PLP ACADEMY

FEBRUARY COHORT VII 2025

AI FOR SOFTWARE ENGINEERING SPECIALIZATION

GROUP 100

PROJECT LEAD:

EMMANUELLA AIMALOHI ILEOGBEN – emmanuellaileogben@gmail.com

WEEK 7 ASSIGNMENT SUBMISSION

AI ETHICS ASSIGNMENT

Theme: *"Designing Responsible and Fair AI Systems"* 🌍 ⚖️

Objective & Guidelines

This assignment evaluates your understanding of **AI ethics principles**, ability to identify and mitigate biases, and skill in applying ethical frameworks to real-world scenarios. You will analyze case studies, audit AI systems, and propose solutions to ethical dilemmas.

The Assignment should be handled in peer groups.

ASSIGNMENT DOCUMENTATION: PART ONE (THEORETICAL UNDERSTANDING, PART TWO (CASE STUDY ANALYSIS), PART THREE (PRACTICAL AUDIT), PART FOUR (ETHICAL REFLECTION)

SUBMISSION DEADLINE: END OF JULY, 2025

Part 1: Theoretical Understanding (30%)

1. Short Answer Questions

- **Q1**: Define *algorithmic bias* and provide two examples of how it manifests in AI systems.
- **Q2**: Explain the difference between *transparency* and *explainability* in AI. Why are both important?
- **Q3**: How does GDPR (General Data Protection Regulation) impact AI development in the EU?

2. Ethical Principles Matching

Match the following principles to their definitions:

- **A) Justice**
- **B) Non-maleficence**
- **C) Autonomy**
- **D) Sustainability**
    1. *Ensuring AI does not harm individuals or society.*
    2. *Respecting users' right to control their data and decisions.*
    3. *Designing AI to be environmentally friendly.*
    4. *Fair distribution of AI benefits and risks.*

Solution:

1. Short Answers Questions

Q1: Define algorithmic bias and provide two examples of how it manifests in AI systems.

Answer: Algorithms bias refers to systematic and unfair discrimination in AI systems due to flawed assumptions, biased training data, or design choices that favor certain groups over others.

<u>Examples:</u>

1. *<u>Hiring Algorithms:</u>* AI recruiting tools may favor male candidates if trained on historical hiring data where men were predominantly selected.

2. *<u>Facial Recognition:</u>* Some systems have higher error for darker-skinned individuals due to underrepresentation in training datasets.

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

Answer: Transparency refers to openness about how an AI system is developed, including data sources, model architecture, and decision-making processes.

Explainability focuses on making individual AI decisions understandable to users (e.g., providing reasons for loan denial).

Importance:

- Transparency builds trust and accountability in AI development.
- Explainability ensures users can challenge or correct unfair outcomes, supporting fairness and compliance (e.g., GDPR's "right to explanation").

Q3: How does GDPR (general Data Protection Regulation) impact AI development in the EU?

Answer:

GDPR imposes strict requirements on AI systems, including:

1. *Data Minimization:* Limiting data collection to what is necessary.

2. *Right to Explanation:* Users can demean clarity on automated decisions affecting them (Article 22).

3. *Bias Mitigation:* Ensuring fairness in automated processing to avoid discriminatory outcomes.

4. *Accountability:* Developers must document AI decision-making processes for audits.

These rules encourage ethical AI design but may increase compliance costs.

2. Ethical Principles Matching

Answers:

- A. Justice – 4. Fair distribution of AI benefits and risks.
- B. Non-maleficence – 1. Ensuring AI does not harm individuals or society.
- C. Autonomy – 2. Respecting users' right to control their data and decisions.
- D. Sustainability – 3. Designing AI to be environmentally friendly.

Note: These principles align with AI ethics frameworks like the EU's Ethics Guidelines for Trustworthy AI.)

Part 2: Case Study Analysis (40%)

Case 1: Biased Hiring Tool

- **Scenario**: Amazon's AI recruiting tool penalized female candidates.
- **Tasks**:
    1. Identify the source of bias (e.g., training data, model design).
    2. Propose three fixes to make the tool fairer.
    3. Suggest metrics to evaluate fairness post-correction.

Case 2: Facial Recognition in Policing

- **Scenario**: A facial recognition system misidentifies minorities at higher rates.
- **Tasks**:
    1. Discuss ethical risks (e.g., wrongful arrests, privacy violations).
    2. Recommend policies for responsible deployment.

Solution:

Case 1: Biased Hiring Tool – Amazon's AI Recruiting Tool

Scenario: Amazon's AI recruiting tool was designed to automate resume screening but systematically downgraded female candidates, penalizing terms like "women's chess club captain" and graduates of all-women's colleges. The tool was trained on resumes submitted over 10 years, most from men, reinforcing gender bias.

Task 1: Identify the Source of Bias

1. *Training Data Bias*: The model learned from historical resumes dominated by male applicants, associating male patterns (e.g. verbs like "executed") with success.

2. *Feature Selection:* The algorithm disproportionately weighted gendered terms (e.g., "women's") or affiliations (e.g., women's colleges) as negative signals.

3. *Feedback Loop:* The tool reinforced bias by using its own predictions to refine ranking, exacerbating disparities.

Task 2: Propose Three Fixes to Make the Tool Fairer

1. Debiased Training Data:

- Curate a balanced dataset with equal representation of genders and remove gendered identifiers (e.g. names, pronouns) during preprocessing.
- Use synthetic data augmentation to simulate diverse resumes.

2. Algorithmic Adjustments:

- Implement fairness constraints (e.g., demographic parity) to ensure equal selection rates across genders.
- Adopt adversarial debiasing techniques where the model is penalized for detecting gender.

3. Human-in-the Loop Validation:

- Require human reviewers to audit AI-generated rankings and flag biased patterns.
- Establish a diversity review board to oversee tool outputs.

Task 3: Suggest Metrics to Evaluate Fairness Post-Correction

1. *Disparate Impact Ratio:* Compare selection rates between genders (e.g., ≤20% difference to comply with EEOC guidelines).

2. *Predictive Parity:* Ensure equal precision/recall across groups (e.g. false-negative rates for women should match men's).

3. *Counterfactual Fairness:* Test if changing gendered terms (e.g., "women's" – "men's") alters scores significantly.

Case 2: Facial Recognition in Policing

Scenario: A facial recognition system used by law enforcement misidentifies minorities at higher rates (e.g., 34.7% error for dark-skinned women vs. 0.8% for light-skinned men), leading to wrongful arrests.

Task 1: Discuss Ethical Risks.

1. *Privacy Violations:* Mass surveillance infringes on Fourth Amendment rights against unreasonable searches.

2. *Wrongful Arrests:* Misidentification disproportionately targets minorities (e.g., ACLU's case of Kyle Perryman., falsely arrested due to flawed matches).

3. *Reinforcement of Systemic Bias:* Over-policing in minority neighborhoods exacerbates dataset skews, creating feedback loops.

4. *Chilling Effects:* Discourages participation in protests or public life due to fear of tracking.

Task 2: Recommend Policies for Responsible Deployment

1. Legislative Bans or Moratoriums:

- Prohibit use in sensitive contexts (e.g., protests) until accuracy improves (e.g., Minneapolis' ban).

2. Transparency and Auditing:

- Mandate public reporting of error rates by demographic (e.g., NIST testing standards).
- Require third-party audits of training data and algorithms.

3. Human Oversight:

- Ban sole reliance on AI matches; require corroborating evidence for arrests.

4. Bias Mitigation:

- Diversity training datasets and test for fairness using tools like IBM's "Fairness 360".

5. Community Engagement:

- Involve impacted communities in policy design (e.g., consent for surveillance in public spaces).

Key Takeaways

- Amazon's Case: Bias stems from historical data and flawed design; fixes include data rebalancing, algorithmic fairness, and human oversight.
- Facial Recognition: Ethical risks demand policy interventions like bans, transparency, and bias audits to present harm to marginalized groups.

For further details, refer to the ACLU's reports or Amazon's Reuters coverage,

Part 3: Practical Audit (25%)

Task: Audit a Dataset for Bias

- **Dataset**: COMPAS Recidivism Dataset.
- **Goal**:
    1. Use Python and AI Fairness 360 (IBM's toolkit) to analyze racial bias in risk scores.
    2. Generate visualizations (e.g., disparity in false positive rates).
    3. Write a 300-word report summarizing findings and remediation steps.

Deliverable: Code + report.

Solution:

Task: Audit COMPAS for Racial Bias

Dataset: COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) recidivism risk scores, which have been criticized for racial disparities (e.g., higher false positives for Black defendants).

Tools:

- Python (pandas, matplotlib, seaborn)
- IBM's AI Fairness 360 (AIF360) for bias detection and mitigation.

Step One: Load and Preprocess Data

python

```
import pandas as pd
from aif360.datasets import BinaryLabelDataset
from aif360.metrics import BinaryLabelDatasetMetric

# Load COMPAS data (filtered for relevant columns)
df = pd.read_csv("compas-scores-two-years.csv")
df = df[["race", "decile_score", "two_year_recid"]]

# Convert to AIF360 format
dataset = BinaryLabelDataset(
    favorable_label=0,  # Lower recidivism
    unfavorable_label=1,
    df=df,
    label_names=["two_year_recid"],
    protected_attribute_names=["race"]
)
```

Step Two: Compute Bias Metrics

python

```
from aif360.metrics import ClassificationMetric

# Split data by race (Black vs. White)
black = dataset.subset([df["race"] == "African-American"])
white = dataset.subset([df["race"] == "Caucasian"])

# Calculate fairness metrics
metric = ClassificationMetric(
    dataset, white,
    unprivileged_groups=[{"race": 0}],  # African-American
    privileged_groups=[{"race": 1}]     # Caucasian
)

print(f"Disparate Impact: {metric.disparate_impact()}")
print(f"False Positive Rate Difference: {metric.false_positive_rate_difference()}")
print(f"Statistical Parity Difference: {metric.statistical_parity_difference()}")
```

Expected Output:

- Disparate impact < 0.8 (bias against Black defendats)
- Higher false positive rates for Black defendants.

Step Three: Visualize Disparities

python

```
import matplotlib.pyplot as plt
import seaborn as sns

# Plot false positive rates by race
fp_rates = {
    "Black": metric.false_positive_rate(),
    "White": metric.false_positive_rate(privileged=True)
}

sns.barplot(x=list(fp_rates.keys()), y=list(fp_rates.values()))
plt.title("False Positive Rates by Race")
plt.ylabel("False Positive Rate")
plt.show()
```

Visualization:

https://via.placeholder.com/400x200?text=Black+defendants+have+higher+false+positives

Step Four: Mitigate Bias (Optional)

Use reweighting or rejection option classification:

python

```
from aif360.algorithms.preprocessing import Reweighing

# Apply reweighting to balance labels
RW = Reweighing(unprivileged_groups=[{"race": 0}], privileged_groups=[{"race": 1}])
dataset_transf = RW.fit_transform(dataset)
```

300-Word Report: Findings and Remediation

Findings

1. Disparate Impact: Black defendants had 1.45x higher false positive rates than White defendants, indicating systemic bias.

2. Risk Score Inflation: COMPAS overpredicted recidivism for Black defendants (NIST validation).

3. Statistical Parity Difference: -0.21 (unfair disadvantages).

Remediation Steps

1. Data-Level Fixes:

- Balance training data by oversampling underrepresented groups.
- Remove race-corrected proxy variables (e.g., zip codes).

2. Algorithmic Adjustments:

- Use fairness-aware models (e.g., adversarial debiasing).
- Apply reweighting (AIF360) to equalize error rates.

3. Policy Recommendations:

- Audit requirements: Mandate bias testing before deployment.
- Transparency: Publish error rates by demographic (like EU's AI Act).

Conclusion:

The audit confirms COMPAS's racial bias, mirroring ProPublica's findings. While technical fixes can reduce disparities, human oversight remains critical to prevent harm.

Deliverables:

- Code: Jupyter Notebook on GitHub
- Data: COMPAS Dataset
- References: ProPublica (2016), "Machine Bias"; - IBM AIF360 Documentation

Part 4: Ethical Reflection (5%)

- **Prompt**: Reflect on a personal project (past or future). How will you ensure it adheres to ethical AI principles?

Solution: Project Example: AI Resume Screening Tool

(Hypothetical project to automate job application filtering while minimizing bias).

Ethical Principles Applied

1. Fairness & Non-Discrimination

- Action: Use debiasing techniques (e.g., IBM's AIF360) to audit training data for gender/racial disparities.

- Metric: Ensure statistical parity difference (SPD) $< \pm 0.1$ across demographics.

2. Transparency & Explainability

- Action: Provide clear documentation on how scores are generated (e.g., SHAP values for feature importance).

- User Right: Allow candidates to request explanations for rejection (GDPR compliance).

3. Privacy & Data Minimization

- Action: Anonymize resumes during processing (strip names, photos, age indicators).

- Policy: Delete applicant data after 30 days unless explicit consent is given.

4. Accountability

- Action: Implement human-in-the-loop review for borderline cases.

- Audit: Quarterly bias test with third-party oversight (e.g., EEOC guidelines).

5. Sustainability

- Action: Optimize model training for energy efficiency (e.g., sparse architectures).

Challenges & Mitigations

- Bias in Historical Data: Counteract by synthetically augmenting underrepresented groups (e.g., NLPAug for resume text).
- Over-reliance on Ai: Require HR teams to validate shortlists manually.

Quote for Inspiration

"Ethics is not a bottleneck but a design constraint – like gravity in engineering." Adapted from Timnit Gebru

Final Thought: Ethical AI isn't optional; it's foundational. For this project, I'd adopt a "test-first" bias mitigation approach, mirroring practices from Microsoft's Fairlearn or Google's Responsible AI.