Part 1: Theoretical Understanding

Q1: Define algorithmic bias and provide two examples of how it manifests in AI systems.

Algorithmic bias refers to systematic and unfair errors in AI systems that lead to disadvantaged or discriminatory outcomes for certain individuals or groups. These biases often arise from imbalanced data, flawed model design, or structural inequalities embedded in the training process.

Example 1 — Facial Recognition Bias:
Facial recognition models trained predominantly on lighter-skinned faces may misidentify individuals with darker skin at a significantly higher rate, leading to unequal accuracy across demographic groups.

Example 2 — Loan or Credit Scoring Bias:
AI models used in banking may learn patterns in historical data that disadvantage lower-income or minority groups, resulting in unfairly higher rejection rates or higher interest offers even when creditworthiness is similar.

Q2: Explain the difference between transparency and explainability in AI. Why are both important?

Transparency refers to how openly an AI system reveals information about its design, data sources, training process, and decision pipeline. It answers: *What does the system contain, and how was it built?*

Explainability, on the other hand, focuses on making the model's **specific predictions** understandable. It answers: *Why did the system make this particular decision?* Explainability tools include methods like SHAP, LIME, and feature importance.

Why both matter:

- Transparency ensures accountability, regulatory compliance, and trust from stakeholders by revealing how an AI system was created.
- Explainability helps users and impacted individuals understand decisions, detect errors, and challenge unfair or harmful outcomes.
  Together, they reduce risk, improve reliability, and support responsible adoption of AI systems.

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

GDPR imposes strict rules on how personal data can be collected, processed, and stored, significantly influencing AI development. Key impacts include:

- **Data minimization:** AI developers must limit data collection to what is strictly necessary, reducing excessive or irrelevant data usage.

- **Lawful basis & consent:** AI systems must have a clear legal basis for processing personal data, often requiring explicit user consent.
- **Right to explanation:** Individuals have the right to understand automated decisions that significantly affect them, pushing AI creators to incorporate explainability features.
- **Right to erasure ("right to be forgotten"):** AI systems must be designed to delete personal data upon request, impacting how training data is stored and updated.
- **Accountability and security:** Developers must implement strong data protection mechanisms and document compliance processes.

GDPR therefore forces AI systems to be more transparent, privacy-preserving, and user-centric.

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 — Amazon Recruiting System

1. Source of Bias

The primary source of bias was the **training data**. Amazon trained its hiring model on historical résumés submitted over a 10-year period—data that reflected a male-dominated tech workforce. Because the dataset contained more successful male applicants, the model learned to favor male-associated patterns (e.g., verbs or keywords more common in men's resumes) and penalize terms linked to women (e.g., "women's chess club").
Additional issues include **feature design choices** (e.g., using keywords correlated with gender) and **lack of fairness constraints** during model training.

2. Three Fixes to Make the Tool Fairer

1. **Debias the training data**
   Remove gender-correlated features, re-balance the dataset, or apply preprocessing techniques such as *reweighing* to ensure equal representation of genders.
2. **Use fairness-aware algorithms**
   Implement in-processing methods like adversarial debiasing or include fairness constraints (e.g., equal opportunity) during model optimization.
3. **Human-in-the-loop review + continuous audits**
   Ensure model recommendations are reviewed by trained HR staff, and conduct periodic fairness checks to monitor drift or emerging biases in decisions.

3. Fairness Metrics to Evaluate After Fixing

- **Disparate Impact Ratio** (selection rate of women vs. men)
- **Equal Opportunity Difference** (TPR difference between genders)
- **Demographic Parity Difference**
- **False Positive/Negative Rate gaps** between gender groups

Case 2: Facial Recognition in Policing

1. Ethical Risks

- **Wrongful arrests and false accusations**
  Higher misidentification rates for minorities can lead to unjust detentions, criminal records, or police violence.
- **Privacy violations**
  Continuous surveillance erodes citizens' rights to anonymity, especially if facial recognition is deployed without consent.
- **Discrimination and disproportionate targeting**
  Communities of color may face over-surveillance, reinforcing existing inequalities in policing.
- **Lack of transparency and accountability**
  Proprietary algorithms make it hard to contest errors or understand how decisions are made.

2. Policies for Responsible Deployment

1. **Independent accuracy and bias audits**
   Require third-party testing of model performance across demographic groups before deployment.
2. **Strict regulations and limited use cases**
   Only allow facial recognition for serious crimes with judicial oversight—not for routine patrol or public crowd scanning.
3. **Human verification requirement**
   No match should lead directly to arrest; officers must verify identities using independent evidence.
4. **Data governance and privacy protections**
   Limit data retention, log all searches, and ban the use of images from social media without consent.
5. **Public transparency and community consultation**
   Police departments must disclose system performance, procurement policies, and allow public input before adoption.

PART 3: Practical Audit

Dataset: COMPAS Recidivism Dataset Goal: Detect racial bias using AI Fairness 360.

Summary: The COMPAS dataset often shows higher false positive rates for Black defendants, meaning

they are predicted to re offend more often than they actually do. After analyzing the dataset using

fairness metrics such as disparate impact, equal opportunity, and false positive rate disparity, results

typically show measurable racial bias. Remediation steps include reweighing, adversarial debiasing,

and post-processing adjustments like equalized odds optimization. Visualization through FPR disparity.

Part 4: Ethical Reflection

First, **fairness and bias mitigation** will be central. The model will be trained on datasets that represent the full diversity of machine screen states to avoid unintended bias toward certain lighting conditions, motion patterns, or device types. I will continuously test the system for performance disparities across different video sources or machine models, ensuring no group or scenario is unfairly disadvantaged.

Second, I will enforce strong **privacy and data protection** standards. All video footage used will be handled securely, anonymized where possible, and stored only for the minimum duration necessary. Any sensitive or personally identifying information will be excluded or blurred. I will follow local data protection laws and implement access control for all datasets and models.

Third, I will prioritize **transparency and explainability**. Even though this is a technical project, I will document the model’s decision logic, feature extraction pipeline, and evaluation metrics. For stakeholders—especially non-technical users—I will provide simplified explanations that clarify how the model decides whether a frame is moving or still.

Fourth, I will apply **accountability mechanisms**. This includes maintaining logs of model predictions, versioning model updates, and creating mechanisms to report errors or unintended outputs. If the system fails or behaves unexpectedly, I will ensure that clear responsibility pathways exist for diagnosing and correcting the issue.

Lastly, I will consider the broader **social impact** of the system. Even a technical solution can have downstream effects—for example, influencing financial auditing processes or regulatory decisions. Ensuring the model is accurate, fair, and used appropriately will be essential to preventing harm.

In one of my recent projects—building a machine learning model to classify images captured from gambling machines—I learned the importance of embedding ethical principles throughout the development life-cycle. As I continue the project and future AI work, I will implement several safeguards to ensure the system aligns with ethical AI principles.

Authors:

Lynn Bitok
Adoh Baraza
Salome Wamaitha
Purity Mutunga
Meshack Omuda
Jusper Ageri