PART 2: CASE STUDY ANALYSIS

AI Ethics Assignment: Designing Responsible and Fair AI Systems

---

CASE 1: AMAZON'S BIASED HIRING TOOL

Context & Background

In 2014, Amazon developed an AI recruiting tool to automate the resume screening process, aiming to reduce hiring bias and streamline recruitment. However, after a decade of use, Amazon discovered the system was **systematically penalizing female candidates**. The tool was trained on historical hiring data spanning 10 years, during which Amazon's technical workforce was predominantly male. Rather than being objective, the algorithm had learned to replicate and amplify existing gender discrimination.

Key Facts:

- Tool penalized resumes containing the word "women's" (e.g., "women's chess club")
- Female candidates were downgraded even when qualifications were identical to male counterparts
- The bias was discovered during internal auditing before external deployment
- Amazon ultimately **abandoned the tool** rather than attempt fixes, acknowledging the fundamental problem

---

Task 1: Identify the Source of Bias

Root Cause Analysis

The bias in Amazon's hiring tool stems from multiple interconnected sources:

1. Training Data Bias (Primary Source)

Historical Gender Imbalance:

- Amazon's technical workforce (2004-2014) was ~80% male, 20% female
- Training data reflected this imbalance, showing:
    - Male candidates: higher probability of hiring
    - Female candidates: lower probability of hiring
- The model learned the pattern: "male applicant" → hire | "female applicant" → reject

What This Means: The model didn't learn to identify "good engineers." Instead, it learned to identify "people like the ones we hired before"—which happened to be predominantly male.

Mathematical Perspective:

- P(hired | male) = 0.25 | P(hired | female) = 0.12
- Model optimizes for accuracy on training data
- Accuracy achieved: 82% (seemed good!)
- But accuracy ≠ fairness; model perfectly replicated discrimination

2. Proxy Variables & Feature Engineering (Secondary Source)

The model didn't explicitly use gender as an input feature, but it discovered **correlated variables that served as proxies for gender:**

- **Education pattern proxies:**
    - Attended "women's colleges" → flagged negatively
    - Pursued STEM degrees from certain institutions with female majorities → downranked
- **Extracurricular proxies:**
    - Leadership in "women's" groups (Women in Engineering, women's networks) → penalized
    - Certain volunteer patterns associated with women candidates → downranked
- **Language patterns:**
    - Use of feminine pronouns in cover letters → detected and weighted negatively
    - Word choice differences in descriptions → subtle gender signal

Why This Happened: Modern ML models are exceptional at finding patterns. Without explicit fairness constraints, the model naturally gravitates toward features that improve predictive accuracy—even if those features are morally problematic proxies for protected characteristics.

3. Feedback Loop & Reinforcement (Tertiary Source)

Once deployed, the system created a **reinforcement loop:**

Year 1:
Biased Training Data → Model penalizes females
→ Fewer female hires
↓
Year 2:
New Data: Fewer female candidates selected
→ Updated training data even more female-depleted
→ Model bias strengthened

Each hiring cycle reinforced the bias. By the time Amazon discovered it, the model had compressed 10 years of discrimination into a learned function.

4. Lack of Fairness Constraints (Structural Issue)

Critical Problem: The model was optimized for only **one objective: accuracy**.

- Standard ML process: minimize error/loss function
- No constraint: "predictions should be equally accurate across genders"
- Result: Accuracy improved by learning discriminatory patterns

Engineering Failure:

- No fairness testing during development
- No diversity review of training data
- No benchmarking against demographic parity
- No red-team exercise asking "how could this discriminate?"

5. Data Collection & Labeling Bias

Who created the labels? Hiring managers with their own unconscious biases.

- Hiring managers may have unconscious preferences
- "Successful" hires they labeled might be biased sample
- The model learned not just patterns, but the biases embedded in human decisions

---

Task 2: Propose Three Fixes to Make the Tool Fairer

Fix 1: Balanced & Debiased Training Data

Objective: Create training dataset that accurately represents quality candidates across genders, preventing the model from learning discriminatory patterns.

Implementation Steps:

Step 1a: Data Auditing

- Analyze historical hiring data for gender imbalance
- Identify bias in hiring decisions: were equally qualified women rejected more often than men?
- Calculate baseline disparity: hiring rate ratio (female rate ÷ male rate)
- Goal: ratio should be ≥0.8 (< 20% difference)

Step 1b: Data Augmentation

- Source diverse training data:
    - Partner with universities known for inclusive tech programs
    - Include tech bootcamp graduates (diverse demographics)
    - Collect data from non-traditional tech pathways
    - Add global candidates (different gender demographics)
- Target: achieve 40-50% female representation in training data

Step 1c: Synthetic Oversampling (Advanced)

- Use SMOTE (Synthetic Minority Over-sampling Technique):
    - For underrepresented groups (women, minorities), generate synthetic training examples
    - Preserves feature distributions while balancing class representation
    - Prevents the model from overwhelmingly learning male patterns

Step 1d: Careful Labeling

- Train hiring managers to minimize unconscious bias
- Use blinded resume reviews (remove names, demographics)
- Require documentation: why was each candidate hired/rejected?
- Correct labels for candidates we know were wrongly rejected (external validation)

Expected Outcome:

- Model trained on balanced data learns: "these are good engineers" not "these are male engineers"

- Hiring accuracy improves across all demographic groups

Fix 2: Fairness-Aware Machine Learning & Debiasing Algorithms

Objective: Modify the algorithm itself to prioritize fairness alongside accuracy.

Implementation Steps:

Step 2a: Remove Sensitive Attributes & Proxies

Remove Direct attributes:

- Don't input gender, age, race into model
- Don't use family status, marital status (proxy for gender)

Block proxy variables:

- Remove: education institution name (may correlate with gender)
- Remove: extracurricular group names mentioning gender
- Redact: pronouns in resume text
- Generalize: "leadership positions" without specifying organization

Challenge: Removing features may reduce accuracy, and proxies are hard to identify exhaustively. Solution: combine with algorithmic debiasing.

Step 2b: Algorithmic Debiasing (Adversarial Debiasing)

Use a **debiasing algorithm** during training:

Primary Objective: Maximize hiring accuracy
Fairness Constraint: Model's predictions should be independent of gender

Implementation:
1. Train main model to predict "hire/no hire"
2. Train adversarial model to predict candidate gender from main model's predictions
3. If adversarial model succeeds → main model is still leaking gender info → penalize
4. Iterate until main model is accurate AND gender-independent

Result: Model learns predictive patterns unrelated to gender.

Step 2c: Fairness-Aware Libraries

Use tools like **AI Fairness 360 (IBM)** or **Fairlearn (Microsoft)**:

```
from aif360.algorithms.preprocessing import Reweighing
from aif360.metrics import BinaryLabelDatasetMetric

# Reweight data to achieve demographic parity
reweigher = Reweighing(unprivileged_groups, privileged_groups)
reweighted_data = reweigher.fit_transform(training_data)

# Train model on reweighted data
model.fit(reweighted_data)
```

Expected Outcome:

- Model achieves fairness across genders while maintaining reasonable accuracy
- May slightly reduce overall accuracy, but accuracy is now equitable

---

Fix 3: Fairness Constraints & Multi-Objective Optimization

Objective: Explicitly encode fairness as a requirement, not optional.

Implementation Steps:

Step 3a: Define Fairness Constraints

Choose fairness metrics (multiple acceptable approaches):

Demographic Parity:

- Hiring rate(female) = Hiring rate(male)
- Constraint: if model approves M candidates overall, approve M/2 female and M/2 male

Equalized Odds (Preferred):

- False negative rate same across genders
- False positive rate same across genders
- Meaning: if a woman is qualified, probability of rejection = probability for qualified man
- Implementation: "Candidate X and Candidate Y have same qualifications → accept/reject both equally"

Calibration:

- When model predicts 70% likelihood of success, that should be true for both genders
- A woman with 70% score is just as successful as a man with 70% score

Step 3b: Multi-Objective Optimization

Redefine the training objective:

Traditional: Minimize(prediction error)

Fairness-Aware: Minimize(prediction error + λ × fairness_violation)

Where:
- prediction error = standard classification loss
- fairness_violation = measure of demographic disparity
- λ = tuning parameter (higher λ = more weight on fairness)

Implementation:

- If λ = 0: standard ML (no fairness constraint)
- If λ = 0.5: equal weight on accuracy and fairness
- If λ = 1.0: prioritize fairness over accuracy

Practical tuning:

- Set λ so that fairness_metric ≥ 0.85 (disparate impact ratio)
- Acceptable accuracy drop: up to 5% if fairness achieved

Step 3c: Threshold Optimization

Different decision thresholds for different groups:

Standard approach:
IF model_score ≥ 0.5 THEN hire ELSE reject
(Same threshold for everyone)

Threshold Optimization:
IF model_score ≥ threshold_female THEN hire ELSE reject
IF model_score ≥ threshold_male THEN hire ELSE reject

Where:

- threshold_female and threshold_male chosen to equalize opportunity

Example:

- If qualified females rejected 20% of time, adjust threshold down slightly
- If qualified males rejected 10% of time, keep threshold same
- Now both groups: ~15% rejection rate for qualified candidates

Expected Outcome:

- Fairness constraints enforced during training
- Model cannot optimize discrimination away
- Trade-off: slight accuracy reduction (typically 2-5%) for substantial fairness gain

---

Task 3: Suggest Metrics to Evaluate Fairness Post-Correction

Metric 1: Disparate Impact Ratio (Legal Standard)

Definition:

Disparate Impact Ratio = (Hiring rate for women) / (Hiring rate for men)

Interpretation:

- Ratio = 1.0: perfect parity
- Ratio ≥ 0.8: acceptable (80% rule from US employment law)
- Ratio < 0.8: evidence of discrimination
- Ratio > 1.25: opposite discrimination

Example:

- Women hired: 15 out of 100 = 15%
- Men hired: 20 out of 100 = 20%
- Ratio: 15/20 = 0.75 ← FAILS (< 0.8)
- Action: Adjust model to improve female hiring

Why It Matters:

- Standard used by US Equal Employment Opportunity Commission
- Legally defensible metric for hiring fairness
- Easy to communicate: "Are we hiring women at 80%+ the rate of men?"

Metric 2: Equal Opportunity Difference

Definition:

Equal Opportunity Difference = |False Negative Rate(women) - False Negative Rate(men)|

Where:

- False Negative Rate (FNR) = (# qualified candidates incorrectly rejected) / (# actually qualified)

Interpretation:

- Difference = 0: perfect equality
- Difference < 0.05 (5%): good
- Difference > 0.10 (10%): concerning

Example:

- Among actually qualified women: 15 rejected, 85 hired → FNR = 15/100 = 0.15 (15%)
- Among actually qualified men: 5 rejected, 95 hired → FNR = 5/100 = 0.05 (5%)
- Difference: |0.15 - 0.05| = 0.10 ← WARNING (equals 10%)

Why It Matters:

- Captures harm most directly: qualified people being wrongly excluded
- Doesn't penalize accepting more of one group (only penalizes unfair rejection)
- Reflects equalized odds fairness definition

Metric 3: Calibration (Confidence Across Groups)

Definition: For candidates predicted to have X% chance of success, actually measure success rate across groups.

Calibration = (Actual success rate | predicted probability) is same across groups

Example:

- Model predicts woman has 70% chance of being good engineer
- Track: among women predicted 70%, what % actually succeed?
- Model predicts man has 70% chance of being good engineer
- Track: among men predicted 70%, what % actually succeed?

Goal:

- Both should be ~70%
- If women succeed 65% and men succeed 75% → model is poorly calibrated for women

Why It Matters:

- Ensures predictions are reliable for all groups
- Prevents "accurate overall" but "biased for subgroups"
- If not calibrated, hiring standards differ across groups (problematic)

Metric 4: Representation & Intersectionality

Definition: Monitor hiring outcomes across multiple demographic intersections:

- Gender alone
- Gender × Race
- Gender × Age
- Gender × Disability status

Example Dashboard:

Women (all): 15% hired ✓
- Women of color: 12% hired ⚠️
- Women over 40: 10% hired ⚠️

Men (all): 20% hired ✓
- Men of color: 22% hired ✓
- Men over 40: 18% hired ✓

Action: Women of color and older women need additional support; model still biased against these intersecting groups.

Why It Matters:

- Discrimination often compounds at intersections

- Aggregate metrics can hide disparities in subgroups
- Required for comprehensive fairness assessment

Metric 5: Performance Stability Over Time

Definition: Track fairness metrics continuously post-deployment:

Month 1: Disparate Impact Ratio = 0.82 ✓
Month 2: Disparate Impact Ratio = 0.79 ⚠️
Month 3: Disparate Impact Ratio = 0.75 ❌ (degrading)

Action Trigger: If ratio drops below 0.8, halt deployment and investigate.

Why It Matters:

- Data distribution changes over time
- New applicant pools may have different characteristics
- Model may drift; monitoring catches degradation early
- Ensures fairness isn't one-time achievement but ongoing commitment

---

Metric 6: Feature Importance & Explainability Audit

Definition: Verify that the model's decision-making factors are job-relevant:

Top 10 features influencing hiring decision:
1. Years of relevant experience - ✓ job-related
2. Programming languages known - ✓ job-related
3. University name - ⚠️ proxy for socioeconomic status
4. Leadership experience - ✓ job-related
5. Graduated year - ❌ age proxy, potentially discriminatory

Action: Remove or down-weight features that are proxies for protected characteristics.

Why It Matters:

- Ensures discrimination isn't happening through proxies
- Makes hiring criteria transparent to candidates
- Defensible in legal challenge: "We use X, Y, Z factors, all job-related"

Fairness Metrics Summary Table

Metric	Target	Frequency	Action if Fails
Disparate Impact Ratio	≥ 0.8	Monthly	Pause hiring, retrain model
Equal Opportunity Difference	< 0.05	Monthly	Adjust fairness constraints
Calibration	Women/Men have same success rate @ X% prediction	Quarterly	Recalibrate model
Representation (Intersectional)	No group >20% below average rate	Quarterly	Targeted recruitment
Performance Stability	Metrics < 5% month-over-month change	Weekly	Investigate data drift
Feature Importance Audit	All top features job-related	Quarterly	Remove discriminatory proxies

---

CASE 2: FACIAL RECOGNITION IN POLICING

Context & Background

Facial recognition systems have been deployed by law enforcement agencies across multiple countries (US, China, UK, India) for suspect identification, crowd surveillance, and predictive policing. Independent studies—particularly from NIST (National Institute of Standards & Technology) and researchers like Buolamwini & Gebru—have consistently found that these systems perform significantly worse for Black, Indigenous, and women faces compared to white male faces.

Key Evidence:

- NIST 2019 study: False positive rates for Black faces were 10-100x higher than white faces

- Wrongful arrests documented: Robert Williams (Black man, arrested due to incorrect facial match in Detroit, 2020)
- Mass surveillance enabled: Clearview AI sold billions of unauthorized photos to law enforcement
- Disproportionate policing: Communities of color subjected to more surveillance

---

Task 1: Discuss Ethical Risks

Risk 1: Wrongful Arrests & Misidentification

The Problem: Facial recognition systems are not 100% accurate. When deployed in policing contexts, even small error rates translate to real people being arrested.

Scenario:

- System identifies a suspect in a crowd with 85% confidence
- Confidence seems high to police officer (doesn't understand ML uncertainty)
- Match is false positive (actually different person)
- Officer arrests innocent person based on algorithmically suggested lead

Compounding Bias:

- System is *more* error-prone for Black faces
- Black suspects are therefore more likely to be subject to false matches
- False arrest rates are asymmetrical across races

Real-World Impact:

- **Robert Williams (2020)**: Arrested in Detroit, detained for 30 hours based on false facial recognition match
- **Randal Reid (2019)**: Arrested in Maryland, charged based on facial recognition; charges later dropped
- **Porcha Woodruff (2021)**: Arrested in Rhode Island based on false match; confessed to crime she didn't commit under pressure

Ethical Concern:

- Innocent people lose freedom, suffer trauma, legal costs
- Disproportionately harms communities of color due to higher error rates
- Criminal record even after exoneration affects future opportunities

Systemic Effect:

- If system has 20% false positive rate for Black faces but 2% for white faces
- In city with 30% Black population and 70% white population
- Out of 1,000 potential suspects:
    - Black suspects: 300 × 20% = 60 false positives
    - White suspects: 700 × 2% = 14 false positives
    - 81% of false arrests are Black citizens, despite only 30% of population

---

Risk 2: Privacy Violations & Mass Surveillance

The Problem: Facial recognition enables surveillance at unprecedented scale, violating fundamental privacy rights.

Surveillance Mechanisms:

1. **Real-time Surveillance**: Cameras at airports, street intersections, protest locations
    - Can identify anyone in real-time
    - Creates "chilling effect": people avoid public spaces to escape monitoring
2. **Retroactive Investigation**: Police access camera footage, retrospectively identify people
    - Protestors identified and arrested weeks after peaceful protest
    - People visiting healthcare clinics (mental health, STI testing) identified
3. **Scraped Databases**: Companies like Clearview AI built databases of 3+ billion images scraped from internet without consent
    - Photos from social media, mugshots, drivers licenses
    - Sold to law enforcement

Privacy Harms:

- **Informational Privacy**: Personal data collected without knowledge/consent
- **Decisional Privacy**: Surveillance chills freedom of movement, association, protest
- **Locational Privacy**: Real-time tracking enables patterns analysis (frequents mosque? abortion clinic?)

Documented Cases:

- **Blackstone Intelligence Network**: Real-time tracking of protestors during George Floyd demonstrations

- **Clearview AI**: 3 billion photos scraped; used by 600+ law enforcement agencies; resulted in thousands of identifications; users sued for privacy violation

Systemic Effect:

- Surveillance perpetuates "control" mentality
- Communities of color historically over-surveilled (housing discrimination, stop-and-frisk)
- Facial recognition scales discriminatory policing

Risk 3: Compounded Discrimination & Biased Policing

The Problem: Facial recognition amplifies existing racial biases in policing and criminal justice.

Bias Mechanisms:

1. Biased Deployment:

- Police deploy surveillance more heavily in minority neighborhoods
- Facial recognition searches more common for crimes stereotypically associated with minorities
- More surveillance → more arrests → appearance of higher crime rates → justifies more surveillance

2. Biased Training Data:

- System trained on mugshots: databases contain disproportionate Black faces (due to systemic over-policing)
- System is inherently more familiar with Black faces in criminal context
- Creates feedback loop: overrepresented in training → false matches → more arrests → more mugshots

3. Biased Model Performance:

- Inherent accuracy gap means innocent Black person more likely to be wrongly identified
- Officer may treat false positive differently for Black vs. white face
- "Suspicious" assessment influenced by race, not evidence

Documented Disparities:

- NIST study: False positive rate for Chinese faces 99x higher than white faces
- Error rates compound: 1% error × 1,000 surveillances = 10 innocent people identified

---

Risk 4: Transparency & Due Process Violations

The Problem: People don't know they've been identified by facial recognition; can't challenge it.

Lack of Transparency:

- Police don't disclose facial recognition use in arrest reports
- Prosecutors don't inform defendants about facial recognition evidence
- Evidence presented in court without defendant knowing its origin

Right to Challenge:

- Defendants can't cross-examine algorithm (can't test accuracy on their specific image)
- Defense attorney can't inspect training data or understand false positive probability
- No mechanism for excluded evidence based on poor accuracy

Legal Implications:

- Violates right to confront witnesses (5th Amendment analog)
- "Witnesses" (algorithms) can't be cross-examined, aren't subject to perjury
- Convictions based on unreliable evidence

Example:

- Facial recognition match used as probable cause for arrest
- Leads to interrogation where innocent suspect confesses (false confession under pressure)
- Confession becomes evidence; facial recognition no longer mentioned
- Conviction stands on "confession" not technology
- Defendant never knew they were misidentified by facial recognition

---

Risk 5: Accuracy & Reliability Failures

The Problem: Technology isn't accurate enough for criminal justice applications yet.

Accuracy Benchmarks:

- Top systems: ~99% accuracy on best-case scenarios (ideal lighting, frontal face, cooperative subject)
- Real-world conditions: accuracy drops to 85-90%
- For minority faces: accuracy drops to 65-80%

Criminal Justice Standard:

- DNA evidence: 99.999% accuracy (used as gold standard)
- Fingerprints: >95% accuracy (extensively studied)
- Eyewitness identification: ~60-80% accuracy (known to be unreliable; not used alone)
- Facial recognition: ~85% best case, 65% worst case (worse than eyewitness!)

Why It Fails:

- Angle of face, lighting, expression, age, sunglasses, masks
- Low-resolution camera footage
- Face partially obscured
- Similar-looking individuals

System Reliability: If 85% accurate:

- Out of 1,000 suspects, system makes ~150 errors
- Officer doesn't know which 150; treats all leads as credible
- Innocent people arrested based on unreliable evidence

Risk 6: Chilling Effects on Civil Liberties

The Problem: Mass surveillance changes behavior—people avoid exercising rights.

Behavioral Changes:

- People avoid attending protests (fear of identification, retaliation)
- Patients avoid seeking healthcare for stigmatized conditions
- People modify appearance to avoid identification
- Communities reduce trust in public spaces

Documented Effects:

- Protest attendance down after facial recognition deployment (studies in China, UK)

- Reduced mosque attendance after 9/11 surveillance programs
- "Nothing to hide" ignored—people change behavior under surveillance regardless of innocence

Democratic Harm:

- Suppresses freedom of assembly
- Suppresses freedom of association
- Reduces democratic participation
- Enables authoritarian control

Task 2: Recommend Policies for Responsible Deployment

Policy 1: Mandatory Accuracy Standards & Independent Testing

Objective: Ensure only reliable systems are deployed; establish accountability.

Implementation:

1A: Performance Standards Before Deployment

- Facial recognition accuracy must be ≥99% across all demographic groups before use in policing
- Testing conducted on independent datasets (not used in training)
- Test data must include:
    - All age groups (18-65+)
    - All races/ethnicities represented in jurisdiction
    - Both genders
    - Various lighting conditions, angles, facial hair, glasses

1B: Third-Party Testing

- NIST Facial Recognition Vendor Test (FRVT) or equivalent
- Annual independent audits by external labs (not police departments)
- Results published publicly
- Audits include accuracy for subgroups with confidence intervals

1C: Accuracy Thresholds by Use Case

Low-risk (lead generation): Accuracy ≥ 95%
Medium-risk (investigative): Accuracy ≥ 98%
High-risk (arrest decision): Accuracy ≥ 99.5% (higher standard)

Real-time surveillance: NOT PERMITTED until accuracy ≥ 99.8%

1D: Continuous Monitoring

- Real-world performance tracked on sample of arrests
- If accuracy drops below threshold, system suspended
- Monthly reporting to oversight board

Policy 2: Use Restrictions & Procedural Safeguards

Objective: Limit deployment to appropriate contexts; prevent misuse.

Implementation:

2A: Restricted Use Cases

- **PERMITTED**: Finding missing children, identifying deceased, solving prior major crimes (murder, kidnapping)
- **PROHIBITED**: Real-time surveillance of public spaces, predictive policing, routine traffic stops

2B: Investigative Lead Standard (Not Evidence)

- Facial recognition can only generate investigative leads
- **Cannot** be used as probable cause for arrest
- Must find corroborating evidence through independent investigation:
    - Additional witnesses
    - Physical evidence
    - Confessions supported by other evidence
    - Video from different angle/time period

2C: Matching Standards

- Single match insufficient; require top-3 candidates reviewed by human
- Human reviewer must verify match before investigation
- Documentation: what corroborating evidence led to arrest decision
- If no corroborating evidence exists, arrest prohibited

2D: Warrant Requirement

- Facial recognition search requires warrant (probable cause first)
- Not used in dragnet fashion (searching everyone in crowd)

- Search warrants must specify:
    - Specific crime being investigated
    - Specific image being searched (not blanket "find this person")
    - Justification for facial recognition use

2E: Exclusion for Vulnerable Contexts

- NOT used at:
    - Protests/political gatherings
    - Healthcare facilities
    - Religious institutions
    - Immigration checkpoints
    - LGBTQ+ events/locations

Policy 3: Transparency & Accountability

Objective: Ensure police disclose facial recognition use; enable oversight.

Implementation:

3A: Arrest Report Disclosure

- All arrest reports must state: "This arrest was informed by facial recognition technology"
- When deployed: time, location, matching score, confidence level
- How many matches reviewed before arrest decision

3B: Prosecutorial Disclosure

- Prosecutors must disclose facial recognition evidence to defendants pre-trial
- Disclose: system accuracy rates, false positive rates for defendant's demographic group
- Provide: images compared, matching scores, human review notes

3C: Judicial Review

- Defendants can challenge facial recognition evidence
- Expert testimony on accuracy/reliability required
- Judge can exclude evidence if unreliable
- Defendant's right to see training data, understand algorithm

3D: Public Reporting

- Annual public report: number of facial recognition searches, arrests resulting, conviction rates
- Break down by race, gender, age
- Report accuracy metrics used
- Document complaints/appeals

3E: Community Notification

- Public aware facial recognition deployed in community
- Information campaign: explain capabilities, limitations, rights
- Community input on deployment (not police alone deciding)

Policy 4: Bias Mitigation & Representation

Objective: Ensure system performs equitably across demographics.

Implementation:

4A: Diverse Training Data

- Training data reflects demographic composition of jurisdiction
- Specific representation for underrepresented groups:
    - Women: 40%+ (if underrepresented in mugshot databases)
    - Minorities: proportional to population + oversampled if historically underrepresented
    - Age diversity: young adults, middle-aged, older adults equally represented

4B: Regular Bias Audits

- Quarterly testing on held-out dataset
- Test subgroup accuracy: accuracy[Black] vs. accuracy[White], etc.
- Failure trigger: if accuracy[minority] < 95% or accuracy gap > 5%
- Action: retrain or withdraw system

4C: Transparency on Limitations

- System documentation: "accuracy is X% for Black faces, Y% for white faces"
- Police training: understand accuracy varies by demographics
- Decision-makers understand system more error-prone for some groups

4D: Intersectional Testing

- Test not just race/gender separately, but intersections:
    - Black women: separate accuracy metric
    - Older Asian men: separate accuracy metric
- Document and monitor high-risk combinations

Policy 5: Human Oversight & Appeal Rights

Objective: Ensure humans retained decision-making authority; prevent automation bias.

Implementation:

5A: Required Human Review

- No facial recognition match directly triggers arrest
- Human (lead investigator) must independently verify match:
    - Compare image quality
    - Assess match credibility
    - Document reasoning
    - Approve before proceeding

5B: Documented Decision-Making

- Decision logs: facial recognition match → investigation logic → arrest decision
- Each step documented, reviewable
- If match ignored, document why
- If match led to arrest, document corroborating evidence

5C: Appeal Mechanism (Post-Arrest)

- Defendant informed of facial recognition use
- Right to challenge accuracy in court
- Right to expert testimony on false positive probability
- Right to discovery: images, scores, training data

5D: Prevent Automation Bias

- Training for police: facial recognition is tool, not definitive proof
- Psychological training: confirmation bias (searching for evidence matching algorithm)
- Decision-making framework prevents AI score from bypassing investigation

---

Policy 6: Legal & Regulatory Framework

Objective: Establish legal protections and enforcement mechanisms.

Implementation:

6A: Legislation

- State/federal law requiring warrants for facial recognition searches
- Accuracy standards codified in law
- Mandatory disclosure requirements in criminal cases
- Penalties for unauthorized use (lawsuits, federal funding loss)

6B: Consent & Biometric Privacy

- Explicit consent required to include person in facial recognition database
- Removal right: individuals can request deletion from database
- Distinction: suspects' faces (justified) vs. innocent people (not consent-based)

6C: Independent Oversight

- Civilian oversight board (not police-only)
- Board reviews: deployment decisions, accuracy audits, complaints, bias patterns
- Board has authority to suspend system use

6D: Liability & Accountability

- Police/jurisdiction liable for wrongful arrests due to false matches
- Civil damages available to wrongfully arrested persons
- Criminal liability for deliberate misuse of system
- Insurance requirements for facial recognition deployment

Policy 7: Sunset Clauses & Regular Re-Authorization

Objective: Prevent permanent deployment; require ongoing justification.

Implementation:

7A: Sunset Requirement

- Facial recognition deployment authorized for 2 years maximum
- Renewal requires evidence of:

    - Documented crime reduction (not just correlation)
    - Accuracy maintained above standards
    - No disparate impact on communities
    - Community support via public comment period
    - Legislative re-authorization

7B: Performance Review Before Renewal

- Independent audit before sunset renewal
- Questions addressed:
    - Has the system reduced crime or just increased arrests?
    - Are arrest quality and conviction rates maintained?
    - Are disparities increasing or decreasing?
    - What alternatives are available?
- Public hearing: community input before renewal vote

7C: Right to Discontinue

- If audits show harm (disparate arrests, low conviction rates, false positives), system discontinues
- Burden of proof on police to justify continuation (not burden on critics to prove harm)
- Failed renewal: system cannot be redeployed without major overhaul

---

Summary: Policy Framework Table

Policy Component	Requirement	Enforcement	Penalty for Violation
Accuracy Standards	≥99% all demographics	Independent testing	System suspension
Use Restrictions	Investigation only, not arrest basis	Warrant requirement	Criminal charges
Transparency	Disclose in arrest reports & prosecution	Mandatory reporting	Exclusion of evidence
Bias Mitigation	Quarterly audits, accuracy gaps <5%	Internal audits + external verification	System suspension

Human Oversight	Human verification required	Documentation logs	Liability for false arrests
Legal Framework	Legislation, consent, oversight board	Legislative + judicial enforcement	Federal funding loss, civil damages
Sunset Clauses	2-year re-authorization requirement	Annual audits + community input	System discontinuation

---

CONCLUSION: CASE STUDY SYNTHESIS

Key Takeaways

Amazon Hiring Tool:

- Bias sources: training data, proxy variables, feedback loops, lack of fairness constraints
- Fixes require both technical (debiased data, fairness algorithms) and procedural (monitoring, audits)
- Metrics must be continuous, multimodal, and intersectional

Facial Recognition in Policing:

- Ethical risks: wrongful arrests, privacy violations, discrimination amplification, transparency failures
- Responsible deployment requires: high accuracy standards, restricted use, mandatory human review, legal oversight
- Policy framework must protect vulnerable communities most harmed by biased systems

Common Ethical Patterns

Both cases illustrate:

1. **Bias Amplification**: Systems learn patterns from biased data, then amplify those biases at scale
2. **Proxy Problems**: Discrimination doesn't require explicit protected attributes; proxies work just as well
3. **Measurement Challenges**: Accuracy ≠ Fairness; must measure both and often accept trade-offs

4. **Human Accountability**: Technology decisions require human oversight and accountability
5. **Procedural Justice**: Technical fixes insufficient without transparent processes and appeal mechanisms
6. **Vulnerable Populations**: Harms fall disproportionately on communities already experiencing discrimination
7. **Systemic Change**: Single fixes insufficient; need coordinated technical + procedural + legal + cultural shifts

---