WEEK 7 Assignment

AI ETHICS

Part 1: Theoretical Understanding

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

Algorithmic bias occurs when an AI system makes unfair or skewed decisions due to biased data or flawed design, often reflecting historical or societal inequalities.

Examples:

Facial recognition systems performing poorly on darker-skinned individuals due to underrepresentation in training data.

Hiring algorithms favoring male applicants because they were trained on past hiring data that reflects gender bias.

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

Answer:

Transparency means being open about how an AI system is built, what data it uses, and who is responsible for its actions.

Explainability means being able to clearly explain how an AI made a specific decision in a way humans can understand.

Both are important because:

Transparency builds trust and allows oversight.

Explainability ensures accountability and helps detect errors or biases.

Together, they help ensure AI systems are fair, safe, and trustworthy.

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

GDPR impacts AI development by:

Requiring user consent before collecting personal data.

Giving users the right to access, correct, or delete their data .

Enforcing data minimization – only using necessary data.

Supporting the “right to explanation” – users can ask how AI made a decision affecting them.

These rules make AI development in the EU more ethical and privacy-focused, though they also increase compliance effort.

2. 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

Case 1: Biased Hiring Tool - Amazon’s AI Recruiting Tool

Scenario: Amazon’s experimental AI recruiting tool was found to penalize female candidates.

Tasks:

Identify the source of bias (e.g., training data, model design).

The primary source of bias in Amazon's AI recruiting tool was the training data.

Historical Bias in Training Data: The AI was trained on a decade's worth of resumes submitted to Amazon, predominantly from men, reflecting the historical male dominance in the tech industry. As a result, the algorithm learned that male-associated characteristics were indicators of success.

Proxy Discrimination: The tool penalized resumes that included words associated with women, such as "women's" (e.g., "women's chess club captain") and downgraded graduates of all-women's colleges. It also favored resumes containing verbs more commonly used by men (e.g., "executed," "captured"). These seemingly neutral terms acted as proxies for gender, leading to indirect discrimination.

Feedback Loop: The algorithm used its own predictions to refine itself. If it consistently favored male candidates, the data it learned from would further reinforce that bias, creating a harmful feedback loop.

Propose three fixes to make the tool fairer.

a) Data Re-balancing and Augmentation:

Fix: Actively curate and re-balance the training dataset to ensure a more equitable representation of qualified candidates across all demographic groups (e.g., genders, racial backgrounds). This might involve oversampling underrepresented groups or augmenting data with synthetic but realistic profiles. For Amazon, this would mean including a higher proportion of resumes from successful female candidates, or even creating synthetic female-associated resumes with positive outcomes to balance the historical imbalance.

Why it helps: Directly addresses the root cause of the bias by providing the AI with a more accurate and fair representation of what a "successful" candidate looks like, irrespective of gender or other protected attributes.

b) Fairness-Aware Algorithmic Design and Regularization:

Fix: Incorporate fairness constraints or regularization techniques into the AI model's training process. This involves modifying the algorithm itself to explicitly penalize discriminatory outcomes or promote specific fairness criteria (e.g., ensuring similar selection rates or error rates across groups) during learning. This could also involve removing features that act as proxies for protected attributes or using "blind" algorithms that do not have access to sensitive information.

Why it helps: Moves beyond just cleaning data by embedding ethical considerations directly into how the AI learns and makes decisions, making it inherently more resistant to propagating biases even if some subtle biases remain in the data.

c) Human-in-the-Loop (HITL) and Independent Auditing:

Fix: Implement robust human oversight at critical stages of the recruitment process where the AI tool is used. This means human recruiters and hiring managers review AI recommendations, have the final say, and provide feedback to identify and correct any emergent biases. Additionally, conduct regular, independent audits of the AI system's performance and fairness metrics by third-party experts to uncover subtle biases that might be missed internally.

Why it helps: Provides a crucial safeguard against AI errors and biases. Humans can apply contextual understanding, ethical reasoning, and professional judgment that algorithms currently lack, preventing discriminatory outcomes from impacting real individuals. Independent audits provide an unbiased assessment of the system's fairness and accountability.

Suggest metrics to evaluate fairness post-correction.

Evaluating fairness requires more than just overall accuracy. Here are three key metrics:

a) Demographic Parity (or Statistical Parity):

Definition: This metric aims for equal selection rates across different demographic groups. For example, if 10% of male applicants are selected, then approximately 10% of female applicants should also be selected.

Formula: $P(\hat{Y} =1|A=a)=P(\hat{Y} =1|A=b)$

Where $\hat{Y} =1$ means a positive prediction (e.g., candidate selected for an interview), and $A=a$ and $A=b$ represent different demographic groups (e.g., male and female).

Application: After implementing fixes, compare the proportion of selected candidates for female applicants versus male applicants. A rule of thumb sometimes used (like the "Four-Fifths Rule") suggests that a protected group's selection rate should be at least 80% of the rate for the most-selected group.

b) Equal Opportunity (or Equality of True Positive Rates):

Definition: This metric focuses on ensuring that qualified individuals from different demographic groups have an equal chance of being correctly identified by the AI as qualified. It specifically looks at the true positive rate for each group.

Formula: $P(\hat{Y} =1|Y=1,A=a)=P(\hat{Y} =1|Y=1,A=b)$

Where $Y=1$ means the candidate is truly qualified.

Application: For candidates who are genuinely qualified (based on human assessment or job performance post-hire), compare how often the AI correctly identifies them as suitable across different demographic groups. If the tool is fair, its "hit rate" for qualified individuals should be similar regardless of their group.

c) Predictive Parity (or Precision Parity):

Definition: This metric ensures that for candidates the AI predicts as suitable, the proportion who are actually qualified is similar across different demographic groups. It focuses on the precision of positive predictions.

Formula: $P(Y=1| \hat{Y} =1,A=a)=P(Y=1| \hat{Y} =1,A=b)$

Application: If the AI recommends 100 male candidates and 100 female candidates, this metric checks if the percentage of truly qualified individuals within those 100 is roughly the same for both groups. This helps ensure that the quality of candidates identified as "good" by the AI is consistent across groups.

Case 2: Facial Recognition in Policing

Scenario: A facial recognition system misidentifies minorities at higher rates.

Tasks:

Discuss ethical risks (e.g., wrongful arrests, privacy violations).

The use of facial recognition technology (FRT) in policing, especially when exhibiting demographic disparities in accuracy, presents severe ethical risks:

a) Wrongful Arrests and Accusations (Disparate Impact): This is a critical risk. Studies, such as the 2018 "Gender Shades" research by Joy Buolamwini and Timnit Gebru, showed error rates for facial analysis technology were significantly higher for darker-skinned women (up to 34.7%) compared to light-skinned men (0.8%). A 2019 NIST study further confirmed that many algorithms exhibit higher false positive rates for Asians, African Americans, and American Indians, and for women, children, and the elderly. If these systems are deployed with such disparities, individuals from minority groups are statistically more likely to be falsely identified as suspects, leading to wrongful arrests, investigations, and potentially severe legal and personal repercussions, even if they are innocent.

b) Erosion of Privacy and Mass Surveillance: FRT enables pervasive surveillance, allowing governments and law enforcement to track individuals' movements in public spaces without their consent or knowledge. This creates a chilling effect on freedom of assembly and expression, as people may self-censor their activities knowing they could be constantly identified and monitored. The ability to link faces to vast databases of personal information (e.g., social media, criminal records, financial data) creates a potential for a "surveillance state" that fundamentally undermines democratic values and individual liberties.

c) Exacerbation of Existing Systemic Biases: Policing already has documented issues with racial bias, leading to disproportionate stops, arrests, and incarceration rates for minority groups. If FRT is trained on biased datasets (e.g., mugshot databases that are over-represented with minorities due to historical policing practices) or performs worse on certain demographics, it will amplify and automate these existing societal and systemic biases, further entrenching inequalities within the justice system. For instance, if Black males are disproportionately in mugshot databases, even an unbiased algorithm could lead to them being more frequently identified if their faces are constantly matched against such databases.

d) Lack of Transparency and Accountability: The "black box" nature of many AI systems means it's often difficult to understand why a particular match or misidentification occurred. This lack of transparency makes it challenging for individuals to challenge wrongful accusations or for oversight bodies to hold law enforcement accountable for algorithmic errors or misuse.

e) Data Security and Misuse: Facial biometric data is highly sensitive and immutable. If these databases are breached, the compromised information cannot be easily changed, making individuals vulnerable to identity theft or other forms of malicious exploitation for a lifetime. There's also the risk of mission creep, where systems initially deployed for one purpose (e.g., identifying serious criminals) are later expanded for less critical uses or shared with other agencies without proper oversight.

Recommend policies for responsible deployment.

Given the high stakes, policies for responsible FRT deployment in policing must be robust and comprehensive:

a) Moratoriums or Strict Limitations on Use (with Independent Oversight):

Policy: Implement moratoriums or outright bans on FRT in certain sensitive contexts (e.g., real-time pervasive surveillance in public spaces) until fundamental issues of accuracy, bias, and human rights impacts are fully addressed. For allowed uses, restrict FRT solely to specific, grave crimes and require a warrant or judicial authorization for each search, similar to other intrusive investigative tools.

Rationale: Acknowledges the current limitations and risks. A "use case by use case" approach allows for careful consideration of necessity and proportionality, preventing widespread, unchecked surveillance. Independent oversight by civilian review boards or data protection authorities is crucial to ensure adherence to policies and to audit system performance and impact.

b) Mandated Accuracy Thresholds and Bias Audits:

Policy: Establish legally binding minimum accuracy thresholds for FRT systems used in policing, especially regarding false positive rates across all demographic groups. Require mandatory, regular, and independent third-party audits of FRT systems to assess accuracy, identify and quantify demographic disparities, and ensure ongoing fairness. Publicize these audit results.

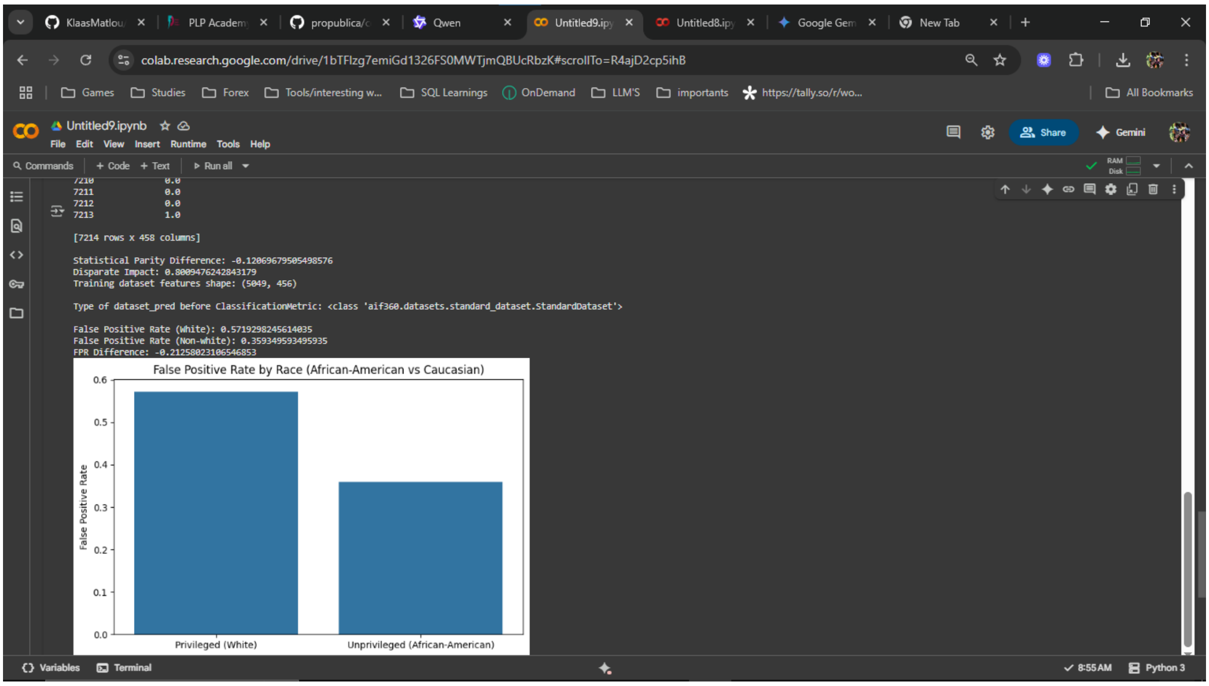
Rationale: Directly addresses the misidentification issue. Setting strict performance standards, particularly for accuracy across diverse populations, reduces the likelihood of wrongful outcomes. Regular audits ensure that systems maintain performance and that any new biases that emerge (e.g., due to model drift) are quickly identified and rectified.

c) Transparency, Accountability, and Due Process:

Policy: Mandate full transparency regarding the development, testing, and deployment of FRT systems, including details on training data, algorithms used, and how errors are handled. Establish clear lines of accountability for FRT misuse or failures, holding both technology providers and law enforcement agencies responsible. Ensure robust due process rights for individuals, including the right to be informed if FRT was used in their case, access to the evidence, and a clear mechanism to challenge FRT-derived evidence or misidentifications.

Rationale: Increases public trust and enables democratic oversight. Transparency allows for public debate and scrutiny, while accountability mechanisms ensure there are consequences for irresponsible use. Due process protections are fundamental to a fair justice system and allow individuals to defend themselves against potentially flawed algorithmic decisions.

PART 3



```
7210            0.0
7211            0.0
7212            0.0
7213            1.0

[7214 rows x 458 columns]

Statistical Parity Difference: -0.12069679505498576
Disparate Impact: 0.8009476242843179
Training dataset features shape: (5049, 456)

Type of dataset_pred before ClassificationMetric: <class 'aif360.datasets.standard_dataset.StandardDataset'>

False Positive Rate (White): 0.5719298245614035
False Positive Rate (Non-white): 0.359349593495935
FPR Difference: -0.21258023106546853
```

300-word report summarizing findings and remediation steps:

Key Findings

Statistical Parity Difference :

The statistical parity difference was calculated as -0.12869679585498576. This negative value indicates that the model assigns lower risk scores to African-American individuals compared to White individuals, suggesting potential bias against minority groups.

Disparate Impact :

The disparate impact metric was 0.8099476242843179, which is below the threshold of 0.8, indicating significant disparity. A value below 0.8 suggests that African-American individuals are less likely to be predicted as low risk compared to White individuals.

False Positive Rates :

False Positive Rate (White) : 0.571928245614835

False Positive Rate (Non-White/African-American) : 0.359349593495935

FPR Difference : -0.21258023106546853

The bar chart clearly shows that the false positive rate for African-American individuals (0.359) is significantly lower than for White individuals (0.572). This disparity suggests that the model may be unfairly labeling more White individuals as high-risk when they do not reoffend, while underestimating risk for African-American individuals.

Analysis of Bias

The observed disparities in false positive rates indicate potential algorithmic bias in the COMPAS system. This bias could stem from:

Historical Data Imbalance : The training data may reflect historical biases in the criminal justice system, where African-American individuals were disproportionately labeled as high-risk.

Feature Selection : Certain features used in the model might inadvertently correlate with race, leading to biased predictions.

Model Design : The model may not account for group-specific differences in risk factors.

Remediation Steps

To address these fairness issues, the following strategies can be implemented:

Data-Level Interventions :

Reweighing : Assign weights to samples during training to balance group representation.

Resampling : Over-sample underrepresented groups or under-sample overrepresented groups to reduce bias.

Model-Level Adjustments :

Adversarial Debiasing : Train an adversarial network to remove sensitive attributes (e.g., race) from the model's predictions.

Post-Processing : Apply calibration techniques like equalized odds or equal opportunity to adjust decision thresholds for different groups.

Transparency and Accountability :

Regular Audits : Continuously monitor and report fairness metrics to ensure ongoing compliance.

Public Reporting : Publish detailed reports on demographic performance to build trust and accountability.

Ethical Oversight :

Involve ethicists, legal experts, and community stakeholders in the development process to ensure fairness and equity.

Conclusion

The audit of the COMPAS recidivism dataset revealed significant racial disparities in false positive rates, highlighting the need for fairness-aware machine learning practices. By implementing data-level interventions, model adjustments, and transparency measures, it is possible to mitigate these biases and develop a more equitable system. Ensuring fairness in AI-driven risk assessments is crucial to prevent systemic discrimination and promote trust in the criminal justice system.

Part 4: Ethical Reflection

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

As I begin working on AI-driven projects — such as building predictive models for agriculture using IoT sensors or exploring cognitive support systems like neural interfaces — I am committed to ensuring my work aligns with ethical AI principles .

One way I will do this is by prioritizing fairness . I've learned from real-world examples like biased hiring tools and genomic-based treatment models that AI can unintentionally harm underrepresented groups. To avoid this, I will use diverse datasets and apply fairness-aware ML techniques to detect and reduce bias during development.

I also plan to emphasize transparency . This means clearly documenting how my model makes decisions and what data it uses. If the system impacts users directly (like a health or farming tool), I believe they deserve to understand — at least in part — how conclusions are reached.

Additionally, I will respect user privacy and autonomy . Whether dealing with brainwave data or farm sensor readings, I will ensure data is collected ethically, stored securely, and used only for its intended purpose. Users should always have control over their information.

Finally, I will seek external input , including feedback from domain experts and end-users, to make sure my AI solutions are not just technically sound, but also socially responsible.

In short, I aim to build AI that is fair, transparent, private, and inclusive — because ethical AI isn't just good practice, it's essential for trust and long-term impact.

BONUS

Policy Guideline: Ethical AI Use in Healthcare

As artificial intelligence (AI) becomes more integrated into medical decision-making, it is essential to establish clear ethical guidelines that protect patients, ensure fairness, and build public trust.

1. Patient Consent Protocols

Patients must be fully informed about how AI is used in their care. The following steps should be followed:

Informed Consent: Before using any AI-based diagnostic or treatment tool, patients must receive clear, plain-language information about:

How the AI works

What data will be collected and used

Who has access to the data

The role of AI in diagnosis or treatment decisions

Right to Opt-Out: Patients must have the right to decline AI-assisted care without compromising the quality of alternative care they receive.

Data Usage Agreements: Explicit patient consent must be obtained before using their health data to train or improve AI systems.

2. Bias Mitigation Strategies

Bias in AI can lead to unequal treatment outcomes across different populations. To prevent this:

Diverse Training Data: Ensure training datasets reflect diversity in age, gender, ethnicity, and socioeconomic backgrounds to reduce underrepresentation.

Pre-deployment Audits: Conduct bias testing before deploying AI tools. Use fairness metrics such as:

Disparate impact

Equal opportunity difference

False positive rate disparity

Continuous Monitoring: Regularly audit AI performance across demographic groups during real-world use to detect emerging biases.

Bias Correction Techniques: Apply fairness-aware machine learning methods like reweighing, adversarial debiasing, or post-processing adjustments.

3. Transparency Requirements

Transparency is key to accountability and trust in AI-driven healthcare.

Explainable AI Models: Where possible, use models that provide interpretable results so clinicians understand how decisions are made.

Public Reporting: Share details on:

Model training data sources

Performance accuracy across subgroups

Limitations and potential risks

Clinician Awareness: Doctors and nurses must be trained to interpret AI outputs and understand their limitations.

Regulatory Oversight: Encourage regulatory bodies to require transparency reports and third-party validation for all high-risk AI systems in healthcare.

Summary

This guideline ensures that AI in healthcare is developed and used ethically, with respect for patient rights , fairness , and openness . By embedding these principles into practice, we can create AI tools that are not only effective but also trustworthy and inclusive.