Case 1: Biased Hiring Tool (Amazon’s AI recruiting tool)

Source of Bias:

The primary source of bias in Amazon's AI recruiting tool was the **training data**. The tool was trained on historical resume data, which predominantly came from male applicants in the tech industry. This historical data reflected existing gender disparities in the industry, and the AI learned to associate certain terms and patterns common in male applicants' resumes (e.g., participation in male-dominated sports, use of specific verbs) with success, while penalizing terms and patterns more common in female applicants' resumes (e.g., attendance at women's colleges, membership in women's organizations). The **model design** itself, if not specifically built with fairness in mind, would perpetuate this bias learned from the data. The algorithm essentially mirrored and amplified the historical biases present in the hiring process.

Proposed Fixes to Make the Tool Fairer:

1. **Data Augmentation and Balancing:** Instead of solely relying on historical data, augment the training data with a more balanced representation of qualified candidates from underrepresented groups. This could involve creating synthetic data (with caution to avoid introducing new biases) or actively seeking out and incorporating resumes from successful female candidates in various tech roles. Techniques like oversampling minority classes or under sampling majority classes can also help balance the dataset.
2. **Bias Mitigation Techniques in Model Training:** Employ algorithmic bias mitigation techniques during the model training phase. These techniques can be applied pre-processing (adjusting the data), in-processing (modifying the training algorithm), or post-processing (adjusting the model's output). Examples include using adversarial debiasing, reweighing data points, or using methods that enforce equalized odds or demographic parity.
3. **Feature Selection and Engineering with Fairness in Mind:** Carefully review and potentially remove or transform features that are highly correlated with protected attributes like gender (e.g., specific names, participation in gender-specific activities) but are not truly indicative of job performance. Focus on skills, experience, and qualifications that are directly relevant to the job requirements, regardless of how they are presented in the resume.

Metrics to Evaluate Fairness Post-Correction:

- **Demographic Parity (or Disparate Impact):** Measure if the selection rate for female candidates is statistically like the selection rate for male candidates. A common rule of thumb is the "80% rule," where the selection rate for the disadvantaged group should be at least 80% of the selection rate for the advantaged group.
- **Equalized Odds:** This metric assesses whether the model has similar false positive rates and false negative rates across different demographic groups. In a hiring context, it would mean the tool is equally likely to incorrectly reject a qualified female candidate as a qualified male candidate, and equally likely to incorrectly flag an unqualified female candidate as a qualified male candidate.
- **Predictive Parity:** This metric checks if the predicted outcome (e.g., likelihood of success) is the same for individuals from different groups who have the same true outcome. In hiring,

this would mean that among truly successful candidates, the model assigns a similar high probability of success regardless of gender.

Case 2: Facial Recognition in Policing

Ethical Risks:

The use of facial recognition systems in policing, especially when they exhibit differential accuracy across demographic groups, poses significant ethical risks:

- **Wrongful Arrests and Incarceration:** Higher misidentification rates for minorities can lead to innocent individuals from these groups being wrongly accused, arrested, and even incarcerated, with devastating consequences for their lives and communities.
- **Exacerbation of Existing Biases:** Such systems can reinforce and amplify existing racial biases within the criminal justice system, leading to disproportionate surveillance and scrutiny of minority communities.
- **Chilling Effect on Civil Liberties:** The pervasive use of facial recognition can create a chilling effect on freedom of assembly and expression, as individuals may be hesitant to participate in protests or public gatherings if they fear being tracked and identified.
- **Privacy Violations:** Constant surveillance through facial recognition infringes on individuals' right to privacy, allowing for the potential tracking of movements and associations without explicit consent or probable cause.
- **Lack of Transparency and Accountability:** Often, the algorithms used in these systems are proprietary, making it difficult to understand how they work, identify sources of error, and hold those responsible accountable for biased outcomes.

Recommended Policies for Responsible Deployment:

- **Accuracy Standards and Independent Testing:** Mandate rigorous independent testing of facial recognition systems to ensure they meet high accuracy standards and demonstrate equal performance across different demographic groups before deployment. Systems that fail to meet these standards should not be used.
- **Clear and Limited Use Cases:** Define specific and narrow use cases for facial recognition in policing, focusing on serious crimes and requiring a reasonable suspicion threshold before activation. Prohibit its use for mass surveillance or in low-level offenses.
- **Human Oversight and Due Process:** Require meaningful human review and verification of any potential matches generated by the facial recognition system. The AI's output should never be the sole basis for an arrest or other significant action. Due process rights must be protected.
- **Transparency and Public Disclosure:** Law enforcement agencies should be transparent about their use of facial recognition technology, including the types of systems used, their capabilities, and how the data is collected, stored, and used. Public reporting on accuracy rates and demographic disparities should be mandatory.
- **Legal and Regulatory Framework:** Establish clear legal and regulatory frameworks governing the use of facial recognition in policing, including guidelines on data retention, access, and the rights of individuals who are subject to surveillance.