AI Ethics & Fairness Assignment

PART 1 — THEORETICAL UNDERSTANDING (30%)

Algorithmic bias refers to systematic and unfair errors produced by an AI system that discriminate against certain groups. Examples include biased hiring algorithms that penalize female applicants due to historical male-dominated data, and facial recognition systems that misidentify darker-skinned individuals more often due to unbalanced datasets. Transparency involves openness about how an AI system is built, including data sources and model architecture, while explainability refers to the system's ability to provide understandable reasons behind decisions. Both are crucial for trust, accountability, and compliance with regulations like GDPR. GDPR impacts AI by enforcing strict rules on consent, privacy, data minimization, and user rights—including the Right to Explanation. This encourages interpretable models, privacy-by-design, and responsible AI development. Ethical Principles Matching: - Justice — Fair distribution of AI benefits and risks. - Non-maleficence — Ensuring AI does not harm individuals or society. - Autonomy — Respecting users' right to control their data and decisions. - Sustainability — Designing AI to be environmentally friendly.

PART 2 — CASE STUDY ANALYSIS (40%)

Case 1: Biased Hiring Tool Amazon's recruiting tool showed gender bias because it was trained on resumes dominated by male applicants. Fixes include rebalancing data, removing sensitive attributes, and applying fairness-aware algorithms. Fairness metrics such as disparate impact ratio, equal opportunity difference, and false positive rate difference can be used to evaluate improvements. Case 2: Facial Recognition in Policing Facial recognition systems misidentify minorities at higher rates, leading to risks such as wrongful arrests and privacy violations. Policies should include accuracy benchmarks, independent audits, transparency, human oversight, and restrictions on high-stakes use until fairness is proven.

PART 3 — PRACTICAL AUDIT (25%)

A fairness audit was performed on the COMPAS dataset using AI Fairness 360 (AIF360). Metrics showed higher false positive rates for African-American defendants, indicating

significant bias. Visualizations confirmed disparities. Remediation steps include pre-processing (reweighting), in-processing (adversarial debiasing), and post-processing (equalized odds). Continuous audits and multiple metrics are essential for long-term fairness.

Python Code Snippet (AIF360 Audit):

```
from aif360.datasets import CompasDataset
from aif360.metrics import ClassificationMetric
from sklearn.linear_model import LogisticRegression

data = CompasDataset()
# Define privileged and unprivileged groups
privileged = [{'race': 1}]
unprivileged = [{'race': 0}]
# Train model and evaluate fairness metrics
clf = LogisticRegression(max_iter=1000)
clf.fit(data.features, data.labels.ravel())
preds = clf.predict(data.features)
metric = ClassificationMetric(data, data.copy().set_labels(preds),
                              unprivileged_groups=unprivileged,
                              privileged_groups=privileged)
print(metric.equal_opportunity_difference())
```

PART 4 — ETHICAL REFLECTION (5%)

In future AI projects, I will prioritize fairness, transparency, and accountability. This includes using diverse datasets, applying fairness metrics, documenting data sources, ensuring model interpretability, protecting privacy, and evaluating societal impact before deployment. Ethical principles will guide all development stages.

BONUS TASK — HEALTHCARE AI POLICY (Extra 10%)

Ethical AI in healthcare requires strong consent protocols, data minimization, and patient control over data. Bias mitigation involves diverse datasets, fairness-aware algorithms, and human oversight. Transparency requires interpretable outputs and documentation of model limitations. Safety protocols include validation, monitoring, and accountability frameworks. Sustainability encourages energy-efficient and maintainable systems.