Case Study Analysis

Case 1 Biased Hiring Tool

Identify the source of bias

- **Training data bias**: Historical hiring data contained gender imbalance and features correlated with gender (e.g., references to men's clubs, certain universities).
- **Label bias**: Past hires reflect discriminatory practices; model treats those labels as ground truth.
- **Feature selection and proxy variables**: Using features strongly correlated with gender (graduation year, affiliations) allowed the model to infer gender indirectly.

Three fixes to make it fairer

1. **Data-level interventions**
    - **Rebalance / resample** training examples (oversample underrepresented groups or collect more diverse labeled examples).
    - **Remove or transform sensitive proxies**: detect and remove features highly correlated with protected attributes; use representation learning to remove sensitive information.
2. **Preprocessing fairness algorithms**
    - **Reweighing**: compute instance weights so protected groups are represented proportionally during training.
    - **Disparate Impact Remover**: alter feature distributions to reduce dependency on sensitive attributes.
3. **Algorithmic constraints / post-processing**
    - **Adversarial debiasing** or fairness-aware training objectives (penalize differences in selected metrics such as FPR/FNR across groups).
    - **Reject-option classifier / threshold adjustment** to equalize error rates as needed.
4. **Process fixes**
    - Human-in-the-loop review for borderline/rejected candidates; robust documentation and appeals process.

Metrics to evaluate fairness post-correction

- **Statistical parity difference / Demographic parity** (difference in positive selection rates).
- **Disparate impact ratio** (ratio of selection rates).
- **Equal opportunity / TPR parity** (difference in true positive rates).
- **Equalized odds** (differences in both FPR and TPR).
- **False Positive Rate (FPR) and False Negative Rate (FNR) differences** between protected groups.
- **Calibration within groups** (are predicted scores equally interpretable across groups).

- **Aggregate utility metrics** (precision, recall) per-group to ensure no large accuracy drop for any group.

Case 2 Facial Recognition in Policing

Ethical risks

- **Wrongful arrests**: Higher misidentification rates for minorities can lead to false suspicion, detentions, arrests severe harm.
- **Privacy violations & surveillance**: Mass deployment may enable continuous tracking, chilling effects on public life and protests.
- **Discrimination & disparate enforcement**: Systemic biases can intensify existing inequalities (over-policing of specific communities).
- **Due process & consent issues**: Use without notice or oversight threatens civil liberties.
- **Mission creep**: Tools deployed for limited use can be repurposed (e.g., from serious-crime detection to low-level surveillance).

Recommended policies for responsible deployment

1. **Scope and limitation**
    - Restrict use to well-defined, high-priority cases (e.g., identifying suspects of serious violent crimes), not for mass surveillance.
2. **Human oversight**
    - Require human confirmation before arrest; facial-recognition output is investigatory evidence, not conclusive proof.
3. **Accuracy thresholds & audits**
    - Minimum performance standards disaggregated by demographic group; fail-safe procedures when accuracy below thresholds.
4. **Transparency & public notice**
    - Public documentation of where, when, and how systems are used; regular public impact reports.
5. **Privacy protections**
    - Data minimization, retention limits, encryption, and strict access controls.
6. **Independent testing and audits**
    - Third-party audits for bias and accuracy; public release of evaluation datasets and results where safe.
7. **Legal oversight & redress**
    - Clear legal frameworks authorizing use, and accessible mechanisms for individuals to challenge and obtain remediation for mistakes.