Part 2: Case Study Analysis (40%)

Case 1: Biased Hiring Tool

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

1. **Identify the source of bias (e.g., training data, model design).**

The bias in Amazon's AI is likely to have emerged from different sources of the AI development workflow. The first source is likely to be the training dataset. The AI was trained on historical data which was heavily imbalanced against women. The data had highest successful recruitments for men than women. As such, the model silently learned and continued to reinforce the prejudice.

As such, the model penalized anything with women or word women based on the training data which also suppressed features with anything to do with women. This means the developed model was an AI model developed by men for men to further dominate women.

This is the reason why there is now a wide call for AI ethics and fairness, to make sure the storm of AI uplifts all boats to the same levels than at the expense of swallowing small ones.

2. **Propose three fixes to make the tool fairer.**

Three fixes to make the tool fairer to everyone is obviously to :

   a. Use balanced and debiased training data. This means train using gender balanced dataset, re-weighting methods to make sure underrepresented groups are fairly represented.

   b. Utilize bias mitigation algorithms – many algorithms since then were introduced including IBM AI Fairness 360.

   c. Introduced transparency and human oversight when building models. This means build explainable models than black box models.

3. **Suggest metrics to evaluate fairness post-correction.**
   Since the introduction of different algorithms and strategies to ethically ensure AI do not further propagate and engrain unfairness and subjugation of any group by another. Different metrics were introduced to evaluate and assess the fairness of AI

models. These metrics would evaluate the actions of the AI post corrections. Three metrics stand out as follows:

Metric	**Explanation**
Disparate Impact Ratio (DIR)	Assess if one group based on the AI decision or model, receives favorable outcomes at a lower rate than the other group. A DIR of <80% potentially signals discrimination
Equal Opportunity Difference	It measures whether the AI model is equally good at making correct positive predictions (true positives) across different groups. This in hiring, are all qualified candidates being selected for interviews at the same rate.
Statistical Parity Difference	Checks on overall outcomes. Do all groups receive positive outcomes at same rate irrespective of their gender, ethnicity or income level etc.

Case 2: Facial Recognition in Policing

Scenario: A facial recognition system misidentifies minorities at higher rates.

Tasks:

Discuss ethical risks (e.g., wrongful arrests, privacy violations).

As an example, police facial recognition has more false positives for criminals on a certain minority group. For example, black over white or Asian over American. This systematic unfairness sparks ethical debates and leads to a whole lot of issues. For instance, it leads to wrongful arrests, which is a human right violation. Continues to propagate fires of racial discrimination which perpetuates silent racism.

Recommend policies for responsible deployment.

Obviously, there will be a need for proper governance and introduction of AI Ethical Policies. These policies should make it an obligation that every system should be explainable, and training data sources be reviewed. Also, each system should undergo continuous audits through human oversight using the AI Fairness 360 and the metrics identified above.