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 discrimination in AI outcomes caused by flawed data, design, or decision-making processes. It occurs when AI models produce results that disadvantage certain individuals or groups.

- **Example 1:** A hiring algorithm trained on past employee data favors male candidates over female candidates, replicating historical gender imbalances.
- **Example 2:** A facial recognition system misidentifies people of colour more frequently than white individuals, leading to disproportionate risks of wrongful arrests.

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

- **Transparency** means making the inner workings of an AI system visible and understandable to stakeholders (e.g., what data is used, what model is applied).
- **Explainability** refers to the ability to provide clear, interpretable reasons for specific AI outputs or decisions.

Importance:

- Transparency builds trust and accountability by revealing how decisions are made.
- Explainability ensures users and regulators can understand, challenge, or contest decisions—especially in high-stakes contexts like healthcare, finance, and justice.

Q3: How does GDPR (General Data Protection Regulation) impact AI development in the EU?
GDPR affects AI development by enforcing **data protection, fairness, and accountability**. Key impacts include:

- **Right to explanation:** Users can request explanations of AI-driven decisions.
- **Data minimization & consent:** AI developers must collect only necessary data with explicit consent.
- **Accountability:** Organizations must demonstrate lawful processing and bias mitigation.
This ensures AI systems respect privacy, fairness, and human rights.

Ethical Principles Matching

- **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.

Case 1: Biased Hiring Tool (Amazon’s AI recruiting tool)

Source of Bias:

The primary source of bias in Amazon's AI recruiting tool was the **training data**. The tool was trained on historical resume data, which predominantly came from male applicants in the tech industry. This historical data reflected existing gender disparities in the industry, and the AI learned to associate certain terms and patterns common in male applicants' resumes (e.g., participation in male-dominated sports, use of specific verbs) with success, while penalizing terms and patterns more common in female applicants' resumes (e.g., attendance at women's colleges, membership in women's organizations). The **model design** itself, if not specifically built with fairness in mind, would perpetuate this bias learned from the data. The algorithm essentially mirrored and amplified the historical biases present in the hiring process.

Proposed Fixes to Make the Tool Fairer:

1. **Data Augmentation and Balancing:** Instead of solely relying on historical data, augment the training data with a more balanced representation of qualified candidates from underrepresented groups. This could involve creating synthetic data (with caution to avoid introducing new biases) or actively seeking out and incorporating resumes from successful female candidates in various tech roles. Techniques like oversampling minority classes or under sampling majority classes can also help balance the dataset.
2. **Bias Mitigation Techniques in Model Training:** Employ algorithmic bias mitigation techniques during the model training phase. These techniques can be applied pre-processing (adjusting the data), in-processing (modifying the training algorithm), or post-processing (adjusting the model's output). Examples include using adversarial debiasing, reweighing data points, or using methods that enforce equalized odds or demographic parity.
3. **Feature Selection and Engineering with Fairness in Mind:** Carefully review and potentially remove or transform features that are highly correlated with protected attributes like gender (e.g., specific names, participation in gender-specific activities) but are not truly indicative of job performance. Focus on skills, experience, and qualifications that are directly relevant to the job requirements, regardless of how they are presented in the resume.

Metrics to Evaluate Fairness Post-Correction:

- **Demographic Parity (or Disparate Impact):** Measure if the selection rate for female candidates is statistically like the selection rate for male candidates. A common rule of thumb is the "80% rule," where the selection rate for the disadvantaged group should be at least 80% of the selection rate for the advantaged group.
- **Equalized Odds:** This metric assesses whether the model has similar false positive rates and false negative rates across different demographic groups. In a hiring context, it would mean the tool is equally likely to incorrectly reject a qualified female candidate as a qualified male candidate, and equally likely to incorrectly flag an unqualified female candidate as a qualified male candidate.
- **Predictive Parity:** This metric checks if the predicted outcome (e.g., likelihood of success) is the same for individuals from different groups who have the same true outcome. In hiring,

this would mean that among truly successful candidates, the model assigns a similar high probability of success regardless of gender.

Case 2: Facial Recognition in Policing

Ethical Risks:

The use of facial recognition systems in policing, especially when they exhibit differential accuracy across demographic groups, poses significant ethical risks:

- **Wrongful Arrests and Incarceration:** Higher misidentification rates for minorities can lead to innocent individuals from these groups being wrongly accused, arrested, and even incarcerated, with devastating consequences for their lives and communities.
- **Exacerbation of Existing Biases:** Such systems can reinforce and amplify existing racial biases within the criminal justice system, leading to disproportionate surveillance and scrutiny of minority communities.
- **Chilling Effect on Civil Liberties:** The pervasive use of facial recognition can create a chilling effect on freedom of assembly and expression, as individuals may be hesitant to participate in protests or public gatherings if they fear being tracked and identified.
- **Privacy Violations:** Constant surveillance through facial recognition infringes on individuals' right to privacy, allowing for the potential tracking of movements and associations without explicit consent or probable cause.
- **Lack of Transparency and Accountability:** Often, the algorithms used in these systems are proprietary, making it difficult to understand how they work, identify sources of error, and hold those responsible accountable for biased outcomes.

Recommended Policies for Responsible Deployment:

- **Accuracy Standards and Independent Testing:** Mandate rigorous independent testing of facial recognition systems to ensure they meet high accuracy standards and demonstrate equal performance across different demographic groups before deployment. Systems that fail to meet these standards should not be used.
- **Clear and Limited Use Cases:** Define specific and narrow use cases for facial recognition in policing, focusing on serious crimes and requiring a reasonable suspicion threshold before activation. Prohibit its use for mass surveillance or in low-level offenses.
- **Human Oversight and Due Process:** Require meaningful human review and verification of any potential matches generated by the facial recognition system. The AI's output should never be the sole basis for an arrest or other significant action. Due process rights must be protected.
- **Transparency and Public Disclosure:** Law enforcement agencies should be transparent about their use of facial recognition technology, including the types of systems used, their capabilities, and how the data is collected, stored, and used. Public reporting on accuracy rates and demographic disparities should be mandatory.
- **Legal and Regulatory Framework:** Establish clear legal and regulatory frameworks governing the use of facial recognition in policing, including guidelines on data retention, access, and the rights of individuals who are subject to surveillance.

COMPAS Dataset Fairness Audit – Report

Summary of Findings

The COMPAS recidivism dataset was analysed to examine potential racial and demographic disparities in algorithmic risk scoring. After filtering, the dataset included 6,172 individuals. Demographic analysis showed that African - American defendants represented the largest group (51.44%), followed by Caucasian defendants (34.07%). Males made up a significant majority (80.96%), and the overall two-year recidivism rate was approximately 45.51%.

Exploring decile risk scores revealed clear disparities: African - American defendants were more frequently assigned higher scores (7–10) compared to Caucasian defendants. A logistic regression model predicting "High Score" demonstrated statistically significant predictors. Race remained influential even after controlling for age, priors, and charge degree. The odds ratio indicated that African - American defendants had ~45% higher odds of receiving a "High Score" than Caucasian defendants. Additionally, individuals under 25 were found to have ~2.5 times greater odds of being classified high risk, showing age as another strong factor.

These findings confirm the presence of algorithmic bias in COMPAS risk assessment, with African - American defendants disproportionately flagged as higher risk. Such disparities raise serious ethical concerns regarding fairness, justice, and trust in algorithmic decision-making within the criminal justice system.

Remediation Steps

1. **Fairness-Aware Machine Learning**: Integrate techniques such as reweighing or adversarial debiasing to minimize disparate impact while maintaining predictive performance.
2. **Ongoing Algorithmic Audits**: Establish regular audits using fairness metrics (e.g., disparate impact ratio, equal opportunity difference) to track bias over time.
3. **Feature Scrutiny**: Reassess input features to ensure they are not proxies for race or other protected attributes.
4. **Transparency & Explainability**: Document scoring logic and provide interpretable outputs so stakeholders can understand how risk scores are generated.
5. **Human Oversight**: Use COMPAS as a supplementary tool, not as the sole determinant of judicial decisions, ensuring contextual review in each case.

Conclusion

This audit highlights how bias can persist in widely used AI systems. Addressing these disparities requires a combination of **technical interventions, policy safeguards, and ethical governance** to ensure that risk assessment tools are fair, accountable, and trustworthy.

Ethical Reflection on My AI Projects

In my past and ongoing projects, such as developing predictive models in health and education, I recognize the importance of embedding ethical AI principles at every stage. My main goal is to ensure fairness, transparency, and accountability while avoiding harm. For example, in a predictive health model I worked on, I would ensure that sensitive attributes such as race, gender, or socioeconomic status are either excluded or carefully audited to prevent indirect bias. To achieve this, I will adopt practices such as bias detection using fairness metrics, continuous dataset auditing, and applying fairness-aware algorithms when disparities are identified. Transparency will be promoted by documenting all preprocessing and modeling decisions, and explainability tools will be used to help stakeholders understand predictions. Importantly, I will prioritize human oversight so that AI recommendations are treated as supportive insights rather than absolute decisions. By following principles of justice, non-maleficence, autonomy, and sustainability, I aim to ensure that my AI systems not only perform well technically but also uphold the values of fairness and human dignity.

Policy Guidelines for Responsible AI Use in Healthcare

Introduction

Artificial Intelligence holds great promise for healthcare, from diagnosis support to resource allocation. However, ethical safeguards are essential to ensure patient safety and fairness. The following guidelines are proposed:

1. Patient Consent Protocols

- Ensure explicit informed consent before using patient data for AI training or deployment.
- Provide patients with the right to opt out without compromising their care.

2. Bias Mitigation Strategies

- Audit datasets regularly for representation imbalances across race, gender, and socioeconomic groups.
- Apply fairness-aware algorithms (e.g., reweighing, adversarial debiasing) to reduce disparities in outcomes.

3. Transparency Requirements

- Mandate explainable AI methods that allow clinicians and patients to understand how predictions are made.
- Publish documentation of model training, limitations, and intended use cases.

4. Human Oversight

- Ensure AI systems act as decision-support tools, not autonomous decision-makers.
- Clinicians must retain final authority in patient care decisions.

5. Sustainability & Accountability

- Evaluate the environmental impact of large-scale healthcare AI systems.
- Establish accountability frameworks where developers and institutions share responsibility for errors or harm.

Conclusion

By integrating consent, fairness, transparency, oversight, and sustainability, these guidelines will promote trust and responsible adoption of AI in healthcare, ensuring technology enhances patient well-being without compromising ethical standards.