Part 2: Case Study Analysis (40%)

Case 1: Biased Hiring Tool – Amazon's AI Recruiting Tool

Scenario: Amazon's internal AI tool used for screening resumes penalized female candidates, favoring male-dominated language and historical hiring patterns.

1. Identify the Source of Bias:

- **Training Data Bias:** The model was trained on **10 years of hiring data** that reflected male-dominated hiring practices in the tech industry, resulting in the model **learning to favor resumes with male-associated terms** and penalizing mentions of "women's" (e.g., "women's chess club").
- **Model Design Bias:** The system reinforced patterns in the data without counterbalancing for gender neutrality or underrepresentation.

2. Three Fixes to Make the Tool Fairer:

a) Bias-Aware Data Preprocessing:

- Remove or anonymize gender-related features or proxies (e.g., names, clubs).
- Balance the dataset by ensuring diverse gender representation.

b) Fairness Constraints in Model Training:

- Apply fairness-aware algorithms such as **prejudice remover** or **adversarial debiasing** from the **AI Fairness 360** toolkit.
- Use algorithms that account for **demographic parity** during training.

c) Human-in-the-Loop Systems:

- Include a diverse panel of recruiters to review AI recommendations.
- Ensure **algorithmic decisions are audited and overruled** when necessary.

3. Fairness Evaluation Metrics Post-Correction:

- **Disparate Impact Ratio (DIR):** Measures the ratio of favorable outcomes between groups (e.g., male vs. female candidates).
- **Equal Opportunity Difference:** Checks if both groups have equal true positive rates.
- **Demographic Parity Difference:** Measures outcome rates between groups to ensure equality.
- **False Positive/Negative Rates by Group:** To avoid one group being unfairly penalized.

Case 2: Facial Recognition in Policing

Scenario: Facial recognition systems used by law enforcement misidentify minorities at higher rates, resulting in wrongful arrests and public distrust.

1. Ethical Risks:

a) Wrongful Arrests & Discrimination:

- Misidentifications disproportionately affect minority communities, leading to wrongful detentions, criminal records, and social stigma.

b) Privacy Violations:

- Continuous surveillance and data collection without consent violate individuals' privacy rights.

c) Erosion of Public Trust:

- Overreliance on flawed technology undermines faith in law enforcement and justice systems.

d) Lack of Accountability:

- Opaque systems make it difficult to challenge or audit decisions, raising issues of transparency and due process.

2. Recommended Policies for Responsible Deployment:

a) Mandatory Accuracy & Bias Audits:

- Require regular third-party testing to assess accuracy across different demographic groups using fairness metrics.

b) Consent & Transparency Requirements:

- Inform the public when and how facial recognition is used.
- Limit deployment to situations where there's **informed public awareness or judicial oversight**.

c) Legal Safeguards & Oversight:

- Restrict the use of facial recognition to **serious cases** and under **judicial warrants**.
- Establish **independent ethics boards** or regulatory bodies to monitor and approve deployment.

d) Alternative Verification Methods:

- Use facial recognition only as a **supporting tool**, not the sole basis for identification or arrest.