PART 1

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

ANS

Algorithmic bias occurs when AI systems produce unfair or discriminatory outcomes due to flawed data, design, or decision-making processes. It often reflects existing societal prejudices or imbalances in training data.

Examples:

1. Facial Recognition: AI may misidentify people of color more often than white individuals due to underrepresentation in training data.

2. Hiring Tools: Algorithms trained on biased historical data may favor male candidates for technical roles, disadvantaging women.

Such biases perpetuate inequality if left unchecked.

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

ANS

Transparency in AI refers to openness about how a system is developed, including data sources, design choices, and potential biases.

Explainability in AI focuses on making individual AI decisions understandable to users, often through clear reasoning or interpretable models.

Transparency builds trust by revealing the system's workings at a high level, while explainability ensures users can follow specific decisions (e.g., why a loan was denied). Both are crucial: transparency holds developers accountable, and explainability empowers users to challenge unfair outcomes. Without them, AI risks being a "black box," eroding trust and perpetuating harm.

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

  ANS

The GDPR (General Data Protection Regulation) significantly impacts AI development in the EU by enforcing strict data protection and ethical standards. Key requirements include:

1. Lawful Basis for Processing – AI systems must have a valid legal reason (e.g., consent or necessity) to use personal data.

2. Data Minimization – AI models can only collect data essential for their purpose, limiting large-scale surveillance or profiling.

3. Transparency & Explainability – Users must be informed about automated decisions and have the right to explanations (Article 22).

4. Right to Explanation – Individuals can contest AI-driven decisions affecting them (e.g., loan denials or job screenings).

5. Bias & Fairness – GDPR's non-discrimination principles push developers to mitigate algorithmic bias in AI systems.

6. Privacy by Design – AI must integrate data protection from the outset, reducing risks like re-identification in anonymized datasets.

7. Accountability & Audits – Companies must document AI decision-making processes and ensure compliance, increasing regulatory scrutiny.

8. Data Subject Rights – Users can access, correct, or delete their data, complicating AI training on historical datasets.

9. Cross-Border Data Flows – AI firms must comply with GDPR even if data is processed outside the EU, affecting global deployments.

These rules encourage ethical AI but also increase compliance costs, pushing developers toward federated learning, synthetic data, or interpretable models to meet legal standards while innovating responsibly.

Ethical Principles Matching

Here’s the correct matching of principles to their definitions:

(A) Justice: Fair distribution of AI benefits and risks.

(B) Non-maleficence: Ensuring AI does not harm individuals or society.

(C) Autonomy: Respecting users’ right to control their data and decisions.

(D) Sustainability: Designing AI to be environmentally friendly.

PART 2

Case Study Analysis: Amazon’s Biased Hiring Tool

Source of Bias

The bias stemmed from historical training data—Amazon’s AI was trained on resumes submitted over a 10-year period, which were predominantly from male applicants. Since the tech industry has been male-dominated, the algorithm learned to favour male candidates by downgrading resumes containing words like "women’s" (e.g., "women’s chess club captain") or graduates of all-women colleges.

Proposed Fixes to Improve Fairness

1. Debias Training Data
   - Use gender-neutral datasets or synthetically balance underrepresented groups.
   - Remove gendered keywords and focus on skills/experience.
2. Adversarial Fairness Testing
   - Implement AI "counterfactual tests" (e.g., tweak only gender indicators in resumes to check for scoring disparities).
3. Human-in-the-Loop Oversight
   - Combine AI with human reviewers to audit shortlisted candidates and flag potential biases.

Fairness Evaluation Metrics

Post-correction, track:

1. Demographic Parity – Equal selection rates across genders.
2. Equal Opportunity – Similar true positive rates (qualified candidates hired) for male vs. female applicants.
3. Disparate Impact Ratio – Ensure scores don’t disproportionately favour one group (e.g., <80% difference is fair).

Case Study Analysis: Facial Recognition in Policing

Ethical Risks

1. **Wrongful Arrests & Discrimination**
   - Higher misidentification rates for minorities (e.g., Black, Asian individuals) can lead to unjust detentions, reinforcing systemic bias.
   - Example: *Robert Williams' wrongful arrest* due to flawed facial recognition.
2. **Privacy Violations & Mass Surveillance**
   - Unregulated use enables unwarranted tracking, chilling free assembly and movement.
   - Risk of misuse (e.g., targeting activists, political dissidents).
3. **Lack of Transparency & Accountability**
   - Many systems are proprietary "black boxes," preventing scrutiny of bias or errors.

Policies for Responsible Deployment

1. **Legislative Bans or Strict Regulation**
   - *Example:* EU's proposed AI Act bans real-time facial recognition in public spaces.
   - Require judicial warrants for use in criminal investigations.
2. **Bias Mitigation & Testing**
   - Mandate third-party audits (e.g., NIST's racial bias benchmarks).
   - Only deploy systems with <1% false-positive rates across demographics.
3. **Transparency & Public Oversight**
   - Publish accuracy disparities by race/gender (like IBM's 2019 disclosure).
   - Establish civilian review boards to approve and monitor usage.
4. **Alternative Solutions**
   - Prioritize non-biased tools (e.g., witness lineups, detective work) over automated recognition.

Key Takeaway: Without strict safeguards, facial recognition risks exacerbating injustice. Policies must balance security needs with civil rights protections.
