Case Study 1: Biased Hiring Tool (Amazon AI Recruiting)

Scenario:
Amazon’s AI recruiting tool penalized female candidates due to historical bias in hiring data.

Tasks & Answers:

1. Identify the source of bias:

- **Training Data Bias:** Historical resumes favored male candidates, causing the AI to learn gender-biased patterns.
- **Feature Selection Bias:** Certain keywords, experiences, or schools correlated with male candidates.
- **Modeling Bias:** The model prioritized historical hiring outcomes without considering fairness constraints.

2. Three fixes to make the tool fairer:

1. **Debias the Training Data:** Remove gendered indicators and ensure balanced representation of all genders.
2. **Fairness-Constrained Modeling:** Implement algorithms that enforce fairness metrics (e.g., equal opportunity or demographic parity).
3. **Continuous Auditing & Feedback:** Regularly test the model for bias and adjust based on audit results.

3. Metrics to evaluate fairness post-correction:

- **Disparate Impact Ratio:** Compares selection rates between genders.
- **Equal Opportunity Difference:** Measures differences in true positive rates across groups.
- **Statistical Parity:** Ensures equal positive decision rates across genders.
- **False Positive / False Negative Rates:** Compare errors for each gender subgroup.

Case Study 2: Facial Recognition in Policing

Scenario:
A facial recognition system misidentifies minorities at higher rates, leading to potential wrongful arrests and privacy violations.

1. Ethical Risks:

- **Wrongful Arrests:** Higher false positive rates for minority groups can result in innocent individuals being accused or detained.
- **Privacy Violations:** Continuous surveillance can infringe on individuals’ rights to privacy and data protection.
- **Discrimination:** Unequal accuracy reinforces societal biases and marginalizes vulnerable groups.
- **Erosion of Trust:** Communities may distrust law enforcement and technology due to biased outcomes.

2. Recommended Policies for Responsible Deployment:

- **Bias Auditing:** Regularly audit the system using representative datasets to measure accuracy across demographic groups.
- **Human Oversight:** Ensure automated decisions are reviewed by trained human officers before taking action.
- **Transparency:** Publicly disclose system limitations, error rates, and deployment scope.
- **Consent & Legal Compliance:** Adhere to data protection laws (like GDPR) and obtain consent where required.
- **Limited Use Cases:** Deploy facial recognition only for well-defined, critical tasks to minimize misuse.