Case 1: Biased Hiring Tool

Background: Amazon created an AI tool to help with hiring, but it was unfairly rejecting female candidates for technical jobs.

1. Source of Bias

The main problem was **biased training data**. Here's what happened:

- Amazon trained their AI using resumes from the past 10 years. Most of these resumes came from men because the tech industry has historically been male-dominated and the AI learned that "successful" resumes looked like men's resumes, It started penalizing resumes that contained words like "women" or came from all-women's colleges. The model basically learned that being male was a qualification for tech jobs.

2. Three Fixes to Make the Tool Fairer

Fix 1: Balance the Training Data

- Collect equal numbers of successful male and female employee resumes.
- If not enough female resumes exist, use techniques like data augmentation (creating slight variations of existing female resumes).

Fix 2: Add Fairness Constraints

- Program the AI to ignore gender-related keywords completely.
- Set rules that the AI must recommend equal proportions of male and female candidates when qualifications are similar.
- Use "blind" evaluation where gender indicators are hidden from the AI.

Fix 3: Diverse Training Team

- Have women and people from different backgrounds involved in building and testing the system.
- Regular bias audits by diverse teams who can spot problems others might miss.
- Include fairness experts in the development process.

3. Metrics to Evaluate Fairness

Demographic Parity: Measure if the AI recommends men and women at equal rates.

- Example: If 40% of applicants are women, 40% of recommendations should be women.

Equal Opportunity: Check if qualified men and women have equal chances of being recommended.

- Example: If 80% of qualified men get recommended, 80% of qualified women should too.

Calibration: Ensure the AI's confidence scores are equally accurate for both groups.

- Example: When the AI says someone has a 90% chance of success, this should be true for both men and women.

Case 2: Facial Recognition in Policing

Background: Police use facial recognition systems that make more mistakes when identifying Black and Hispanic people compared to white people.

1. Ethical Risks

Wrongful Arrests

- Innocent people get arrested because the AI misidentified them and this is especially dangerous for minorities who are already over-policed as a false match can ruin someone's life, career, and reputation.
- Example: Robert Julian-Borchak Williams was wrongfully arrested in Detroit because of a facial recognition error.

Privacy Violations

- People can't go in public without being tracked and identified which creates a surveillance state where everyone is constantly monitored and this violates the right to anonymity in public spaces.
- Data about people's movements and activities gets stored without consent violating privacy.

Amplifying Existing Bias

- The system makes policing even more unfair than it already is by reinforcing racist stereotypes by generating more false alerts for minorities.
- Creates a feedback loop where minorities get stopped more, creating more data that makes the bias worse.

2. Policies for Responsible Deployment

Strict Accuracy Requirements

- Ban use of facial recognition unless it meets high accuracy standards for ALL demographic groups.
- Require regular testing on diverse populations.
- Mandate that any system with accuracy differences above 5% between groups cannot be used.

Human Oversight Rules

- Facial recognition can only be used as a lead for investigation, never as sole evidence for arrest.
- Require human officers to verify all matches before taking action.
- Officers must be trained on the system's limitations and bias issues.

Transparency and Accountability

- Police departments must publicly report how often they use facial recognition.
- Publish accuracy statistics broken down by demographic group
- Allow independent audits of the system's performance
- Create clear procedures for people to challenge false identifications

Limited Use Cases

- Only allow facial recognition for serious crimes (not minor infractions)
- Prohibit use for general surveillance or crowd monitoring.

- Require warrants for facial recognition searches, just like other surveillance tools.