Theoretical Ethics & Concepts

Q1: Define Algorithmic Bias and give Examples

Definition:
Algorithmic bias refers to systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one group over another. These biases often arise from biased training data, flawed assumptions in model design, or unintentional human influence.

Example 1: Facial Recognition Bias
Some facial recognition systems have shown higher error rates for individuals with darker skin tones. A 2019 study by MIT found that commercial facial recognition software had error rates as high as 35% for dark-skinned women, while accuracy for light-skinned men was nearly 100%. This demonstrates how biased datasets can lead to discriminatory technology.

Example 2: Gender Bias in Job Recruitment Tools
Amazon's experimental AI recruiting tool was found to favor male candidates by downgrading resumes that included the word "women's" (like "women's chess club captain"). This happened because the tool was trained on past hiring data that reflected gender imbalances, reproducing historical bias.

Q2: Transparency vs Explain ability

Transparency:
Transparency in AI refers to how much we can "see into" the internal workings of a system. This includes knowing what data was used, how the model was trained, and what logic or algorithms are applied. A transparent system openly shares how it functions.

Explain ability:
Explain ability focuses on how well the system's decisions can be understood by humans. Even if the internal workings are complex (like in deep learning models), the system should offer clear reasoning or justification for its outputs.

Difference in Summary:

Transparency is about system visibility: *"How does it work?"*

Explainability is about interpretability: *"Why did it do that?"*

Both are crucial for building **trustworthy AI** but serve different roles in ethics and accountability.

Q3: GDP R's Impact on AI in the EU

The **General Data Protection Regulation (GDPR)**, implemented in the EU in 2018, has reshaped how AI is developed and used. It places **strict obligations** on organizations to ensure AI systems respect human rights, privacy, and fairness.

Key Impacts on AI

Right to Explanation:
Individuals have the right to receive meaningful information about the logic involved in automated decisions (Article 22). This limits the use of "black-box" models without human oversight.

Data Minimization and Consent:
AI must only use data that is adequate, relevant, and limited to what is necessary. Clear, informed consent is required for personal data usage.

Accountability and Fairness:
Organizations are required to demonstrate that their AI systems do not discriminate and are regularly tested for fairness, accuracy, and transparency.

Conclusion:
GDPR encourages the development of **human centric, ethical AI**, promoting transparency, user control, and fairness in automated decision-making.

Match 4 Ethical Principles to Definitions

Ethical Principle	Definition
Autonomy	Respecting individuals' right to make informed choices about how their data is used.
Justice	Ensuring fairness and equal treatment for all, avoiding discrimination in AI outputs
Non-maleficence	Avoiding harm to individuals or groups as a result of AI actions or decisions
Beneficence	Promoting well-being and designing AI to provide benefits to individuals and society.