**Q1: Define algorithmic bias and give two examples.**

**Definition:** Algorithmic bias occurs when an AI system produces systematically unfair outcomes for certain groups due to biased training data, model design, or deployment context. It leads to disparate accuracy, false positives/negatives, or resource allocation across protected groups.

**Examples:**

1. **Hiring model penalizing women** — trained on historical hiring data where males were hired more often; model learns and reproduces that preference.
2. **Facial recognition misidentifying darker-skinned people** — dataset underrepresents darker-skinned faces so the model has lower accuracy and higher false positives for those groups.

**Q2: Difference between transparency and explainability; why both matter.**

* **Transparency** means that stakeholders can access information about the AI system (e.g., data sources, training procedures, versioning, evaluation metrics). It’s about *openness and documentation*.
* **Explainability** (or interpretability) means the ability to provide understandable reasons for **specific** predictions or decisions (e.g., which features drove a diagnosis).

**Why both matter:**

* **Transparency** builds trust and enables auditing and governance.
* **Explainability** allows affected individuals and practitioners to understand, contest, and act upon outputs. Together they enable accountability, detection of bias, and regulatory compliance.

**Q3: How GDPR impacts AI development in the EU**

* **Lawful basis & consent:** Personal data processing often requires explicit consent or another lawful basis.
* **Data minimization:** Collect only what's necessary; avoid unnecessary features.
* **Right to explanation / access:** Individuals can request access to data about them and meaningful information about automated decisions.
* **Data subject rights:** Right to rectification, erasure (right to be forgotten), and portability — impacts data pipelines and model retraining.
* **Privacy by design:** Developers must embed data protection principles early in design and document compliance (impact assessments may be required).

**2. Ethical Principles Matching**

Match lines:

* 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**

**Source(s) of bias:**

* **Training data bias:** Historical hiring data favored men; hiring decisions reflect past bias.
* **Feature proxies:** Using proxies (e.g., college, job gaps) that correlate with gender leads to disparate impact.

**Three fixes to make it fairer:**

1. **Re-sample or reweight training data** to reduce historical imbalance (oversample underrepresented class or use group-aware weighting).
2. **Remove or neutralize proxy features** that unfairly correlate with protected attributes (or use fairness-aware feature selection).
3. **Apply fairness-aware learning / post-processing:** e.g., use algorithms that constrain statistical parity or equalized odds, or apply calibrated adjustments to predictions.

**Metrics to evaluate fairness post-correction:**

* **Statistical parity difference** (difference in positive outcome rates between groups).
* **Equalized odds / False Positive Rate parity** (difference in FPR and FNR between groups).
* **Disparate impact ratio** (ratio of selection rates).

**Facial Recognition in Policing**

**Ethical risks:**

* **Wrongful arrests / wrongful suspicion** — disproportionate harm to misidentified groups.
* **Privacy violations** — pervasive surveillance without consent.
* **Chilling effects** — reduced freedom of movement/association.
* **Disparate enforcement** — criminal justice disparities amplified.

**Policies for responsible deployment:**

1. **Strict use-case limits:** Only use for well-justified, high-risk scenarios with human oversight.
2. **Mandatory human-in-the-loop review:** Automated matches must be verified by trained humans before action.
3. **Audit & transparency:** Regular independent audits for accuracy and bias; publish aggregate performance broken down by race/gender.
4. **Data governance & consent:** Limit retention, restrict access, and obtain legal/ethical approvals.
5. **Ban until safe:** Consider moratoria or strict regulation for use-cases with unacceptable risk.