Untitled1

July 11, 2025

Short Answer Questions Q1: Define algorithmic bias and provide two examples of how it manifests in AI systems. Answer: Algorithmic bias refers to systematic and repeatable errors in a computer system that create unfair outcomes, especially affecting marginalized or underrepresented groups. These biases often stem from the data used to train models or design flaws in the algorithms themselves. Examples: • Hiring Algorithms: An AI trained on historical resumes that favored male applicants may learn to penalize resumes with female-identifying terms, replicating past bias. • Facial Recognition Systems: Studies have shown some systems misidentify individuals from minority groups at higher rates due to imbalanced datasets that underrepresent those populations. Q2: Explain the difference between transparency and explainability in AI. Why are both important? Answer: • Transparency refers to the openness of an AI system's design, data sources, and decision-making processes. It allows stakeholders to see how the system operates. • Explainability is the ability to understand and interpret the decisions made by an AI, often in a human-friendly way. Importance: Both are crucial for trust and accountability. Transparency helps ensure responsible development, while explainability allows users to challenge or validate decisions—especially in sensitive domains like healthcare or criminal justice. Q3: How does GDPR (General Data Protection Regulation) impact AI development in the EU? Answer: GDPR enforces strict data protection rules, significantly influencing how AI systems are built and deployed. Key impacts include: • Consent Requirements: AI systems must obtain explicit consent before processing personal data. • Right to Explanation: Individuals can request an explanation of decisions made by automated systems. • Data Minimization & Privacy: Developers must ensure data is used minimally and securely, promoting privacy-by-design in AI architectures. Ethical Principles Matching Principle Definition 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. When applied in real-world development, these principles help shape AI that is equitable, safe, respectful, and future-conscious. Case 1: Amazon's AI Recruiting Tool Penalized Female Candidates 1. Source of Bias The bias originated from: • Training Data: The model was trained on historical resumes, predominantly from male applicants, reinforcing existing gender biases. • Feature Selection: Terms or qualifications more commonly used by women may have been undervalued. • Model Design: Lack of fairness constraints allowed biased patterns to influence decision-making. 2. Proposed Fixes To mitigate bias: • Rebalance Training Data: Include diverse, representative samples of qualified candidates across gender identities. • Apply Fairness Constraints: Integrate algorithms that detect and reduce bias during training (e.g., reweighting or adversarial debiasing). • Audit Feature Importance: Remove or revise features that correlate with gender but aren't relevant to job qualifications. 3. Fairness Metrics Useful post-correction metrics: • Demographic Parity: Check if selection rates are equal across gender groups. • Equal Opportunity: Compare true positive rates for male and female candidates. • Disparate Impact Ratio: Evaluate ratio of positive outcomes between groups—ideal range is typically between 0.8 and 1.25. Case 2: Facial Recognition System Misidentifies Minorities 1. Ethical Risks Facial recognition systems, especially when deployed in policing, raise significant ethical concerns: • Wrongful Arrests: Misidentifications can lead to innocent people being detained or prosecuted. • Privacy Violations: Continuous surveillance can infringe on personal privacy and autonomy. • Bias Amplification: Existing societal biases may be reinforced through algorithmic errors, disproportionately affecting marginalized communities. • Loss of Public Trust: Inaccurate and biased use undermines confidence in both law enforcement and AI technology. 2. Recommended Policies for Responsible Deployment To ensure ethical use: • Bias Audits: Mandate regular, third-party evaluations to assess accuracy across demographic groups. • Human Oversight: Require human verification before taking any action based on AI predictions. • Transparency Reports: Law enforcement agencies should publish performance metrics, use cases, and error rates. • Public Consultation & Consent: Engage communities to understand concerns and preferences before implementation. • Restricted Use Cases: Limit usage to high-priority tasks with appropriate safeguards (e.g., identifying missing persons, not mass surveillance). Task: Audit a Dataset for Bias using AI Fairness 360 Dataset: COMPAS Recidivism Dataset Objective: Analyze racial bias in predicted risk scores and recommend fairer approaches. Python Code Setup (AI Fairness 360 Toolkit) python

[ ]:

```python
# Install and import necessary packages
!pip install aif360
import pandas as pd
from aif360.datasets import CompasDataset
from aif360.metrics import BinaryLabelDatasetMetric, ClassificationMetric
from aif360.explainers.metrics import MetricTextExplainer
import matplotlib.pyplot as plt

# Load COMPAS dataset
dataset = CompasDataset()

# Measure dataset bias
metric = BinaryLabelDatasetMetric(dataset, privileged_groups=[{'race':␣
 ↪'Caucasian'}],
                                  unprivileged_groups=[{'race':␣
 ↪'African-American'}])

# Visualize disparate impact
print("Disparate Impact:", metric.disparate_impact())
print("Mean Difference:", metric.mean_difference())

# (Optional) Visualization for false positive rates
fig, ax = plt.subplots()
ax.bar(['Caucasian', 'African-American'],
       [metric.base_rate(privileged=True), metric.base_rate(privileged=False)])
ax.set_title('Base Rates by Race')
plt.show()
```

Bias Audit Summary Report (300 words) The COMPAS dataset was audited using IBM's AI Fairness 360 toolkit to uncover racial bias in criminal risk assessments. The analysis compared base rates and disparate impact between privileged (Caucasian) and unprivileged (African-American) groups.

Results showed a disparate impact ratio below the acceptable threshold of 0.8, suggesting that African-American individuals were disproportionately labeled as high-risk compared to Caucasian peers. Mean differences further highlighted this imbalance, raising concerns about fairness and ethical integrity. To remediate bias, several steps are recommended: • Preprocessing Techniques: Reweigh instances or resample data to ensure equitable representation. • Algorithmic Fairness Constraints: Implement debiasing algorithms like adversarial debiasing or reject option classification. • Post-processing Evaluation: Adjust model outputs to align with fairness metrics such as equal opportunity and predictive parity. Additionally, continuous bias monitoring and transparent reporting must be standard practice in deploying such systems, especially in high-stakes domains like criminal justice. Ultimately, bias audits are critical to protecting vulnerable populations, upholding legal fairness, and restoring trust in AI-assisted decision-making. Prompt: Reflect on a personal project (past or future). How will you ensure it adheres to ethical AI principles? For my reflection, I'll imagine I'm building a chatbot to assist with mental health support — empathetic, resourceful, but not a substitute for therapy. To ensure ethical AI principles are embedded in its development: • Autonomy: I'll design the bot to respect user consent at every stage. It won't record sensitive information unless users explicitly opt in, and they'll always be informed about what data is collected and why. • Non-maleficence: Guardrails will be built to avoid harmful responses. If a user expresses distress, the bot will gently redirect to professional resources without judgment. • Justice: I'll train the bot with diverse datasets representing various cultural, social, and emotional expressions to avoid skewed responses and to reflect inclusivity. • Transparency & Explainability: Users will have access to a clear FAQ or explainer on how the bot works, what it can do, and its limitations. If it makes a recommendation, it'll provide reasoning in plain language. • Sustainability: The deployment infrastructure will prioritize energy-efficient practices—possibly running on servers powered by renewable energy and minimizing resource-intensive processes. Ultimately, ethical design isn't a checklist—it's a mindset woven into every code commit and design decision. It's about humility, foresight, and responsibility in the face of complexity. 1-Page Guideline for Ethical AI Use in Healthcare Title: "Compassionate Intelligence: Ethical AI Guidelines for Healthcare" 1. Patient Consent Protocols • Ensure informed, explicit consent is obtained before collecting or analyzing patient data. • Use plain language when explaining AI involvement in diagnostics or treatment. • Enable patients to opt out of AI-based decisions without prejudice or penalty. 2. Bias Mitigation Strategies • Conduct regular bias audits across demographic categories to uncover disparities. • Use diverse training datasets reflecting age, race, gender, and health conditions. • Integrate fairness-aware algorithms, such as adversarial debiasing or reweighing. • Prioritize feedback from underserved communities during model evaluation and updates. 3. Transparency Requirements • Provide clear documentation on how the AI system functions, its limitations, and decision rationale. • Disclose the source of training data, model accuracy, and known error rates. • Maintain an open channel for third-party review and real-time monitoring of critical decisions. • Use explainable AI methods to ensure clinicians and patients understand the outcomes.