Part 2: Case Study Analysis

Addressing AI Bias

Case 1: Biased Hiring Tool (Amazon)

1.Source of Bias:

- Primarily Training Data: The tool was trained on resumes submitted to Amazon over a 10-year period, predominantly from male applicants reflecting the historical male dominance in the tech industry.
- Model Design Reinforcement: The model learned patterns associated with successful hires from that biased dataset. This included penalizing:
- Resumes containing words like "women's" (e.g., "women's chess club captain").
- Graduates from all-women's colleges.
- Potentially undervaluing skills or experiences more commonly found on female applicants' resumes.
- Feedback Loop: The tool's initial biased outputs, if used to screen candidates, would have further reinforced the bias by limiting the diversity of resumes considered successful, creating a harmful feedback loop.

2.Proposed Fixes for Fairer Tool:

a). Comprehensive Data Remediation & Augmentation:

- Audit & Clean: Rigorously audit historical data for gender bias indicators (e.g., gendered keywords, institution bias, skewed representation). Remove or anonymize problematic identifiers where appropriate.
- Augment with Balanced Data: Actively source and incorporate high-quality resumes from diverse female candidates (and other underrepresented groups) who would have been qualified historically but may not have applied or been hired. Partner with women-in-tech organizations.
- Synthetic Data (Cautiously): Generate carefully curated synthetic resumes representing diverse backgrounds and qualifications to

balance the training set, ensuring the synthetic data accurately reflects real qualifications and avoids introducing new biases.

b). Algorithmic Debiasing Techniques:

- Pre-processing: Apply techniques like reweighting or resampling the training data to ensure balanced representation of different demographic groups before model training.
- In-processing: Implement fairness constraints directly within the learning algorithm (e.g., adversarial debiasing where a secondary model tries to predict the protected attribute like gender from the main model's outputs, forcing the main model to learn features invariant to gender).
- Post-processing: Adjust the model's output scores or ranking thresholds specifically for different demographic groups to achieve fairness metrics (e.g., equal opportunity).

c). Human-Centric Design & Oversight:

- Transparent Features: Explicitly remove or neutralize the influence of features highly correlated with gender (e.g., specific college names, certain extracurricular keywords) unless demonstrably job-relevant and validated not to cause disparate impact.
- Focus on Skills & Outcomes: Base predictions primarily on verifiable skills, experiences, and quantifiable achievements relevant to the job description.
- Mandatory Human Review: The tool should never be the sole decision-maker. Its role is to surface potentially qualified candidates for human review, especially candidates near the threshold or flagged for potential bias. Diverse hiring panels are essential.

3.Fairness Evaluation Metrics (Post-Correction):

- Demographic Parity: Compare the selection rate (proportion recommended for interview/hire) between male and female applicants. Significant disparities indicate potential bias ($\Delta SR = |SR_{male} - SR_{female}|$).
- Equal Opportunity: Compare the true positive rate (proportion of actually qualified candidates recommended) between male and female applicants. This ensures qualified candidates from both

groups have an equal chance of being selected (ΔTPR = |TPR_male - TPR_female|). This is often preferred over Demographic Parity for hiring.

- Predictive Parity/Calibration: Check if the predicted probability of success (e.g., "hireability score") is equally reliable for both groups. For example, does a score of 80% for a female candidate mean the same likelihood of being a good hire as an 80% score for a male candidate?
- Impact Ratio: (Selection Rate for Females) / (Selection Rate for Males). A ratio significantly less than 1 indicates adverse impact against females.
- Disaggregated Accuracy Metrics: Report standard accuracy, precision, recall, and F1 scores separately for male and female applicants to identify performance gaps.
- Qualitative Audits: Regularly sample recommended and non-recommended resumes (stratified by gender) for human experts to assess fairness and relevance of the tool's decisions.

Case 2: Facial Recognition in Policing

1.Ethical Risks:

- Wrongful Arrests & Detentions: The most immediate and severe risk. Misidentification, especially at higher rates for minorities, leads to innocent people being arrested, jailed, traumatized, and suffering reputational damage, potentially based on faulty AI evidence.
- Erosion of Due Process & Presumption of Innocence: Over-reliance on "match" results can shortcut investigations and shift the burden of proof onto the accused to disprove the AI's output.
- Exacerbation of Systemic Bias: Higher error rates for minorities amplify existing racial disparities in policing and the justice system, leading to further over-policing and disproportionate incarceration of minority communities.
- Chilling Effects & Privacy Violations: Mass surveillance using FR enables persistent tracking of individuals in public spaces without consent or suspicion, infringing on privacy rights and discouraging lawful activities (e.g., protests, seeking social services) due to fear of being tracked or misidentified.

- Lack of Transparency & Accountability: "Black box" algorithms make it difficult to understand why a misidentification occurred, hindering recourse for victims and accountability for errors. Biased training data often reflects historical policing biases.
- Function Creep: Initial deployment for specific, high-risk use cases (e.g., finding terror suspects) can easily expand to routine surveillance, monitoring protests, or identifying low-level offenders without adequate oversight.
- Undermining Community Trust: Deployment, especially with documented bias, severely damages trust between law enforcement and minority communities, making policing less effective and cooperative overall.

2.Policies for Responsible Deployment:

- Legislative Bans/Restrictions: Advocate for or implement laws prohibiting FR use for:
- Real-time mass surveillance of public spaces.
- Routine police investigations or identifying low-level offenses.
- Any use where the error rate for any demographic group exceeds a strict, legislated threshold (e.g., FNR/FPR < 1% per group).
- Strict Use Case Limitations: If deployment is permitted at all, strictly limit it to:
- Narrow, High-Value Scenarios: e.g., Identifying a specific, violent suspect from a prior high-quality image where there's probable cause and imminent threat, after exhausting traditional investigative methods.
- Ex Post Facto Analysis: Using FR to generate leads after a serious crime has occurred, using high-quality evidence (e.g., clear security footage), never as the sole evidence.
- Mandatory Rigorous Auditing & Transparency:
- Independent Third-Party Audits: Require regular, public audits by independent experts to assess accuracy (disaggregated by race, gender, age), bias metrics, and performance in real-world scenarios.
- Performance Transparency: Mandate public reporting of system performance statistics (error rates per demographic group), usage

logs (frequency, purpose, outcomes), and instances of known misidentifications.

- Algorithmic Transparency (where feasible): Require vendors to disclose training data sources, methodologies, and potential limitations.

➢ High Accuracy & Bias Mitigation Requirements:

- Minimum Performance Standards: Set legislated minimum accuracy thresholds (e.g., 99.9% true match rate) that must be met and maintained across all major demographic groups before deployment is even considered.
- Bias Testing & Mitigation: Require extensive pre-deployment bias testing using diverse benchmark datasets and ongoing monitoring. Mandate the use of state-of-the-art debiasing techniques in model development and training.

➢ Robust Operational Safeguards:

- Human Verification & Corroboration: FR "matches" must always be considered a lead, never probable cause alone. Require confirmation by a trained human examiner and independent corroborating evidence before any detention or arrest.
- Clear Chain of Custody & Documentation: Strictly document the FR process: input image quality, algorithm used, version, confidence score, examiner's analysis, and corroborating evidence.
- Judicial Authorization: Require a warrant based on probable cause for FR searches in non-emergency situations, similar to other digital surveillance tools.
- Right to Challenge & Recourse: Establish clear procedures for individuals to be notified if FR was used against them, access the evidence, and challenge inaccurate results effectively.
- Training: Mandate comprehensive training for officers on FR limitations, bias risks, legal requirements, and operational protocols.

➢ Community Engagement & Oversight: Establish independent civilian oversight boards with the power to review FR use policies, audit logs, and investigate complaints. Engage communities in discussions about potential deployment and its impacts. The strongest policy may often be a complete moratorium on police use of live FR until bias is provably eliminated and robust safeguards are legally enforceable.