PART 1: THEORETICAL UNDERSTANDING

Case 1: Algorithmic Bias

Algorithmic bias refers to systematic errors in AI systems that produce unfair or discriminatory outcomes, often reflecting existing social, racial, or gender inequalities.

Examples:

- **Hiring tools:** Amazon's recruiting AI penalized female candidates because training data reflected male-dominated hiring patterns.
- **Facial recognition:** Systems misidentify minorities at higher rates due to underrepresentation in training datasets.

Transparency vs. explainability

Transparency: Refers to openness about how an AI system is built, trained and deployed, including disclosure of datasets, algorithms and decision processes.

Explainability: Refers to the ability to interpret and understand why an AI system made a specific decision, often using tools like SHAP or LIME.

Importance:

- Transparency builds accountability and trust by making systems auditable.
- Explainability ensures users and regulators can challenge or validate outcomes, preventing "black box" risks. Together, they are essential for trustworthy AI, ensuring systems are both open and understandable.

GDPR's impact on AI development in the EU

The General Data Protection Regulation (GDPR) profoundly shapes AI development by prioritizing data protection and individual rights.

Key impacts:

- Requires a lawful basis for AI systems to process personal data.
- Grants individuals the right to explanation and the ability to opt out of automated decision-making.
- Enforces privacy by design and data minimization, limiting excessive data collection.
- Creates compliance burdens that can slow innovation, but also fosters ethical, human-centric AI.

Case 2: Ethical Principles Matching

Principle → 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 (40%)

Case 1: Biased Hiring Tool

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

Source of bias

- The primary source of bias was training data.
- The system was trained on historical resumes from Amazon, which reflected a male-dominated workforce in tech roles. As a result, the model learned to favor male-associated terms and penalize resumes mentioning women’s colleges or female-related activities. This bias was not due to malicious design but rather the inheritance of systemic inequalities embedded in the dataset.

Proposed fixes

(i) Balanced training data: Curate datasets with equal representation of male and female candidates, ensuring diverse educational and career backgrounds.

(ii) Feature engineering & debiasing: Remove or neutralize biased features (penalizing terms like “women’s” in resumes). Apply fairness-aware preprocessing techniques such as reweighing or sampling adjustments.

(iii)Fairness-aware algorithms: Implement in-processing methods like adversarial debiasing, which penalize discriminatory patterns during model training. Combine with post-processing calibration to equalize outcomes across genders.

Fairness Metrics (Post-Correction)

To evaluate fairness after corrections, the following metrics should be applied:

- Demographic parity: Ensures selection rates are equal across genders.
- Equal opportunity: Confirms equal true positive rates (qualified candidates selected) across groups.
- Disparate impact ratio: Measures whether one group is disproportionately disadvantaged compared to another.
- Calibration across groups: Ensures predicted probabilities of success are equally reliable for male and female candidates.

Case 2: Facial Recognition in Policing

Scenario: A facial recognition system misidentifies minorities at higher rates.

Ethical Risks

(i) Wrongful arrests:

Misidentification can result in innocent individuals, particularly from minority groups, being wrongly accused or detained. This undermines justice and perpetuates systemic discrimination.

(ii) Privacy violations:

Facial recognition often involves mass surveillance, capturing biometric data without consent. This erodes individual privacy and raises concerns about constant monitoring in public spaces.

(iii)Erosion of trust:

Communities disproportionately targeted by misidentifications may lose trust in law enforcement and public institutions, deepening social divides.

(iv)Disproportionate harm:

Marginalized groups bear the brunt of errors, reinforcing existing inequalities and exposing them to higher risks of legal and social consequences.

Policies for responsible deployment

(i) Independent audits:

Require third-party audits of facial recognition systems to assess accuracy across demographic groups before deployment.

(ii) Human oversight:

Ensure that AI outputs are advisory only. Final decisions (e.g., arrests) must involve human review and corroborating evidence.

(iii)Transparency reports:

Mandate public disclosure of system performance, error rates, and demographic disparities to maintain accountability.

(iv)Restricted use cases:

(v) Limit deployment to high-stakes scenarios where benefits outweigh risks (serious crimes), and prohibit use in routine surveillance or low-level policing.

(vi)Community consultation:

Engage affected communities in decision-making processes to ensure policies reflect public concerns and ethical standards.

PART 4: ETHICAL REFLECTION

Ensuring ethical AI principles in personal projects

Personal Project: COMPAS Bias Audit

For this reflection, we will consider the COMPAS bias audit project we recently completed. This project involved analyzing racial bias in the COMPAS recidivism risk assessment tool using IBM's AI Fairness 360 toolkit. The analysis revealed that African-American defendants were incorrectly flagged as high-risk at nearly double the rate of Caucasian defendants (47.1% vs 24.2% false positive rate), despite similar overall accuracy.

Reflection: What we learned

- Bias is invisible until measured: The COMPAS system appeared objective but contained significant racial bias that only became apparent through systematic analysis using fairness metrics.
- Technical accuracy is not equal to fairness: The system had reasonable overall accuracy but failed dramatically on fairness metrics, particularly false positive rates.
- Context matters critically: In criminal justice, false positives have severe consequences, incorrectly labeling someone as high-risk can lead to longer sentences and reduced opportunities.
- Proactive measurement is essential: Simply removing protected attributes (like race) is insufficient; bias can persist through proxy variables and must be actively measured and mitigated.
- Transparency enables accountability: By making bias visible through metrics and visualizations, we enable stakeholders to demand change and hold systems accountable.

Ensuring ethical AI in future projects

- Fairness and non-discrimination: Training data for representativeness will be audited, apply fairness metrics from the start, and require bias audits before deployment. Continuous monitoring dashboards will track performance by demographic group, with automated alerts for bias drift. Example: the COMPAS audit revealed a 94.5% disparity in false positive rates across races, underscoring the need for group-specific evaluation.
- Transparency and explainability: Involves creating model cards documenting purpose, data, and limitations, and implement explainability tools (e.g., SHAP, feature importance). Reports and visualizations will be written in accessible language to ensure stakeholders understand system behavior.
- Accountability and responsibility: Clear ownership will be established for system outcomes, with ethical review checkpoints at each phase. Audit trails and user feedback mechanisms will ensure issues are logged and addressed, including retraining or retiring biased models.
- Privacy and data protection: Involves minimizing data collection, apply anonymization or differential privacy, and comply with regulations such as GDPR and CCPA. Users will retain control over their data, with rights to view, correct, or delete information.

- Robustness and safety: Models will be tested against diverse scenarios and adversarial conditions, with confidence scores guiding safe fallback to human judgment when uncertainty is high.
- Human agency and oversight: AI will augment, not replace, human decision-making. Mandatory human review will be required for high-stakes outcomes, with override mechanisms and escalation processes in place.

BONUS TASK

Policy Proposal: Ethical AI Use in Healthcare

Artificial Intelligence (AI) has the potential to transform healthcare by improving diagnostics, treatment planning, and patient outcomes. However, its deployment must adhere to strict ethical standards to protect patients, ensure fairness, and maintain trust. This guideline outlines protocols for patient consent, bias mitigation, and transparency requirements.

Patient consent protocols

- Informed Consent: Patients must be clearly informed when AI systems are used in their care, including the purpose, risks, and limitations.
- Opt-In/Opt-Out Options: Patients should have the right to decline AI-driven decisions or request human review.
- Data Privacy: All patient data must be collected and processed in compliance with GDPR and HIPAA, ensuring confidentiality and secure storage.
- Continuous Consent: Consent should be revisited if AI systems change significantly (e.g., new algorithms or expanded data use).

Bias mitigation strategies

- Diverse training data: AI models must be trained on datasets that represent varied demographics (age, gender, ethnicity, socioeconomic status).
- Fairness audits: Regular audits using tools like AI Fairness 360 should assess disparities in predictions across patient groups.
- Algorithmic adjustments: Apply preprocessing (e.g., reweighing), in-processing (fairness-aware models), and post-processing (calibration) to reduce bias.
- Stakeholder oversight: Involve clinicians, ethicists, and patient advocates in reviewing AI systems for fairness and equity.

Transparency requirements

- Explainability: AI systems must provide interpretable outputs, enabling clinicians and patients to understand the reasoning behind recommendations.
- Documentation: Developers must disclose datasets, model architectures, and evaluation metrics used in system design.
- Performance reporting: Publish regular transparency reports detailing accuracy, error rates, and demographic disparities.
- Human oversight: AI should support, not replace, clinical judgment. Final decisions must remain with qualified healthcare professionals.