Part 1: Theoretical Understanding

Q1: Define algorithmic bias and provide two examples

Algorithmic Bias Definition: Algorithmic bias happens when an AI system produces unfair or discriminatory results that favor certain groups over others. This occurs when the AI learns patterns from biased data or is designed in ways that systematically disadvantage specific populations.

Examples:

Example 1: Credit Scoring Systems

- AI credit systems sometimes give lower credit scores to people from certain zip codes
- This happens because the AI learns that people from low-income neighborhoods have historically defaulted more often
- The problem is that this punishes individuals based on where they live, not their actual ability to pay back loans
- It creates a cycle where people in these areas can't get loans to improve their situation

Example 2: Resume Screening Software

- Some AI hiring tools automatically reject resumes with names that sound African American
- The AI learned this bias from historical hiring data where people with these names were less likely to be hired
- This perpetuates workplace discrimination even when human recruiters might be trying to be fair
- Qualified candidates get rejected before a human even sees their resume

Q2: Explain the difference between transparency and explainability in AI

Transparency in AI:

- Transparency means being open about how an AI system works overall. It's about sharing information like what data was used, who built it, and what it's designed to do.
- Example: A bank tells customers they use AI for loan decisions and explains they consider income, credit history, and employment

Explainability in AI:

- Explainability means the AI can explain why it made a specific decision. It's about understanding the reasoning behind individual outcomes
- Example: The AI tells a loan applicant "You were rejected because your debt-to-income ratio is 60%, which is above our 40% threshold"

Why Both Are Important:

Transparency builds trust as people need to know AI systems exist and how they generally work to trust them. Without transparency, people feel like decisions are being made in secret.

Explainability enables accountability - When someone is affected by an AI decision, they deserve to know why. This helps them understand if the decision was fair and gives them information to appeal if needed.

Together, they enable informed consent - People can only meaningfully agree to AI use if they understand both how systems work (transparency) and how decisions affecting them are made (explainability).

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

Data Collection and Processing:

- AI developers must get explicit consent before using personal data to train models. They can't just collect data without telling people what it's for as people have the right to know if their data is being used to train AI systems. In cases of companies, they must have a legal basis for processing data, not just because it's useful for AI

Right to Explanation:

- GDPR gives people the right to understand automated decision-making that significantly affects them eg If an AI system rejects your job application or loan, you can demand an explanation. This forces AI developers to build explainable systems, not just accurate ones as companies must be able to explain their AI's logic in human-understandable terms

Data Subject Rights:

- **Right to Access**: People can ask what data an AI system has about them
- **Right to Rectification**: If the AI has wrong information, people can demand it be corrected
- **Right to Erasure**: People can ask for their data to be deleted from AI systems
- **Right to Portability**: People can take their data and move it to different AI services

Impact on AI Development:

- Developers must build "privacy by design" - considering privacy from the start, not as an afterthought. AI systems need data governance features like audit trails and deletion capabilities
- Companies must conduct privacy impact assessments before deploying AI .This makes AI development more complex but also more trustworthy

2. Ethical Principles Matching

A) Justice → 4. Fair distribution of AI benefits and risks

- Justice in AI means ensuring that benefits and harms are distributed fairly across all groups in society

B) Non-maleficence → 1. Ensuring AI does not harm individuals or society

- Non-maleficence means "do no harm" - the AI equivalent of the medical principle requires actively preventing AI systems from causing damage to people or communities

C) Autonomy → 2. Respecting users' right to control their data and decisions

- Autonomy means respecting people's freedom to make their own choices, this means not manipulating people and giving them control over how AI affects their lives

D) Sustainability → 3. Designing AI to be environmentally friendly

- Sustainability in AI means considering the environmental impact of training and running AI systems by making AI energy-efficient and considering its carbon footprint