Part 1: Theoretical Understanding (30%)

1. Short Answer Questions

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 systems that arise from the design, data, or assumptions made during the development of the system. It occurs when an AI system produces results that are prejudiced against certain individuals or groups, often due to historical inequalities, non-representative data, or flawed models.

Examples:

1. Facial recognition bias: AI systems may perform poorly on certain demographic groups, particularly darker-skinned individuals, due to underrepresentation in training datasets. This bias can lead to misidentification or misclassification, disproportionately affecting minority groups.
2. Risk assessment tools (e.g., COMPAS): In criminal justice, AI tools like COMPAS have shown bias in predicting recidivism risk. These tools may overestimate the likelihood of re-offending for Black defendants and underestimate it for white defendants, due to biased data used to train the system.

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

Transparency refers to the openness and accessibility of an AI system's inner workings, including the data, algorithms, and decisions made by the system. It means providing clear insights into how the system operates and what data it uses.

Explainability refers to the ability to explain the reasons behind the decisions made by an AI model in a manner understandable to humans. It answers the "why" of a particular output or decision, allowing users to comprehend the rationale behind it.

Importance:

- Transparency helps build trust with users by showing them how the system is functioning and allows for external audits.
- Explainability ensures accountability by providing clear explanations for AI decisions, making it easier to identify errors or biases in the system and enabling users to make informed choices.

Q3: How does GDPR (General Data Protection Regulation) impact AI development in the EU?

The GDPR is a regulation in the European Union that focuses on data protection and privacy. It directly impacts AI development in the EU in several ways:

- Data Consent: AI systems must obtain explicit consent from users before processing their personal data, which affects how AI models gather and use data for training.

- Right to Explanation: Under GDPR, individuals have the right to an explanation about decisions made by automated systems, which directly impacts AI models that make decisions like credit scoring or hiring.
- Data Minimization: AI systems should only collect and use the data necessary for the specific task, limiting excessive data gathering and reducing the potential for misuse or bias.
- Accountability and Transparency: Organizations are required to demonstrate how their AI systems are compliant with GDPR, ensuring that AI models are transparent and auditable.

2. 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.

Part 2: Case Study Analysis

Case 1: Biased Hiring Tool

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

Source of Bias:
The bias in Amazon's AI recruiting tool stems from **training data** and the **model design**:

- **Training Data**: The tool was trained on resumes submitted to Amazon over a 10-year period, and these historical resumes reflected gender imbalances in the tech industry. As a result, the model learned patterns that favored male candidates, who were historically overrepresented in the hiring process.
- **Model Design**: The algorithm did not account for gender or diversity, inadvertently reinforcing biases from the training data. It used factors such as keywords and past hiring patterns that were skewed against women in tech, penalizing resumes that included terms often associated with female candidates (e.g., "women's" or "female").

Proposed Fixes:

1. **Bias-aware Training**:
   Retrain the algorithm with a diverse and balanced dataset that includes a fair representation of both male and female candidates. The data should be adjusted to ensure that it does not reflect historical biases, such as gender imbalances in specific roles or industries.
2. **Bias Mitigation Algorithms**:
   Use fairness-enhancing algorithms such as reweighing or adversarial debiasing to minimize the impact of gender-related patterns in the data. These algorithms can help neutralize the bias by adjusting the training process to ensure fairness for all demographic groups.
3. **Bias Audits and Transparency**:
   Regularly conduct bias audits and make the results of these audits public. This will allow Amazon to track the effectiveness of the fairness corrections, demonstrate transparency, and build trust in the AI system. This can be paired with explainable AI tools to ensure that decisions made by the system are understandable.

Metrics to Evaluate Fairness Post-Correction:

1. **Disparate Impact Ratio**:
   Measure the disparate impact ratio between male and female candidates to ensure that the tool does not disproportionately penalize female candidates in the recruitment process.
2. **Equal Opportunity Difference**:
   Evaluate the equal opportunity difference, particularly focusing on whether the AI system treats male and female candidates with similar qualifications equally, ensuring no systemic disadvantage for women.

3. **Representation of Gender**:
   Measure the representation of male and female candidates in the final pool of hires and compare it with the overall applicant pool. The hiring tool should result in an equitable distribution of genders across all stages of recruitment.

Case 2: Facial Recognition in Policing

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

Ethical Risks:

1. **Wrongful Arrests**:
   The higher misidentification rate for minorities increases the risk of wrongful arrests. Individuals could be wrongly flagged as suspects, leading to detainment, legal challenges, or reputational harm, especially for those in marginalized communities who are already disproportionately affected by the criminal justice system.
2. **Privacy Violations**:
   The use of facial recognition technology raises significant concerns about privacy violations. Individuals may be surveilled without consent, and sensitive data could be collected and misused. The widespread deployment of such systems without proper regulation may lead to constant monitoring, infringing on individuals' right to privacy.
3. **Discrimination and Systemic Bias**:
   The systemic bias in facial recognition systems, which often misidentify minorities and women more than white males, can perpetuate social inequities. This could disproportionately affect people of color, leading to unjust outcomes and reinforcing harmful stereotypes.
4. **Lack of Accountability**:
   The reliance on automated systems in policing can reduce accountability, as it may be difficult to pinpoint responsibility when misidentifications occur. This could erode public trust in law enforcement agencies and their ability to protect citizens fairly.

Recommended Policies for Responsible Deployment:

1. **Transparency and Explainability**:
   Law enforcement agencies must adopt policies that promote transparency in how facial recognition systems are used. This includes informing the public about the use of these technologies, the data being collected, and how the system works. The deployment of these systems should be accompanied by explainable AI to ensure that decisions made based on facial recognition can be understood and justified.
2. **Strict Accuracy Standards and Testing**:
   Before deploying facial recognition technology, agencies should ensure the system meets strict accuracy standards. This includes testing the system for bias and ensuring it performs equally well across all demographic groups. If a system shows any disproportionate rates of misidentification for minority groups, it should be discontinued or improved before further use. This would include regular bias audits and independent testing to monitor performance across different racial, gender, and age groups.

3. **Consent and Regulation**:
   Establish strict consent policies that limit the use of facial recognition technology. Individuals should be informed and give consent before being monitored by facial recognition systems. Additionally, policies should ensure that the use of such technology is regulated by government agencies to prevent abuse and ensure compliance with privacy rights.
4. **Human Oversight and Accountability**:
   Require human oversight for all decisions made by facial recognition systems, especially when it comes to arrests or surveillance. There must be a mechanism for accountability to ensure that errors made by the system do not lead to unfair treatment of individuals. This could include a process for individuals to challenge misidentifications and have their cases reviewed by a human operator.
5. **Public and Stakeholder Engagement**:
   Engage with the public, civil rights groups, and technology experts to discuss the ethical concerns around facial recognition and come up with policies that protect vulnerable communities. This includes fostering public dialogue on the balance between security and privacy, and the potential risks of mass surveillance.

By addressing these ethical risks and implementing robust policies, law enforcement agencies can mitigate the harmful consequences of biased facial recognition systems, ensuring that these technologies are deployed in a responsible, transparent, and fair manner.