Theme: Designing Responsible and Fair AI Systems

Q1: Define algorithmic bias and provide two examples from real-world AI

Definition: Systematic and unjust discrimination ingrained in an AI system's output is known as algorithmic bias. It happens when the system favors one group over another in its predictions or choices because of unfair model design, faulty assumptions, or biased training data. Social injustices that are inadvertently encoded during development are frequently reflected in algorithmic bias.

Example 1: Tool for COMPAS Recidivism
The U.S. criminal justice system's recidivism risk was predicted using the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm. Despite having similar criminal histories, the system disproportionately gave Black defendants higher risk scores than white offenders, according to a 2016 ProPublica investigation. This prejudice resulted from the use of past crime statistics that showed racial differences in sentencing and police.

Example 2: Amazon's AI Hiring Tool
An AI hiring tool that was discovered to punish resumes that contained the phrase "women's" (e.g., "women's chess club") was withdrawn by Amazon in 2018. The model learned and magnified gender disparities in tech hiring after being trained on ten years' worth of hiring data. The algorithm's unreasonably lower ranking of female candidates illustrates how systemic exclusion can be sustained by biased training data.

In conclusion, these illustrations show how algorithmic bias can have negative effects in the real world, such as fostering prejudice, stifling opportunity, and undermining trust. At every level of development, prejudice must be proactively identified and fixed in ethical AI design.

Q2: Distinguish between transparency and explainability in AI and their significance.

Transparency refers to the openness of an AI system's design, objectives, data sources, and development processes. It allows stakeholders to inspect and understand how and why an AI system was created.

Explainability, on the other hand, focuses on how well the decisions or predictions made by the AI can be understood by humans—especially end users. It seeks to make the outputs of "black box" models interpretable through explanations.

Key Differences:

Features	**Transparency**	**Explainability**
Focus	System/process-level clarity	Output/decision-level clarity

Audience	Regulators, developers	Users, decision-makers
Tools/Methods	Open documentation, audits	SHAP, LIME, feature importance

Why Both Matter:

Accountability: Transparency enables external audits, while explainability helps assign responsibility for decisions.

User Trust: Explainable models build user confidence by revealing reasoning behind outcomes.

Compliance: Regulations like the GDPR's "right to explanation" mandate explainability in automated decisions.

Fairness Checks: Both are critical in identifying and mitigating hidden biases.

Conclusion:
Together, transparency and explainability form the foundation for ethical, human-centric AI. One reveals the system's construction; the other reveals its logic in action.

Q3: Analyze how GDPR affects AI development in the EU.

An important piece of EU legislation that regulates the collection, storage, and processing of personal data is the **General Data Protection Regulation (GDPR)**. It has wide-ranging and significant effects on the advancement of AI.

Principal Effects on AI:

Data Minimization (Article 5): AI developers must make sure that just the personal information that is required is gathered, which restricts overzealous data collection techniques and promotes simple, goal-driven models.

Right to Explanation (Article 22): People have the right to know the reasoning behind, importance of, and ramifications of choices that are made only through automated means. This deters opaque "black box" models and forces **explainability** in AI systems.
Articles 6 and 7: Consent and Lawful Processing: AI systems must have a valid reason for processing personal data, usually a contract, legitimate interest, or consent. Sensitive data requires explicit, informed user consent.

Rights of Data Subjects: People have the right to request that their data be accessible, corrected, deleted, or made portable. This affects the curation and upkeep of datasets and adds another level of control over training data.

Accountability and Impact Assessments (Article 35): To assess and reduce privacy concerns, AI systems that pose a high risk to individual rights must undergo a **Data Protection Impact Assessment (DPIA)**. This encourages ethical foresight and proactive risk evaluation.

Conclusion:
GDPR has made privacy and ethical compliance central to AI development in the EU. Developers must now integrate data protection by design, ensuring fairness, transparency, and respect for individual rights from the ground up.

Ethical Principles Matching

Principle	**Definition**
A) Justice	4) Fair distribution of AI benefits and risks
B) Non-maleficence	1) Ensuring AI does not harm individuals or society
C) Autonomy	2) Respecting user's right to control their data and decisions
D) Sustainability	3) Designing AI to be environmentally friendly

References

European Commission. (2019). Ethics Guidelines for Trustworthy AI.

ProPublica. (2016). Machine Bias: Risk Assessments in Criminal Sentencing.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a “right to explanation”.

ACM Code of Ethics and Professional Conduct. (2018).

Wired. (2018). Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women.