Part 1: Theoretical Understanding

1. Short Answer Questions

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

  Algorithmic Bias: refers to systematic and unfair discrimination in the outcomes of AI systems, often resulting from biased data or flawed model design.

  Example 1: A hiring algorithm trained on past employee data may favor male candidates if historical hiring was biased against women.

  Example 2: A facial recognition system may struggle to accurately identify people with darker skin tones if it has been trained primarily on images of lighter-skinned individuals.

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

  Transparency: Refers to being open about how an AI system is built-its data source, algorithm, and design decisions. Explainability, on the other hand, refers to how easily a human can understand why the AI made a specific decision.

  They are both important because transparency builds trust and accountability, while explainability helps users and stakeholders understand, challenge, or correct AI outcomes, especially in high-stakes settings like healthcare or criminal justice.

- **Q3: How does GDPR (General Data Protection Regulation) impact AI development in the EU?**

  GDPR requires that AI systems handling personal data ensure user consent, data protection, and accountability. It also grants individuals the right to understand and consent to automated decisions. This impacts AI development by encouraging responsible design, limiting the use of certain data, and requiring transparency in decision-making processes.

2. Ethical Principles Matching

Match the following principles to their definitions:

- **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 data and decisions.**
- **D) Sustainability- Designing AI to be environmentally friendly.**

1. *Ensuring AI does not harm individuals or society.*
2. *Respecting users' right to control their data and decisions.*
3. *Designing AI to be environmentally friendly.*
4. *Fair distribution of AI benefits and risks.*

Part 2: Case Study Analysis

Case 1: Biased Hiring Tool

- **Scenario: Amazon's AI recruiting tool penalized female candidates.**
- **Tasks:**

1. **Identify the source of bias (e.g., training data, model design).**
2. **Propose three fixes to make the tool fairer.**
3. **Suggest metrics to evaluate fairness post-correction.**

1. Source of Bias

- **Training Data Bias**:
  The AI model was trained on resumes submitted to Amazon over a 10-year period. Since the tech industry has been male-dominated, the historical data reflected gender imbalance, leading the model to associate male candidates with successful hires more often than female ones.

- **Proxy Variables**:
  The model picked up on proxies for gender (e.g., women's colleges, gendered language) and penalized resumes that contained them, despite gender not being an explicit input.

2. Three Fixes to Make the Tool Fairer

Debias the Training Data:

- Use a more balanced dataset that includes equal representation of resumes from male and female candidates.
- Remove or neutralize gender-related indicators (like “women’s chess club”) that could act as proxies.

Introduce Fairness Constraints in the Model:

- Integrate fairness-aware machine learning techniques (e.g., equal opportunity constraints) to prevent the model from learning discriminatory patterns.

Regular Auditing and Human Oversight:

- Implement a process for regular bias audits using synthetic and real-world test data.
- Keep human recruiters in the loop to validate and override AI recommendations, especially when edge cases or potential bias is detected.

3. Fairness Evaluation Metrics (Post-Correction)

Demographic Parity:
Measure whether candidates from different gender groups are being recommended for interviews at similar rates.

Equal Opportunity:
Check whether qualified female and male candidates have equal chances of being selected by the model.

Disparate Impact Ratio:
Calculate the ratio of selection rates between groups. A ratio below 0.8 (80%) typically indicates potential bias (based on the "four-fifths rule").

False Positive/Negative Rates by Group:
Compare how often the model wrongly accepts or rejects resumes across different gender groups.

Case 2: Facial Recognition in Policing

- **Scenario: A facial recognition system misidentifies minorities at higher rates.**
- **Tasks:**
    1. **Discuss ethical risks (e.g., wrongful arrests, privacy violations).**
    2. **Recommend policies for responsible deployment.**

1. Ethical Risks

Wrongful Arrests

- Misidentification can lead to innocent individuals—particularly from minority communities—being wrongly detained or arrested.
- This undermines trust in law enforcement and can have serious personal, legal, and emotional consequences for the affected individuals.

Systemic Discrimination

- Bias in facial recognition reinforces existing inequalities by disproportionately targeting marginalized groups.
- It can exacerbate over-policing in communities already subject to surveillance and enforcement pressure.

Privacy Violations

- The use of facial recognition often occurs without individuals' consent or awareness.
- Mass surveillance may infringe on rights to privacy and freedom of movement, especially in public spaces.

Lack of Accountability

- Opaque decision-making processes and lack of transparency in AI use make it hard for individuals to contest errors or seek redress.

2. Policies for Responsible Deployment

Independent Auditing and Bias Testing

- Require regular third-party audits to test for accuracy across different demographic groups.
- Mandate public reporting of performance metrics disaggregated by race, gender, and age.

Strict Use Restrictions

- Prohibit real-time facial recognition in public spaces unless a warrant or judicial authorization is obtained.
- Limit usage to specific, high-stakes scenarios (e.g., investigating serious crimes), not routine surveillance.

Human-in-the-Loop Decision-Making

- Ensure facial recognition is used only as a decision-support tool, with human officers required to verify matches before taking action.
- Provide officers with mandatory training on how to interpret and verify AI-generated results.

Transparency and Public Accountability

- Implement clear public policies about how, when, and why facial recognition is used.
- Allow individuals to request information on whether they have been subject to facial recognition and provide channels to dispute misidentification.

Community Oversight and Consent

- Engage with affected communities to determine acceptable use cases.
- Establish oversight bodies that include civil rights advocates and community representatives.

Part 3: Practical Audit

Task: Audit a Dataset for Bias

- **Dataset**: COMPAS Recidivism Dataset.
- **Goal**:
    1. **Use Python and AI Fairness 360 (IBM's toolkit) to analyze racial bias in risk scores.**
    2. **Generate visualizations (e.g., disparity in false positive rates).**
    3. **Write a 300-word report summarizing findings and remediation steps.**

Deliverable: Code + report.

AI Fairness Audit Report

Project: COMPAS Recidivism Dataset Fairness Audit

Tools Used: Python, IBM AI Fairness 360 (AIF360), Pandas, NumPy

Objective

To assess and document algorithmic fairness in the COMPAS recidivism risk score dataset using AIF360's bias detection metrics. The focus is on detecting potential racial or ethnic bias in how risk is predicted.

Dataset Overview

Three versions of the dataset were analyzed:

Dataset: df1

- Protected Attribute: race

- Label Definition: is_recid (reoffended)

- Notes: Base COMPAS format

Dataset: df2

- Protected Attribute: Ethnic_Code_Text

- Label Definition: DecileScore >= 5

- Notes: Custom threshold proxy

Dataset: df3

- Protected Attribute: Reconstructed from one-hot race columns

- Label Definition: Derived from cleaned format

Current Focus: df1

Variables Used

- Label: is_recid (0 = no recidivism, 1 = recidivism)

- Protected Attribute: race

- Privileged Group: Caucasian

- Unprivileged Group: African-American

Preprocessing Summary

- Filtered for valid racial groups: African-American, Caucasian, Hispanic

- Verified label distribution:

- 0: 12,040 instances

- 1: 2,822 instances

- Ensured both label values are present for fairness metrics.

Fairness Metrics Computed

- Mean Difference (Disparate Impact)

- Statistical Parity Difference

- Equal Opportunity Difference

- True Positive Rate (TPR)

- Accuracy by group

- Label balance by race

Paused here due to a ValueError when one class label was missing or invalid in another dataset

version.

Key Challenges

- Some datasets (like df2) had unbalanced or missing label classes (e.g., no 1s or 0s).

- One-hot encoding in df3 required reconstruction of categorical attributes.

- Occasional AIF360 compatibility issues with input format (e.g., label/protected attr mismatch).

Part 4: Ethical Reflection

- **Prompt**: Reflect on a personal project (past or future). How will you ensure it adheres to ethical AI principles?

Ethical Reflection:
One project I'm working on is a mobile platform that connects people to local service providers—like plumbers, electricians, and tech support—especially in areas where access to reliable professionals is limited. As I build this, I want to make sure I'm not just focusing on speed and features, but also on doing the right thing ethically.

First, I'll make sure the system treats everyone fairly. That means checking that the algorithm doesn't favor one group over another, like giving more visibility to certain providers just because of where they're from or how often they're online. Everyone should have an equal chance.

Second, I'll be clear about how the platform works. If someone is matched with a provider, I want them to understand why—whether it's based on location, rating, or availability. No hidden decisions.

I'll also respect people's privacy. Users should always know what data is being collected and have control over it. That includes the ability to delete their data if they no longer want to use the app.

Safety matters too. I'll build in ways for users to report problems or bad experiences, and I'll verify providers to help build trust on both sides.

Lastly, I'll try to keep the project environmentally responsible, like choosing green hosting options and encouraging providers to use sustainable materials or transport where possible.

Overall, I believe that building tech with ethics in mind isn't just good practice—it makes the product better for everyone.

Bonus Task

- **Policy Proposal: Draft a 1-page guideline for *ethical AI use in healthcare*. Include:**
    - **Patient consent protocols.**
    - **Bias mitigation strategies.**
    - **Transparency requirements.**

1. Introduction

This policy provides ethical guidelines for the responsible development and deployment of Artificial Intelligence (AI) in healthcare. It aims to protect patient rights, ensure safety, and promote fairness in AI-supported clinical decision-making, diagnostics, and care delivery.

2. Patient Consent Protocols

- Informed Consent:
  Patients must be clearly informed when AI is used in their diagnosis, treatment, or health management. Explanations should include what the AI does, its limitations, and how it affects their care.

- Opt-In Approach:
  AI involvement should require explicit, opt-in consent. Patients must be allowed to decline AI involvement without negative consequences to their care.

- Data Use Transparency:
  Patients must be informed of how their medical data will be used, stored, and shared.
  De-identified data used for model training must follow strict data protection laws (e.g., HIPAA, GDPR).

3. Bias Mitigation Strategies

- Diverse Training Data:
  AI systems must be trained on datasets that represent the diversity of the patient population, including variations in race, gender, age, and socioeconomic status.

- Bias Audits:
  Conduct regular audits to evaluate the AI's performance across demographic groups. Disparities in prediction or treatment outcomes must be documented and addressed.

- Clinical Oversight:
  AI tools should always support—not replace—clinician judgment. Final decisions must rest with qualified healthcare professionals, especially when outcomes may vary across patient groups.

4. Transparency Requirements

- Explainability:
  AI tools must provide understandable explanations for their recommendations or decisions. Clinicians and patients should be able to understand the reasoning behind outputs.

- Model Documentation:
  Developers must publish detailed documentation about model design, training data, validation procedures, and known limitations.

- Regulatory Compliance:
  All AI systems must comply with healthcare regulations and undergo independent validation before deployment (e.g., FDA approval in the U.S.).

5. Accountability & Continuous Monitoring

- Assign clear responsibility for AI decisions and maintain logs of system outputs.

- Implement feedback loops for clinicians to report errors or unexpected behavior.

- Continuously monitor real-world performance and update models as new data becomes available.