AI Ethics – Short Answer Questions and Ethical Principles

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

Algorithmic bias refers to the phenomenon whereby an artificial intelligence system produces results that are systematically prejudiced due to erroneous assumptions in the machine learning process. These biases often stem from the data used to train the model, the way the algorithm is structured, or how decisions are interpreted.Bias can be introduced at various stages, including data collection, feature selection, and model interpretation.

Examples include:

1. Predictive Policing: AI-driven predictive policing tools disproportionately target minority communities due to biased historical data, creating a harmful feedback loop.
2. Healthcare Diagnostics: Algorithms underestimating the needs of Black patients due to using healthcare spending as a proxy for need, which reflects systemic inequality.

This issue is not merely technical—it’s deeply social, requiring interdisciplinary solutions such as algorithmic audits, inclusive datasets, and transparent design practices.

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

Transparency refers to openness in how AI systems are built, including access to algorithms, training data, and intended use. Explainability means the ability for humans to understand why an AI system made a specific decision.

Both are vital: transparency allows for scrutiny and legal compliance, while explainability enables users to interpret and challenge decisions. In high-stakes scenarios such as loan approvals or criminal sentencing, lack of either leads to ethical opacity and undermines public trust.

Example: A university's AI admissions system must not only disclose its data sources (transparency) but also explain why a student was rejected (explainability). Balancing accuracy and interpretability is key to ethical AI.

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

GDPR reshapes AI development by embedding data rights and privacy into its core. Key provisions include:

- Informed Consent: AI systems must collect and use personal data only with clear, explicit user consent.
- Right to Explanation: Individuals can demand justification for decisions made by AI, reinforcing explainability.
- Data Minimization and Privacy by Design: Developers must limit data use to what is strictly necessary and embed privacy into the system from the beginning.

Example: An AI hiring tool in the EU must explain its decisions and use only relevant, consented data. GDPR ensures that ethical principles like autonomy and accountability are upheld in AI systems.

Principle	Definition	Explanation
Non-maleficence	Ensuring AI does not harm individuals or society.	Avoiding direct or indirect harm from AI systems such as misinformation, discrimination, or manipulation.
Autonomy	Respecting users' right to control their data and decisions.	Users must have the ability to consent, opt-out, and challenge decisions made by AI.
Sustainability	Designing AI to be environmentally friendly.	Developing models that are energy-efficient and reduce carbon footprint.
Justice	Fair distribution of AI benefits and risks.	Ensuring that AI outcomes are equitable across different populations, avoiding bias and exclusion.

Ethical Principles Matching – With Explanations