrict use** to critical scenarios with human verification (e.g., not for petty crimes).
  3. **Transparency & accountability** through public reporting and independent oversight.
  4. **Informed consent** and data minimization in surveillance applications.

Part 4: Ethical Reflection (5%)

Project: *HighProbabilityTrader.mq5* — An AI-powered expert advisor (EA) designed for high-probability trading in financial markets.

Ethical AI Principles and Safeguards:

1. **Transparency**:
   I will document the decision-making logic behind each trading strategy (e.g., trend-following, momentum) and provide clear user guidelines explaining how predictions are made and under what conditions the EA trades.
2. **Fairness**:
   To prevent favoring particular market conditions or data sources, I will test the EA across diverse datasets (e.g., emerging and developed markets) and ensure it doesn't rely solely on biased patterns that disadvantage certain trader profiles.
3. **Accountability**:
   I will incorporate a human-in-the-loop feature where users can override automated trades and access logs of all AI decisions to maintain control and oversight.
4. **Data Privacy**:
   Any user-specific financial data will be encrypted and anonymized to comply with data protection principles like GDPR.
5. **Non-maleficence**:
   I will integrate risk management modules to prevent reckless trades, such as limiting drawdowns and avoiding high-leverage scenarios that could cause users financial harm.

Bonus Task: Ethical AI Use in Healthcare — Policy Proposal (1 Page)

Guidelines for Ethical AI Deployment in Healthcare Systems

---

1. Patient Consent Protocols

- **Informed Consent**:
  Patients must be informed clearly and accessibly if an AI system is used in diagnosis, treatment planning, or patient monitoring. Consent must be obtained in writing.
- **Right to Opt-Out**:
  Patients should be allowed to opt out of AI-driven decisions and request human oversight without discrimination or reduced care quality.

- **Data Control**:
  Patients must be able to access, modify, or delete their personal health data in compliance with GDPR or equivalent frameworks.

---

2. Bias Mitigation Strategies

- **Diverse Training Data**:
  Ensure datasets represent varied demographics across ethnicity, gender, age, and socio-economic backgrounds to avoid algorithmic bias.
- **Fairness Audits**:
  Conduct regular fairness assessments using metrics such as Equal Opportunity and Disparate Impact to identify disparities in diagnosis or treatment recommendations.
- **Stakeholder Review Boards**:
  Include ethicists, patient advocates, and healthcare professionals to review and approve AI system deployment.

---

3. Transparency Requirements

- **Model Explainability**:
  Use interpretable AI models or post-hoc explanation tools (e.g., SHAP, LIME) to clarify decisions to clinicians and patients.
- **Audit Logs**:
  Maintain traceable logs of AI decisions to allow auditing and appeals in cases of misdiagnosis or harm.
- **Public Documentation**:
  Publish model performance metrics, limitations, training data sources, and known risks to ensure institutional and public accountability.

Report: COMPAS Bias Audit (300 words)

Overview:
The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset is widely used to predict recidivism risk. However, it has raised concerns about racial bias. Using IBM’s AI Fairness 360 toolkit, we audited the dataset for fairness, specifically analyzing the difference in false positive rates between African-American (unprivileged) and Caucasian (privileged) groups.

Findings:
After training a logistic regression classifier on the COMPAS dataset, we observed a significant disparity in the **false positive rate** (FPR) between racial groups. African-American individuals were classified as high-risk more frequently even when they did not reoffend, resulting in a

higher FPR compared to their Caucasian counterparts. Specifically, the FPR for the unprivileged group was markedly higher—this aligns with known criticisms of COMPAS's racial bias.

The bias metric confirms that the system unfairly penalizes African-American defendants, potentially affecting parole decisions or bail eligibility. This reflects algorithmic discrimination, likely due to biased historical data, where systemic inequities are encoded in the features used for prediction (e.g., number of prior offenses, age, charge degree).

Remediation:

To address this issue, we applied the **Reweighing algorithm**, a pre-processing technique that adjusts instance weights to reduce bias before training. While this doesn't eliminate bias entirely, it helped reduce the FPR gap between racial groups.

In production, further steps should include:

1. Post-processing methods (e.g., calibrated equal odds).
2. Model monitoring with fairness metrics across deployments.
3. Inclusion of explainability tools for stakeholder trust.
4. Involving legal and ethical oversight during deployment.

Conclusion:

Bias in risk assessment tools like COMPAS can perpetuate injustice. Auditing with AI Fairness 360 reveals tangible disparities, and reweighing provides a practical path toward more equitable outcomes. However, technical fixes alone are insufficient—ethical deployment requires transparency, diverse representation, and accountability.