AI Ethics Assignment – Written Answers

Part 1: Theoretical Understanding (30%)

Q1: Define algorithmic bias and provide two examples

Definition:

Algorithmic bias is systematic and unfair discrimination produced by an AI system due to biased data, flawed modeling choices, or deployment context that disadvantages certain groups.

Examples:

- A hiring model trained on past resumes undervalues women by penalizing terms associated with women's colleges, because historical hiring favored men.
- A credit scoring model yields higher false negatives for historically marginalized neighborhoods due to proxy variables (e.g., ZIP code) correlating with race and socioeconomic status.

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

- Transparency: Openness about how an AI system is built and deployed—data sources, objectives, model type, training processes, governance, and limitations.
- Explainability: The ability to provide understandable reasons for specific model outputs—why this prediction, for this person, under these features.
- Importance: Transparency enables accountability and external scrutiny; explainability enables contestability, error correction, and meaningful user consent.

Q3: How does GDPR impact AI development in the EU?

- Lawful basis & purpose limitation: AI processing of personal data must have a lawful basis and a defined, legitimate purpose.
- Data minimization & storage limitation: Collect only what's necessary; don't retain data longer than needed.
- Rights and safeguards: Individuals have rights to access, rectification, erasure, and to object to automated decision-making with significant effects; organizations must enable human review.

- Privacy by design & DPIAs: Build privacy into systems and conduct Data Protection Impact Assessments for high-risk processing.

- Accountability: Maintain records, contracts, and governance to demonstrate compliance.

Ethical Principles Matching

- 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 (40%)

Case 1: Biased Hiring Tool

- Source of bias: Historical data bias (male-dominated hiring), proxy variables (gender-encoded features), evaluation bias (accuracy-only metrics).

- Three fixes:

  1. Data rebalancing and feature auditing (remove gender proxies, rebalance training data).

  2. Fairness-aware learning (algorithms enforcing fairness constraints).

  3. Structured human oversight (blind screening, diverse validation panels).

- Fairness metrics: Disparate impact ratio, equal opportunity difference, false positive/negative rate parity, calibration within groups.

Case 2: Facial Recognition in Policing

- Ethical risks: Wrongful arrests, privacy violations, disparate impact, function creep.

- Policies for responsible deployment:

  - Strict use limitations (require corroborating evidence, judicial authorization).

  - Mandatory bias audits with public reporting.

  - Transparency & community oversight.

- Human-in-the-loop and appeals mechanisms.
- Strong data governance (limit retention, prohibit nonconsensual expansion).

Part 3: Practical Audit (25%)

The COMPAS dataset exhibits racial disparities in error rates and selection outcomes. Baseline evaluation with a logistic regression model shows unequal false positive and false negative rates across privileged and unprivileged groups, alongside gaps in statistical parity and equal opportunity. Disparate impact deviates from parity (ideal ≈ 1), indicating that predicted risk classifications are not distributed evenly across racial groups. These disparities matter: higher false positives for an unprivileged group mean individuals are more likely to be labeled “high risk” incorrectly, imposing greater burdens and potential downstream harms (e.g., stricter supervision), while higher false negatives can undermine legitimate public safety concerns.

We applied the AIF360 Reweighing pre-processing algorithm, which adjusts instance weights to balance the influence of privileged and unprivileged groups during training. Post-remediation, parity metrics improved: disparate impact moved closer to 1; equal opportunity difference narrowed; and average odds difference and statistical parity difference showed reduced gaps. Although not perfect, these shifts demonstrate that fairness-aware preprocessing can meaningfully mitigate bias while maintaining usable predictive performance.

Remediation should be multi-pronged. First, adopt fairness-aware training (e.g., reweighing, adversarial debiasing) and monitor subgroup performance continuously. Second, calibrate decision thresholds per group to align true positive rates and reduce error disparities, while transparently reporting any adjustments. Third, establish governance: periodic independent audits, documentation of data lineage and feature choices, and stakeholder review to assess impact. Finally, ensure transparency and contestability—provide explanations for individual risk scores and accessible channels for review. Fairness is not a one-off fix; it requires ongoing evaluation, inclusive oversight, and clear accountability to align predictive systems with ethical and legal standards.

Part 4: Ethical Reflection (5%)

I plan to embed ethical AI principles into my projects by design and governance. I’ll define a clear purpose and limit data collection to what’s strictly necessary, applying privacy-by-design and conducting impact assessments before deployment. I’ll prioritize representativity in data, audit features for proxies that encode sensitive attributes, and track fairness metrics like disparate impact and equal opportunity difference across relevant groups. Explanations will be user-centered, with concise, accessible rationales for decisions and avenues to challenge outcomes. I’ll institute human-in-the-loop review for high-stakes decisions, document assumptions, and publish model cards detailing performance and limitations. Finally, I’ll commit to ongoing monitoring, community feedback, and periodic external audits to ensure the system remains fair, accountable, and aligned with societal values.