

Bias in AI



Bias Sources

AI inherits biases from training data, risking unfair discrimination.



Real-World Examples

Facial recognition often misidentifies minorities; hiring algorithms favor certain groups.



Mitigation Methods

Techniques include data rebalancing and adversarial debiasing to reduce bias.



Persistent Challenges

Hidden biases in complex AI models remain difficult to fully measure or eliminate.

Fairness in AI

Fairness ensures AI decisions do not disadvantage groups. This requires inclusive datasets and equitable algorithm design.



Demographic Parity

Ensures outcomes are independent of sensitive attributes like race or gender.



Equal Opportunity

Focuses on equal true positive rates across groups to prevent discrimination.



Social Justice

Promotes non-discriminatory outcomes consistent with UNESCO's ethical principles.

AI Ethics & Responsible Development Assignment

Part 1: Theoretical Understanding

1. Short Answer Questions

Q1: Define algorithmic bias and provide two examples

Algorithmic bias refers to systematic and repeatable errors in AI systems that create unfair outcomes, often disadvantaging certain groups based on protected characteristics like race, gender, age, or socioeconomic status. This bias typically stems from biased training data, flawed model design, or prejudiced implementation decisions.

Two examples:

1. **Criminal Justice Risk Assessment:** The COMPAS algorithm has shown higher false positive rates for Black defendants, incorrectly predicting they would reoffend at higher rates than white defendants, leading to harsher sentencing recommendations.
2. **Healthcare Resource Allocation:** An algorithm used by US hospitals to identify patients needing extra medical care systematically underestimated the needs of Black patients because it used healthcare spending as a proxy for health needs. Since Black patients historically had less access to healthcare and lower spending, they appeared "healthier" to the algorithm despite having greater medical needs.

Q2: Explain the difference between transparency and explainability

Transparency refers to openness about an AI system's existence, purpose, data sources, and overall functioning. It answers questions like: "What data was used to train this model?", "Who developed it?", "What is its intended purpose?" Transparency is about disclosure and accountability at the system level.

Explainability (or interpretability) refers to the ability to understand and articulate how an AI system makes specific decisions. It answers: "Why did the model make this particular prediction?", "Which features influenced this outcome?" Explainability operates at the decision level.

Why both are important:

- **Transparency** builds trust, enables oversight, and allows stakeholders to identify potential conflicts of interest or inappropriate applications
- **Explainability** enables debugging, validates decision-making processes, supports legal compliance (especially for consequential decisions), and empowers users to contest unfair outcomes
- Together, they enable meaningful accountability—transparency tells you what exists, explainability tells you how it works

Q3: How does GDPR impact AI development in the EU?

The General Data Protection Regulation significantly impacts AI development through several key provisions:

1. **Right to Explanation (Article 22):** Individuals have the right not to be subject to decisions based solely on automated processing that significantly affects them. This requires AI systems making consequential decisions to provide meaningful explanations.
2. **Data Minimization and Purpose Limitation:** AI developers must collect only necessary data for specific, explicit purposes, limiting the common practice of collecting vast datasets "just in case."
3. **Privacy by Design:** AI systems must incorporate data protection from the design phase, requiring techniques like federated learning, differential privacy, or data anonymization.
4. **Consent Requirements:** Training AI on personal data often requires explicit, informed consent, complicating data acquisition for machine learning.
5. **Data Subject Rights:** Individuals can access, correct, delete, or port their data, meaning AI systems must be able to handle data removal requests (the "right to be forgotten"), which challenges model retraining.
6. **Accountability and Documentation:** Developers must maintain records of data processing activities and conduct Data Protection Impact Assessments for high-risk AI applications.

2. Ethical Principles Matching

- **A) Justice** → 4. Fair distribution of AI benefits and risks
- **B) Non-maleficence** → 1. Ensuring AI does not harm individuals or society
- **C) Autonomy** → 2. Respecting users' right to control their data and decisions
- **D) Sustainability** → 3. Designing AI to be environmentally friendly

Part 2: Case Study Analysis

Case 1: Biased Hiring Tool

1. Identify the source of bias:

The primary source of bias in Amazon's AI recruiting tool was **biased training data**. The system was trained on resumes submitted to Amazon over a 10-year period, during which the tech industry (and Amazon specifically) was predominantly male, especially in technical roles. The model learned to penalize patterns associated with women, including:

- Resumes containing the word "women's" (e.g., "women's chess club captain")
- Graduates of all-women's colleges
- Language patterns more common in female-written resumes

Secondary factors include **model design choices** that failed to account for historical discrimination and **lack of fairness constraints** during optimization.

2. Three fixes to make the tool fairer:

Fix 1: Rebalance and Augment Training Data

- Collect or synthesize resumes to achieve gender balance across all job categories
- Include successful candidates from underrepresented groups, even if historically fewer were hired
- Remove gender-identifiable information (names, pronouns, gendered organizations) during training to prevent the model from learning gender associations
- Use techniques like SMOTE (Synthetic Minority Oversampling) to balance representation

Fix 2: Implement Fairness Constraints

- Add demographic parity or equalized odds constraints to the optimization objective
- Use adversarial debiasing techniques where a secondary model tries to predict gender from resume embeddings, and the primary model is penalized if gender is predictable
- Set threshold adjustments to ensure equal selection rates or equal false positive rates across genders

Fix 3: Human-in-the-Loop with Blind Review

- Use AI only as a screening aid, not the sole decision-maker
- Implement structured blind review processes where gender-identifiable information is removed
- Require human recruiters to review AI recommendations with diversity guidelines
- Conduct regular audits comparing AI recommendations against diverse human panels

3. Metrics to evaluate fairness post-correction:

Demographic Parity Metrics:

- **Selection Rate Ratio:** Ratio of candidates selected from each gender group (ideally close to 1.0)
- **Adverse Impact Ratio:** Percentage of women selected / percentage of men selected (should be ≥ 0.8 per the "four-fifths rule")

Equal Opportunity Metrics:

- **True Positive Rate (TPR) Parity:** Among qualified candidates, equal percentage recommended across genders
- **False Negative Rate (FNR) Parity:** Equal rate of qualified candidates being rejected across groups

Predictive Parity:

- **Positive Predictive Value (PPV):** Of those recommended, equal percentage actually succeed in role across genders
- **Calibration:** Confidence scores should correspond to actual success rates equally across groups

Intersectional Metrics:

- Evaluate fairness across intersections (e.g., women of color, women with disabilities)
- Use subgroup fairness analysis to identify hidden disparities

Implementation Monitoring:

- Track actual hiring outcomes, not just recommendations
- Monitor retention and promotion rates to validate quality of recommendations
- Regular A/B testing comparing AI-assisted vs. traditional hiring outcomes

Case 2: Facial Recognition in Policing

1. Discuss ethical risks:

Accuracy Disparities and Wrongful Arrests:

- Higher false positive rates for minorities can lead to wrongful arrests, investigations, and incarceration
- NIST studies show error rates up to 100 times higher for Asian and Black faces compared to white faces in some algorithms
- False positives can have devastating consequences: loss of employment, legal fees, trauma, and criminal records
- Compounding effect: misidentification → arrest → coerced confession or plea deals

Privacy Violations:

- Mass surveillance without consent as cameras capture faces in public spaces
- Chilling effect on freedom of assembly and protest when demonstrators fear identification
- Function creep: systems deployed for one purpose (finding violent criminals) expand to monitoring protests or tracking individuals
- Lack of data protection: biometric data, once captured, persists indefinitely

Due Process Violations:

- Lack of transparency in how systems work prevents meaningful challenge
- Defendants often unaware that facial recognition was used in their investigation
- Reliability questions: misidentification evidence may be inadmissible but still influences investigations

Discrimination and Targeting:

- Disproportionate deployment in minority communities amplifies existing over-policing
- Reinforces systemic racism by automating discriminatory practices
- Creates feedback loops: more surveillance → more arrests → more "crime" data → justifies more surveillance

Erosion of Trust:

- Communities lose trust in law enforcement when biased technology is deployed
- Particularly harmful in communities with historical trauma from policing
- Reduces cooperation with legitimate investigations

2. Recommend policies for responsible deployment:

Policy 1: Mandatory Accuracy Standards and Testing

- Require independent third-party testing showing <1% false positive rate across ALL demographic groups before deployment
- Continuous monitoring with public reporting of accuracy metrics disaggregated by race, gender, and age
- Immediate suspension if disparities exceed defined thresholds
- Require annual recertification as systems and datasets update

Policy 2: Strict Use Case Limitations

- Permit use ONLY for serious violent crimes (homicide, kidnapping, terrorism) with judicial oversight
- Prohibit use for: minor offenses, protest surveillance, immigration enforcement, identifying anonymous sources/whistleblowers
- Ban real-time facial recognition in public spaces
- Require that facial recognition can only be used as an investigative lead, never as sole probable cause for arrest

Policy 3: Transparency and Accountability Requirements

- Public disclosure when systems are deployed, including vendor contracts and accuracy audits
- Mandatory notification to defendants when facial recognition played any role in their case
- Establish independent oversight boards with community representation
- Create accessible complaint mechanisms with legal remedies for harm

Policy 4: Data Protection and Consent

- Strict limits on database sources (e.g., only booking photos, not driver's licenses or social media)
- Retention limits: delete non-match results immediately, matches after case resolution
- Prohibit commercial facial recognition databases
- Require explicit consent for inclusion in databases (except booking photos)

Policy 5: Community Input and Opt-Out Rights

- Require community approval via public referendum before deployment
- Allow municipalities to ban facial recognition in their jurisdictions
- Mandate impact assessments considering historical context of policing in affected communities
- Establish community advisory boards with power to suspend use

Policy 6: Officer Training and Documentation

- Training on technology limitations, bias, and proper use protocols
- Mandatory documentation of every facial recognition search with justification
- Disciplinary procedures for misuse
- Regular audits of search patterns to identify inappropriate use

Policy 7: Sunset Provisions and Regular Review

- Automatic expiration of authorization unless renewed with demonstrated benefits and no harm
- Regular reviews by civil liberties experts, technologists, and affected communities
- Requirement to consider less invasive alternatives