Part 1: Theoretical Understanding (30%)

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 repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users over another. It is not a technical glitch but often a reflection of historical and social biases embedded in the data or design choices.

Two examples:

1. **Biased Hiring Tools:** Amazon's recruiting engine, trained on a decade of resumes from a male-dominated tech industry, learned to penalize applications that contained words like "women's" (e.g., "women's chess club captain"). This manifested as a bias against female candidates.
2. **Biased Criminal Justice Tools:** The COMPAS algorithm, used to predict the likelihood of a defendant re-offending, was found to have a higher false positive rate for Black defendants than for white defendants, incorrectly labeling them as high-risk at a disproportionate rate.

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

- **Transparency** is about the openness of the AI system's design, data, and processes. It answers the question, "How was this system built?" A transparent system provides information about its training data, the model's architecture, and the overall development process.
- **Explainability** is about the ability to understand and articulate the reasoning behind a specific AI decision or output. It answers the question, "Why did the model make this particular decision for this specific case?"

Both are important because:

- **Transparency** builds trust with stakeholders (regulators, users, the public) and allows for external auditing and accountability.

- **Explainability** is crucial for practical use. For example, if a loan application is denied by an AI, the bank must be able to explain the reason to the customer. It also helps developers debug and improve their models.

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

GDPR imposes significant legal constraints and obligations on AI development, primarily through:

1. **Lawfulness, Fairness, and Transparency (Article 5):** Requires AI systems to be transparent about how personal data is used and ensure processing is fair and lawful.
2. **Purpose Limitation:** Data collected for one purpose cannot be repurposed for AI training without further consent.
3. **Data Minimization:** AI systems should only use data that is strictly necessary, discouraging the "collect everything" approach.
4. **Right to Explanation (Articles 13-15 & 22):** Grants individuals the right to obtain meaningful information about the logic involved in automated decision-making, including profiling, that significantly affects them. This directly mandates a degree of explainability.
5. **Right to Erasure ('Right to be Forgotten'):** Individuals can request their data be deleted, which complicates the maintenance of large, static training datasets.

These principles force EU AI developers to build privacy and fairness into the design phase (Privacy by Design) and prioritize interpretable models over unaccountable "black boxes."

2. Ethical Principles Matching

- **(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

- **Source of Bias:** The primary source was **biased training data**. The model was trained on resumes submitted to Amazon over a 10-year period, which were overwhelmingly from male applicants. The AI learned to associate male candidates with being successful hires and thus penalized patterns and keywords found more commonly in female applicants' resumes.
- **Three Proposed Fixes:**

1. **Debias the Training Data:** Curate a more balanced dataset that represents a diverse range of qualified candidates, not just historical hires. Use techniques like re-sampling or re-weighting to reduce the influence of the male-dominated data.
2. **Remove Protected Attributes and Proxies:** Actively identify and remove not only direct identifiers like gender and race but also proxy variables that could lead to bias (e.g., names, university affiliations, hobbies, or words strongly correlated with a specific demographic).
3. **Implement a "Human-in-the-Loop" (HITL) System:** Redesign the tool to be an assistant, not a decision-maker. The AI would screen for skills and qualifications, presenting a ranked shortlist to human recruiters who make the final call, with built-in checks for demographic diversity in the shortlist.

- **Metrics to Evaluate Fairness:**

  - **Demographic Parity:** Check if the rate of being shortlisted is similar across gender groups.
  - **Equalized Odds:** Compare **False Positive Rates** (rate at which qualified female candidates are incorrectly rejected) and **False Negative Rates** (rate at which unqualified male candidates are incorrectly advanced) across groups. The goal is for these rates to be statistically similar.
  - **Predictive Parity:** Ensure that the precision (the proportion of selected candidates who are actually qualified) is similar for both male and female candidates.

Case 2: Facial Recognition in Policing

- **Ethical Risks:**

  - **Wrongful Arrests and Convictions:** The higher misidentification rate for minorities can lead to innocent people being detained, arrested, or even convicted, eroding trust in the justice system and causing severe personal harm.
  - **Mass Surveillance and Privacy Violations:** The pervasive use of FR systems can create a surveillance state, chilling freedom of assembly and movement, and fundamentally altering the relationship between citizens and the state.
  - **Exacerbation of Systemic Bias:** If deployed predominantly in minority neighborhoods, the technology can reinforce existing policing biases, leading to a feedback loop where more data from these areas further biases the models.

- **Policies for Responsible Deployment:**

1. **A Moratorium or Strict Regulation:** Ban the use of "live" or real-time facial recognition for general surveillance and restrict its use to investigating serious crimes where a warrant has been obtained.
2. **Mandatory Third-Party Bias Audits:** Require independent, transparent audits of the algorithms for accuracy across demographics before any procurement or deployment.
3. **Transparency and Public Accountability:** Publish clear policies on the system's use, its documented error rates, and the procedures for handling misidentification. Establish an independent oversight board.
4. **Robust Officer Training:** Train law enforcement on the limitations of the technology, emphasizing that it should be used only as an investigative lead and never as sole evidence for an arrest.