Part 1 Theoretical Understanding

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

Definition: Algorithmic bias occurs when a model’s outputs systematically and unfairly favor or harm members of particular groups (e.g., defined by race, gender, age) due to data, modeling choices, or deployment context, producing disparate outcomes that reflect or amplify unfair societal patterns.

Examples:

1. **Training-data bias:** A recruitment model trained on historical hires at a firm where few women were hired learns patterns correlated with male-dominated resumes and penalizes female applicants (as in Amazon’s recruiting tool).
2. **Measurement / label bias:** A predictive policing model trained on arrest records reflects policing practices (over-policing some neighborhoods), so higher predicted risk in those neighborhoods stems from biased labels rather than true underlying crime rates.

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

- **Transparency**: Openness about how a system is designed, what data it uses, model architecture, training processes, and governance — i.e., the “what” and “how” at a system/process level.
- **Explainability**: The ability to give understandable reasons for a specific model prediction or decision (e.g., feature contributions for a single prediction).

Why both matter: Transparency builds trust and accountability at the system level (auditors, regulators, stakeholders understand data sources, collection processes, governance). Explainability supports affected individuals and operators to understand particular decisions, enabling recourse, error correction, and debugging. Together they reduce harm, enable compliance, and improve adoption.

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

- **Lawful basis & purpose limitation:** Personal data processing for AI must have a legal basis (consent, contract, legitimate interests) and must be fair and limited to intended purposes.
- **Data minimization & storage limitation:** Collect only necessary data and keep it no longer than required.
- **Rights of individuals:** Right to access, rectify, erase, restrict processing, portability; plus the right not to be subject to decisions based solely on automated processing that produce legal or similarly significant effects (Article 22) triggers transparency/explainability needs.

- **Data protection impact assessments (DPIAs):** Required for high-risk processing (many AI systems), forcing privacy risk analysis and mitigation.
- **Privacy by design/default & accountability:** Developers must implement safeguards, document processing, and demonstrate compliance.

Ethical Principles Matching

Match the definitions:

- Ensuring AI does not harm individuals or society. → **B) Non-maleficence**
- Respecting users' right to control their data and decisions. → **C) Autonomy**
- Designing AI to be environmentally friendly. → **D) Sustainability**
- Fair distribution of AI benefits and risks. → **A) Justice**