# PART 1 — THEORETICAL UNDERSTANDING

## 1. Short Answer Questions

**Q1. Define algorithmic bias and provide two examples of how it manifests in AI systems.**
Algorithmic bias is the systematic and unfair discrimination produced by an AI system due to skewed data, flawed model design, or biased assumptions embedded in the algorithm.

Examples:

i. Hiring systems penalizing certain genders or ethnic groups because historical training data reflects biased human decisions (e.g., Amazon penalizing CVs with “women’s” indicators).

ii. Facial recognition systems misidentifying people of color at higher rates due to under-representation in the training data or biased feature extraction.

**Q2. Explain the difference between transparency and explainability in AI. Why are both important?**
Transparency refers to how open and accessible information about an AI system is—its data sources, model architecture, training process, and decision pathways.

Explainability is the ability of the AI system to articulate why it made a particular decision, often through interpretable outputs or reasoning.

*Why both matter:*

Transparency builds trust, allows external scrutiny, and supports oversight.

Explainability helps users understand, challenge, or correct decisions, especially in high-stakes domains like hiring, lending, and policing.

Together, they strengthen accountability and reduce the risks of hidden discrimination or misuse.

**Q3. How does GDPR impact AI development in the EU?**
GDPR imposes strict rules on how personal data can be collected, stored, processed, and used, significantly shaping AI development.

Key impacts include:

i. Data minimization: AI systems must use only the data necessary for their function.

ii. Consent + lawful basis: Developers must obtain clear consent or establish legitimate grounds for data processing.

iii. Right to explanation: Users can request explanations for automated decisions, encouraging interpretable models.

iv. Right to be forgotten: AI systems must be designed so user data can be removed upon request.

v. Accountability & documentation: Developers must maintain detailed records and conduct Data Protection Impact Assessments (DPIAs) for high-risk AI.

Overall, GDPR pushes AI development toward privacy-preserving, transparent, and user-rights-focused systems.

**2. Ethical Principles Matching**

| Principle | Correct Definition |
|---|---|
| A) Justice | Fair distribution of AI benefits and risks. |
| B) Non-maleficence | Ensuring AI does not harm individuals or society. |
| C) Autonomy | Respecting users’ right to control their data and decisions. |
| D) Sustainability | Designing AI to be environmentally friendly. |

# Part 2: Case Study Analysis

## Case 1: Biased Hiring Tool

### Scenario: Amazon’s AI recruiting tool penalized female candidates.

1. Source of Bias:
   Training Data Bias: Historical hiring data favored male candidates, so the AI learned discriminatory patterns.
   Model Design Bias: Features correlated with gender (e.g., words in resumes) may have influenced decisions.

2. Proposed Fixes:
   Data Balancing: Ensure equal representation of genders; remove gender-related features.
   Bias Mitigation Algorithms: Apply techniques like reweighting, adversarial debiasing, or fairness-constrained learning.
   Human-in-the-Loop Review: Introduce human oversight for AI recommendations.

3. Metrics to Evaluate Fairness:

Statistical Parity Difference: Compare selection rates across genders.

Equal Opportunity / TPR Parity: Ensure equally qualified candidates have similar selection chances.

Disparate Impact Ratio: Ratio of positive outcomes for underrepresented vs. majority group (~1.0 is fair).

## Case 2: Facial Recognition in Policing

### Scenario: System misidentifies minorities at higher rates.

1. Ethical Risks:

Wrongful Arrests: Innocent individuals may be detained or prosecuted.

Discrimination & Inequality: Minorities face disproportionate impact.

Privacy Violations: Unauthorized use of facial data.

Erosion of Trust: Communities lose confidence in law enforcement and technology.

2. Policies for Responsible Deployment:

Accuracy Standards & Audits: Test on diverse datasets; audit performance regularly

Human Oversight: AI should assist, not replace, human decision-making.

Transparency &Accountability: Keep logs; allow complaints and review of AI decisions.

Consent & Legal Compliance: Use only in lawful contexts; comply with privacy laws.

Bias Mitigation: Retrain models with representative data to reduce demographic disparities.

# Part 3: Practical Audit

## COMPAS Notebook Report

Generated: 2025-12-04T13:08:26.892651Z

### 1) Dataset

Original dataframe (df) shape: (7214, 12)
Features used (X) shape: (7214, 11)
Training set: (5771, 11)
Testing set: (1443, 12)
Encoded protected attribute: race (encoding map available)

### 2) Preprocessing

Several irrelevant or duplicate columns were dropped.
Object/string race values were normalized and encoded into integers.
Label encoding applied to remaining categorical columns.

### 3) Model

Model: Logistic Regression
Test accuracy: 0.9757 (higher values indicate good overall classification accuracy)
Note: accuracy alone can hide subgroup performance disparities.

### 4) Fairness evaluation (AIF360)

Computed metrics (where available):

- Disparate Impact Ratio: 1.3781262057595847
- Mean Difference: 0.14600912895667129
- Equal Opportunity Difference: 0.0
- Theil Index: 0.0

Interpretation:

- Disparate Impact close to 1.0 suggests parity in favorable outcome rates; deviations indicate potential disparate treatment.
- Mean difference and equal opportunity difference near 0.0 indicate low group-level bias in these aspects.
- Theil index summarizes prediction distribution inequality.

### 5) Group-level performance (by race)

False Positive Rates (FPR) by race:

- Other (0): 0.0182
- African-American (1): 0.0501
- Caucasian (2): 0.0312
- Hispanic (3): 0.0361
- Native American (4): 0.5000
- Asian (5): 0.2500

True Positive Rates (TPR / Sensitivity) by race:

- Other (0): 1.0000
- African-American (1): 0.9973
- Caucasian (2): 1.0000
- Hispanic (3): 1.0000
- Native American (4): 1.0000
- Asian (5): 1.0000

Observation: - Even with high overall accuracy, FPR and TPR vary across race groups. This indicates disparate impact in error types across protected groups.

### 6) Key findings

i. The trained logistic regression achieves very high test accuracy (0.9757) on this split.
ii. There are measurable differences in false positive rates across races (see section 5), which may disadvantage certain groups despite overall accuracy.
iii. Some fairness metrics (disparate impact, mean difference) indicate mild bias in favorability; equal opportunity difference and Theil index may be low depending on the fold.

### 7) Recommendations

i. Report both overall metrics and per-group metrics (FPR, TPR, precision, recall) for transparency.
ii. Consider mitigation strategies:
    a. Preprocessing: reweighing (AIF360 Reweighing is already available).
    b. In-processing: use fairness-aware training algorithms (e.g., constrained optimization).
    c. Post-processing: calibrate or adjust decision thresholds per group ensuring legal/ethical compliance.
    d. Validate fairness on multiple random splits / cross-validation folds to ensure stability.
    e. If deploying, gather stakeholder input and perform impact assessments for affected groups.

### 8) Next steps (code pointers)

i. Apply AIF360 Reweighing on the training dataset and retrain the classifier.
ii. Evaluate fairness-aware algorithms available in AIF360 /Fairlearn.
iii. Produce per-group confusion matrices and visualize distributions (ROC, calibration).

This is a summary of the preprocessing, modeling, and fairness analysis performed in the notebook. For reproducibility, rerun the notebook cells in order and then regenerate this report.

## Part 4: Ethical Reflection

Project: Predicting student performance using AI.

Ethical Considerations &Actions:

1. Fairness:

    Ensure the dataset represents students from all backgrounds equally to prevent bias.

    Regularly audit the model's predictions to identify and correct disparities.

2. Transparency:

    Make the AI decision-making process understandable to educators and students.

    Provide clear explanations for predictions, e.g., why a student is flagged as at-risk.

3. Privacy & Consent:

    Collect only necessary data and anonymize sensitive student information.

    Obtain explicit consent from students or guardians before data collection.

4. Accountability:

    Assign a responsible team member to monitor model outputs and address errors.

    Maintain logs of decisions and interventions for auditing.

Summary:

By combining fairness, transparency, privacy, and accountability, the project will promote responsible AI use that benefits students without causing harm or discrimination.

# Ethical AI Use in Healthcare

## Ethical AI Guidelines for Healthcare Applications

1. Patient Consent Protocols:

Obtain explicit, informed consent before collecting or using patient data.

Explain clearly how data will be used, including AI predictions or analyses.

Allow patients to withdraw consent at any time without penalty.

2. Bias Mitigation Strategies:

Ensure datasets include diverse patient populations across age, ethnicity, gender and medical history.

Regularly audit AI outputs for disparities in treatment recommendations.

Use bias detection and correction tools to reduce systemic errors.

3. Transparency Requirements:

Provide clear explanations for AI-driven recommendations to patients and clinicians.

Document model design, training data, and limitations for accountability.

Make reporting channels available for clinicians and patients to flag unexpected or harmful AI outputs.

4. Accountability & Oversight:

Assign responsible teams to monitor AI performance and intervene when necessary.

Maintain audit logs for all AI decisions that impact patient care.

Goal:

Ensure AI tools enhance healthcare delivery safely, fairly, and responsibly, while protecting patient rights and promoting trust.