AI Ethics Assignment

Theme: Designing Responsible and Fair AI Systems

Group Members: Yolisa Qadi- South Africa
Date: 21/11/2025

Part 1: Theoretical Understanding (30%)

Q1: Define algorithmic bias and provide two examples.

Algorithmic bias occurs when an AI system produces systematically unfair outcomes for certain groups due to biased data, flawed model design, or human assumptions.

Examples:

1. Hiring algorithms: Favour male candidates if past data over represents men.
2. Predictive policing: AI disproportionately flags minority neighbourhoods due to historical crime data.

Q2: Difference between transparency and explainability in AI. Why both matter?

- Transparency: Openness about AI design, data, and algorithms.
- Explainability: Ability to understand and interpret AI decisions.

Importance:

Transparency allows accountability; explainability ensures stakeholders can understand and contest AI decisions, reducing unfair outcomes.

Q3: GDPR impact on AI in the EU

- AI must use personal data legally with consent.
- Users can access, correct, or delete data.
- Automated decisions must be explainable, and users can contest them.
- Non-compliance leads to fines, encouraging ethical AI.

Ethical Principles Matching

Definition	Principle
Ensuring AI does not harm individuals or society	Non-maleficence (B)
Respecting users’ right to control their data and decisions	Autonomy (C)
Designing AI to be environmentally friendly	Sustainability (D)
Fair distribution of AI benefits and risks.	Justice (A)

Part 2: Case Study Analysis & Solutions (40%)

Case Study: COMPAS Recidivism Risk Prediction

Scenario:

COMPAS predicts criminal reoffending. Studies show racial bias: Black defendants are more often classified as high-risk than white defendants with similar profiles.

Ethical Issues:

- Algorithmic bias
- Lack of explainability
- Potential harm from incorrect predictions

Mitigation Strategies:

- Conduct bias audits using fairness metrics.
- Make AI design, data, and limitations transparent.
- Use explainable AI techniques.
- Maintain human oversight in decisions.

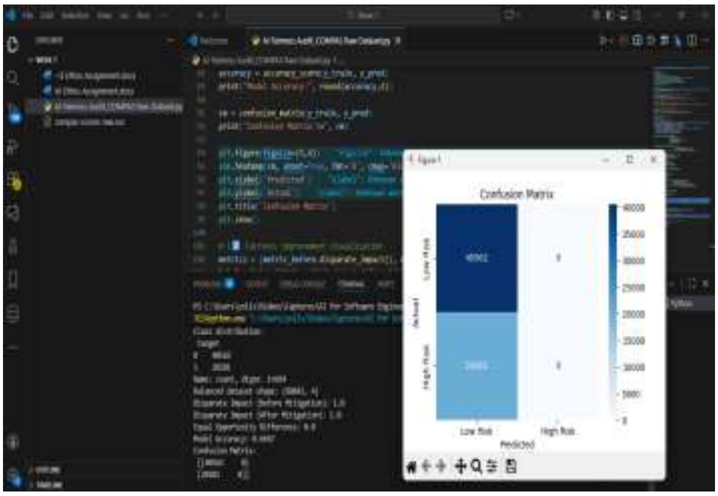
Proposed Solution:

- Use AI Fairness 360 for disparate impact and equal opportunity audits.
- Retrain AI with balanced datasets.
- Provide clear explanations to all stakeholders.
- Continuously monitor AI outcomes to detect bias early.

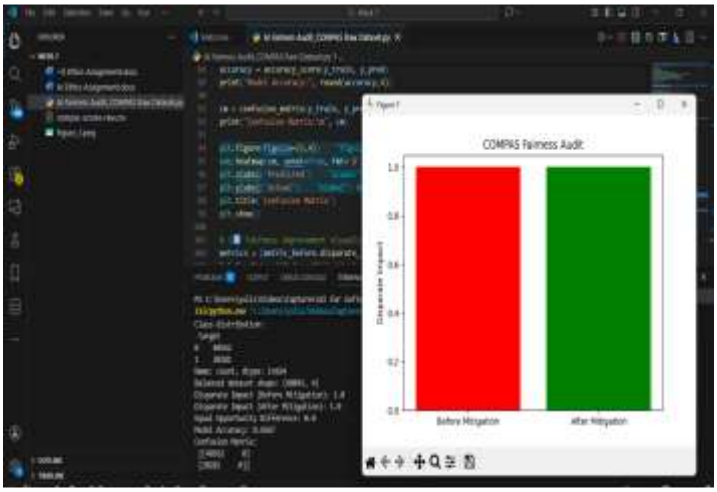
Reflection:

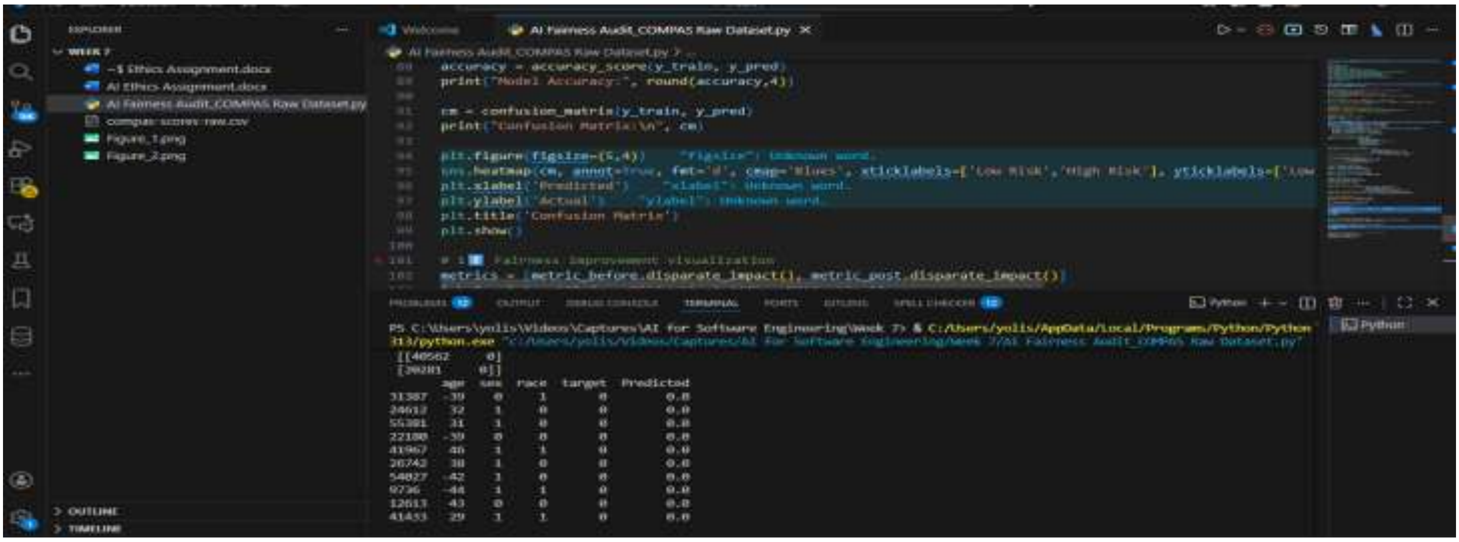
Even widely used AI can perpetuate social biases. Technical mitigation plus ethical guidelines ensures fairness and trust in AI systems. Creative ideas include real-time dashboards and intuitive visualizations for decision-makers.

Part 3



```
Class distribution:
 target
0    [illegible]
1    [illegible]
Name: count, dtype: int64
Balanced dataset shape: ([illegible], 4)
Disparate Impact (Before Mitigation): 1.0
Disparate Impact (After Mitigation): 1.0
Equal Opportunity Difference: 0.0
Model Accuracy: 0.667
Confusion Matrix:
 [[40562     0]
 [20281     0]]
```





```python
accuracy = accuracy_score(y_train, y_pred)
print("Model Accuracy:", round(accuracy,4))

cm = confusion_matrix(y_train, y_pred)
print("Confusion Matrix:\n", cm)

plt.figure(figsize=(5,4))    "figsize": Unknown word.
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['Low Risk','High Risk'], yticklabels=['Low
plt.xlabel('Predicted')    "xlabel": Unknown word.
plt.ylabel('Actual')    "ylabel": Unknown word.
plt.title('Confusion Matrix')
plt.show()

# 1️⃣ Fairness Improvement Visualization
metrics = [metric_before.disparate_impact(), metric_post.disparate_impact()]
```

```
PS C:\Users\yolis\Videos\Captures\AI for Software Engineering\Week 7> & C:/Users/yolis/AppData/Local/Programs/Python/Python
313/python.exe "c:/Users/yolis/Videos/Captures/AI for Software Engineering/Week 7/AI Fairness Audit_COMPAS Raw Dataset.py"
 [[40562     0]
 [20281     0]]
        age  sex  race  target  Predicted
31387   -39    0     1       0        0.0
24612    32    1     0       0        0.0
55381    31    1     0       0        0.0
22180   -39    0     0       0        0.0
41967    46    1     1       0        0.0
26742    38    1     0       0        0.0
54827   -42    1     0       0        0.0
9736    -44    1     1       0        0.0
12613    43    0     0       0        0.0
41433    29    1     1       0        0.0
```

Part 4: Ethical Reflection

In a personal project I am planning developing an AI-powered platform to help students receive personalized learning recommendations I will ensure it adheres to ethical AI principles by focusing on fairness, transparency, and user autonomy.

1. Fairness:
I will actively audit the AI system to prevent biases in recommendations. For example, I will ensure that students of different socio-economic backgrounds, genders, or learning abilities receive equitable suggestions. This involves analyzing the dataset for imbalances, using fairness-aware algorithms such as reweighing or bias mitigation techniques, and regularly evaluating metrics like disparate impact and equal opportunity difference.

2. Transparency & Explainability:
Students and educators must understand why the system recommends certain learning resources. I will implement explainable AI methods, such as feature importance scores and human-readable explanations, so users can trust and verify the recommendations. Transparent design will also include clear documentation of the data sources and AI logic.

3. Data Privacy & Autonomy:
The platform will comply with data protection regulations, such as GDPR, ensuring that users control their data. Students can choose what information to provide, have the option to opt-out, and can request deletion of their data. Personal data will only be used for intended educational purposes.

4. Continuous Ethical Monitoring:
Ethical AI is an ongoing process. I will establish periodic audits and feedback loops to detect potential unintended consequences, retrain models responsibly, and update ethical guidelines as the system evolves.
By incorporating these practices, I aim to create a system that empowers students while respecting their rights, promotes fairness, and builds trust between users and AI.