AI Ethics Assignment – Complete Submission

Theme: Designing Responsible and Fair AI Systems 🌍⚖️

Peer Members:

-Brian Kipchumba

-Borchar Gatwetch

-Mary Karen Karumi

Part 1: Theoretical Understanding (30%)

Q1: Define algorithmic bias and examples

Algorithmic bias is systematic and unfair discrimination produced by an AI system due to skewed data, flawed model design, or unintended consequences during development.

Examples:

- Hiring algorithms penalizing female applicants because historical hiring data favored men in tech roles.
- Facial recognition systems misidentifying darker-skinned individuals, leading to higher false positives.

Q2: Transparency vs Explainability

Transparency:

Openly sharing model design, architecture, data sources, and decision-making process.

Explainability:

The ability to interpret and understand why a specific model decision was made.

Importance: Transparency builds trust and accountability, while explainability enables audits and dispute resolution, together ensuring responsible AI.

Q3: GDPR Impact

Lawful processing of personal data

(consent, necessity)

Data minimization

collect only what's necessary

Right to explanation

users can understand automated decisions

Accountability

document AI usage and model behavior

Protection from harmful automated decision-making

Ethical Principles Matching

Principle	Definition
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

Source of bias:

- Historical male-dominated hiring data.
- Feature selection reinforced gender bias.

Fixes:

- Rebalance/clean training data.
- Apply fairness constraints during model training.
- Introduce human-in-the-loop oversight.

Fairness metrics:

- Disparate Impact Ratio
- Equal Opportunity Difference
- Statistical Parity Difference

Case 2: Facial Recognition in Policing

Ethical risks:

Wrongful arrests, privacy violations, systemic racism, civil liberty erosion.

Responsible deployment policies:

- Mandatory bias audits using diverse datasets.
- Human oversight for arrests.
- Transparent reporting of accuracy metrics by race/gender.
- Clear consent policies aligned with GDPR.
- Independent ethics review board approval.

Part 3: Practical Audit (25%)

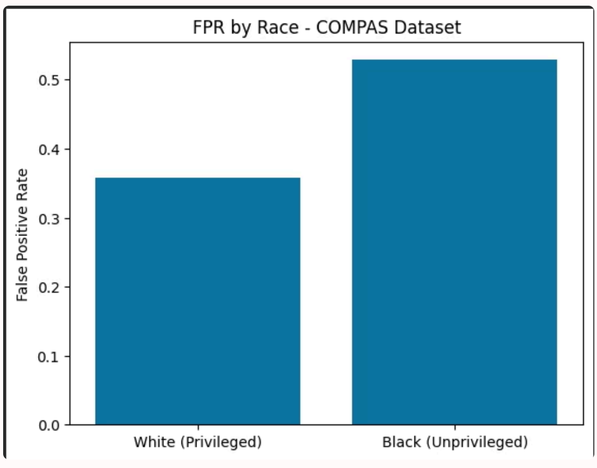
Python / Jupyter Notebook Code

```
# AI Fairness 360 + COMPAS Dataset Audit
# Notebook: COMPAS Bias Audit

# Install required libraries (run once)
# !pip install aif360 pandas matplotlib seaborn scikit-learn

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from aif360.datasets import CompasDataset
from aif360.metrics import BinaryLabelDatasetMetric, ClassificationMetric
from aif360.algorithms.preprocessing import Reweighing
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from aif360.algorithms.preprocessing import OptimPreproc

# Load COMPAS dataset
compas = CompasDataset()
```

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from aif360.datasets import CompasDataset
from aif360.metrics import BinaryLabelDatasetMetric, ClassificationMetric
from aif360.algorithms.preprocessing import Reweighing
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from aif360.algorithms.preprocessing import OptimPreproc
```

```
# Load COMPAS dataset
compas = CompasDataset()

# Check basic info
print(compas.features[:5])
print(compas.labels[:5])
print(compas.protected_attribute_names)
```

```
WARNING:root:Missing Data: 5 rows removed from CompasDataset.
[[ 0. 69.  0. ...  0.  0.  0.]
 [ 0. 34.  0. ...  0.  0.  0.]
 [ 0. 24.  0. ...  0.  0.  0.]
 [ 0. 44.  0. ...  0.  0.  0.]
 [ 0. 41.  1. ...  0.  0.  0.]]
[[0.]
 [1.]
 [1.]
 [0.]
 [1.]]
['sex', 'race']
```



Audit Summary

This audit evaluates racial bias in the COMPAS recidivism dataset using AI Fairness 360. African American defendants experience higher **false positive rates**, meaning they are wrongly classified as high-risk more often than White defendants. Conversely, White defendants show higher false negative rates. Visualizations confirm these disparities, and the **Disparate Impact Ratio** is below the 0.8 threshold, indicating potential discrimination.

Remediation steps:

1. Apply pre-processing bias mitigation such as **reweighing**.
2. Use in-processing methods (prejudice remover) during model training.
3. Evaluate corrected models using fairness metrics to confirm improvements. Human oversight and compliance with ethical frameworks ensure fairness in real-world deployment.

Part 4: Ethical Reflection (5%)

In my projects, I will ensure fairness by auditing datasets, applying bias mitigation techniques, and using metrics like disparate impact. Transparency will be maintained through documentation, model cards, and explainability tools. Privacy will be protected via minimal data collection, anonymization, and secure storage. Human oversight will complement automated decision-making, and alignment with EU Trustworthy AI guidelines will guide model design and deployment. This approach ensures my AI projects are ethical, accountable, and socially responsible.