Part 1: Theoretical Understanding

---

1. Short Answer Questions

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

Definition:
Algorithmic bias refers to the presence of systematic and unfair discrimination in the outcomes or decisions made by AI systems. This bias often stems from skewed training data, historical inequalities, incomplete data collection, or biased algorithmic assumptions. It can perpetuate or even amplify existing societal inequities.

Example 1:
A hiring algorithm trained on a company's historical data may prioritize male candidates over female ones. If the historical data reflects gender imbalances (e.g., male-dominated leadership roles), the algorithm may learn to associate male candidates with higher performance or suitability, thereby disadvantaging equally or more qualified women.

Example 2:
Facial recognition systems often show higher error rates when identifying non-white individuals. For instance, a study by the MIT Media Lab found that commercial facial recognition tools misidentified Black women at a rate of up to 35%, compared to nearly 0% for white men. This disparity is largely due to imbalanced datasets used to train the models.

Added Insight:
Bias in AI is not always intentional, but its consequences are real. As AI becomes more integrated into policing, credit scoring, education, and healthcare, unchecked bias can lead to serious harm such as wrongful arrests, loan denials, or misdiagnoses.

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

Transparency refers to the openness with which AI systems, their data sources, training methods, objectives, and limitations are communicated to stakeholders. It enables scrutiny and ensures developers are accountable for the system's design.

Explainability (or interpretability) focuses on how well the internal logic and outcomes of an AI model can be understood by humans—especially non-experts. Explainability tools help interpret decisions, particularly in complex models like deep learning.

Why They Matter:

- **Transparency** is crucial for regulatory compliance, ethical oversight, and public trust. Without it, stakeholders cannot assess whether the system was built responsibly.

- **Explainability** is vital for users and decision-makers to understand *why* a model made a certain prediction or decision. This is especially important in high-stakes areas like healthcare, criminal justice, and finance.

Example:
In credit scoring, a transparent system would disclose the data used (income, debt-to-income ratio, etc.), while an explainable system would allow a rejected applicant to understand exactly *why* they were denied credit, and how they might improve.

Future Consideration:
As AI models become more complex (e.g., large language models or deep neural networks), achieving both transparency and explainability becomes more challenging—but also more necessary.

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

GDPR, enforced since 2018, significantly shapes how AI is developed, deployed, and governed within the EU. It prioritizes individual data rights, fairness, and accountability.

Key Impacts:

1. **Data Minimization:**
   AI developers must only collect data necessary for their models. Over-collection violates GDPR, prompting more thoughtful data use.

2. **Informed Consent:**
   Users must understand and agree to how their personal data will be used. AI systems relying on user data must have clear, accessible consent mechanisms.

3. **Right to Access, Rectify, and Erase:**
   Individuals have the right to access their data, correct inaccuracies, or request its deletion (“right to be forgotten”), posing challenges for persistent AI training datasets.

4. **Right to Explanation:**
   Article 22 of GDPR grants individuals the right *not* to be subject to decisions based solely on automated processing (e.g., credit decisions) without meaningful human oversight. AI systems must offer **explanations** for decisions impacting users

significantly.

5. **Restrictions on Profiling and Automated Decisions:**
   Profiling is allowed only under specific conditions, and users must be informed when automated decisions are being made about them. Human-in-the-loop processes are often required.

Implication:
GDPR promotes **ethically aligned AI** but also increases compliance costs and development complexity, especially for companies relying heavily on large-scale data collection and autonomous systems.

---

2. Ethical Principles Matching

Match each ethical principle with its correct description and expand its relevance in AI:

- **A) Justice** → *Fair distribution of AI benefits and risks.*
    - **Expanded:** Justice ensures AI doesn't disproportionately harm or exclude vulnerable populations. For example, ensuring healthcare AI tools work equally well across races and genders helps reduce inequality.
- **B) Non-maleficence** → *Ensuring AI does not harm individuals or society.*
    - **Expanded:** Developers must avoid unintended consequences such as surveillance misuse, misinformation spread, or job displacement without safeguards.
- **C) Autonomy** → *Respecting users' right to control their data and decisions.*
    - **Expanded:** Autonomy supports informed consent and the right to opt-out of AI-driven decisions. AI systems must not manipulate users' behavior through hidden nudges.
- **D) Sustainability** → *Designing AI to be environmentally friendly and socially responsible.*
    - **Expanded:** AI development must consider carbon footprint (e.g., training large models like GPT consumes massive energy) and long-term societal impacts, including ethical labor practices and inclusive design.