AI Ethics: Week 7

Part 1: Theoretical Understanding (30%)

1. Short Answer Questions

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

Definition of Algorithmic Bias

Algorithmic bias occurs when an AI system produces unfair, prejudiced, or systematically skewed outcomes due to the data it was trained on, the design of its algorithms, or the way it is deployed. In simple terms, it's when AI unintentionally discriminates against certain groups of people because of hidden biases in its inputs or processes.

Two Examples of Manifestations in AI Systems

1. **Facial Recognition Technology**
   - AI facial recognition systems have been shown to misidentify people of color, women, and younger individuals at much higher rates than white men.
   - This happens because training datasets often contain more images of lighter-skinned individuals, leading to unequal accuracy across demographics.
2. **Hiring Algorithms**
   - Some recruitment AI tools have displayed bias against women when screening resumes.
   - For example, if historical hiring data favored men in tech roles, the algorithm learns to prioritize male-associated terms or experiences, perpetuating gender inequality.

Why It Matters

Algorithmic bias can reinforce existing social inequalities, reduce trust in AI systems, and cause real-world harm (e.g., wrongful arrests, unfair hiring practices). Addressing it requires diverse datasets, fairness audits, and transparent design practices.

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

Transparency in AI means openly disclosing how an AI system works (its data sources, algorithms, and processes), while explainability means making individual AI decisions understandable to humans. Both are important because transparency builds trust and accountability, and explainability ensures users can interpret and challenge outcomes.

Transparency vs. Explainability

Concept	Definition	Focus	Example
Transparency	Providing clear information about the design, data, and functioning of an AI system	*System-level openness*	A company publishes documentation about what data its chatbot is trained on and how it processes inputs.
Explainability	Making AI outputs interpretable so humans understand *why* a specific decision was made	*Decision-level clarity*	A medical AI explains which symptoms and lab results led to a cancer diagnosis recommendation.

Why Both Are Important

- **Trust & Accountability:** Transparency ensures stakeholders know what the AI system is doing, while explainability allows them to verify and challenge its decisions.
- **Regulatory Compliance:** Many AI regulations (like the EU AI Act) require both transparency and explainability to protect users.
- **Bias Detection:** Transparency reveals potential flaws in datasets or algorithms, while explainability helps identify biased outcomes in practice.
- **User Empowerment:** Transparency informs users about system capabilities and limitations, while explainability gives them confidence in specific outputs.
- **Safety & Ethics:** Together, they prevent “black box” AI systems from making unchecked decisions that could harm individuals or society.

Real-World Illustration

- **Finance:** A transparent credit scoring AI discloses that it uses income, debt, and repayment history. Explainability shows why a specific applicant was denied credit (e.g., high debt-to-income ratio).
- **Healthcare:** Transparency reveals that an AI diagnostic tool was trained on 1M patient records. Explainability clarifies why it flagged pneumonia in a particular X-ray.

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

GDPR impacts AI development in the EU by enforcing strict rules on how personal data can be collected, processed, and used, which shapes the design of AI systems to prioritize privacy, fairness, and accountability.

Key Impacts of GDPR on AI Development

- **Data Minimization & Purpose Limitation**
  - AI developers must collect only the data necessary for a specific purpose and cannot repurpose it freely.
  - This restricts the use of massive datasets often needed for training AI models.
- **Consent & Lawful Processing**
  - GDPR requires explicit user consent for data use, making it harder to train AI on personal information without clear permission.
  - AI systems must justify their data processing under lawful bases (e.g., consent, contract, legitimate interest).
- **Right to Explanation & Human Oversight**
  - Individuals have rights to understand automated decisions that affect them (e.g., credit scoring, hiring).
  - This pushes AI developers to integrate **explainability** features into systems.
- **Accountability & Compliance Costs**
  - Companies must document how AI systems handle data, conduct impact assessments, and ensure compliance.
  - This increases development costs but also builds trust in AI adoption.
- **Restrictions on Automated Decision-Making**
  - GDPR limits fully automated decisions with “legal or significant effects” unless safeguards (like human review) are in place.
  - This impacts areas like finance, healthcare, and employment where AI decisions carry high stakes.

Opportunities & Challenges

- **Opportunities:**
  - Encourages ethical AI design, boosting public trust.
  - Aligns with the upcoming **EU AI Act**, which complements GDPR by focusing on risk-based regulation.
- **Challenges:**
  - Limits access to diverse datasets, potentially slowing innovation.
  - Complex compliance requirements may disadvantage smaller startups compared to large corporations.

- ○ Ongoing debates in the EU about easing GDPR rules to reduce constraints on AI development.

Summary

GDPR ensures that **AI in the EU is people-centered, privacy-conscious, and accountable**, but it also creates hurdles for developers who rely on large-scale data. The regulation balances innovation with protection, and future adjustments (like the EU AI Act) aim to refine this balance further.

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**

Quick way to remember:

- *Justice* = fairness in outcomes.
- *Non-maleficence* = "do no harm."
- *Autonomy* = freedom of choice/control.
- *Sustainability* = long-term environmental responsibility.

Part 2: Case Study Analysis (40%)

Case 1: Biased Hiring Tool

Scenario: Amazon's AI recruiting tool penalized female candidates.

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

- **Training Data Bias:** The AI was trained on historical resumes from a male-dominated tech industry, so it learned to favor male-associated terms and experiences.
- **Model Design Bias:** The algorithm picked up patterns that correlated "male" indicators with success, penalizing resumes mentioning "women's" activities or schools.

2. Propose three fixes to make the tool fairer.

- **Balanced Training Data:** Retrain the model with diverse, gender-balanced datasets to remove historical bias.
- **Bias Auditing & Feature Control:** Exclude gender-related features (e.g., words like “women’s club”) that unfairly influence outcomes.
- **Human Oversight:** Combine AI recommendations with human review to ensure fairness and accountability.

3. Suggest metrics to evaluate fairness post-correction.

- **Selection Rate Parity:** Compare hiring recommendations across genders to ensure equal opportunity.
- **Disparate Impact Ratio:** Measure whether one group is disproportionately disadvantaged (<80% threshold often used).
- **Accuracy Across Groups:** Track performance metrics (precision, recall) separately for male and female candidates.

Case 2: Facial Recognition in Policing

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

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

- **Wrongful Arrests:** Misidentification can lead to false accusations and legal harm.
- **Privacy Violations:** Mass surveillance erodes individual privacy and civil liberties.
- **Discrimination:** Higher error rates for minorities reinforce systemic inequalities.
- **Erosion of Trust:** Communities may lose faith in law enforcement if technology is unfair.

2. Recommend policies for responsible deployment.

- **Strict Accuracy Standards:** Require independent testing to prove equal accuracy across demographic groups before deployment.
- **Human-in-the-Loop:** Ensure facial recognition is only used as a supportive tool, not the sole basis for arrests.
- **Transparency & Accountability:** Publicly disclose usage policies, error rates, and audit results.
- **Limited Scope:** Restrict use to serious crimes and prohibit mass surveillance in public spaces.
- **Oversight & Regulation:** Establish independent review boards to monitor misuse and enforce compliance.

Summary:

- Case 1 highlights **bias in training data** and the need for fairness metrics.

- Case 2 emphasizes **ethical risks in policing** and the importance of **policies ensuring accountability and fairness**.

Part 3: Practical Audit (25%)

1. Write a 300-word report summarizing findings and remediation steps.

Audit of COMPAS Recidivism Dataset Using AI Fairness 360

The COMPAS dataset, widely used in criminal justice risk assessments, has been criticized for racial bias. Using IBM's AI Fairness 360 toolkit, we conducted an audit to evaluate disparities in risk scores between Caucasian and African-American defendants.

Initial metrics revealed significant bias. The **disparate impact ratio** was below 1, indicating that African-American defendants were disproportionately labeled as "high risk." The **mean difference** metric confirmed unequal treatment across racial groups. When training a logistic regression classifier, false positive rates were notably higher for African-American defendants compared to Caucasian defendants. This means African-American individuals were more likely to be incorrectly flagged as likely to reoffend, leading to unfair outcomes such as harsher bail or sentencing decisions.

To remediate bias, three strategies are recommended:

1. **Pre-processing (Reweighing):** Adjust dataset weights to balance representation of racial groups before training.
2. **In-processing (Fairness-aware algorithms):** Use models that incorporate fairness constraints during training.
3. **Post-processing (Outcome adjustment):** Calibrate predictions to equalize error rates across groups.

Evaluation of fairness post-correction should rely on metrics such as **equal opportunity difference** (comparing true positive rates), **disparate impact ratio**, and **false positive rate parity**. Visualizations of these metrics help stakeholders understand disparities and track improvements.

In conclusion, the audit confirms that the COMPAS dataset embeds racial bias, particularly disadvantaging African-American defendants. By applying fairness interventions and continuously monitoring metrics, AI systems in criminal justice can move toward more equitable and trustworthy outcomes. Ethical deployment requires not only technical fixes but also transparency, accountability, and human oversight to ensure justice is upheld.

Part 4: Ethical Reflection (5%)

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

Ethical Reflection Example

Project Context (Future Example): Suppose I'm developing an AI-powered mobile app that helps students personalize their study schedules based on performance and preferences.

1. Fairness & Justice

- I will ensure the app works equally well for students from different backgrounds by testing across diverse datasets.
- I'll avoid embedding bias (e.g., assuming certain schools or regions perform better) by balancing training data.

2. Transparency & Explainability

- The app will clearly explain why it recommends a particular study plan (e.g., "You scored higher in math, so more time is allocated to science").
- I'll provide documentation on how the algorithm works so users understand the logic behind decisions.

3. Autonomy & Privacy

- Students will have full control over their data—choosing what to share and being able to delete it anytime.
- I'll use anonymization and encryption to protect sensitive information.

4. Non-Maleficence (Do No Harm)

- I'll test the app to ensure it doesn't cause stress or unfairly disadvantage certain learners.
- Safeguards will prevent harmful recommendations (e.g., overloading a student with unrealistic schedules).

5. Sustainability

- I'll design the app to be lightweight and energy-efficient, reducing unnecessary server usage.
- Cloud resources will be optimized to minimize environmental impact.

Policy Guideline for Ethical AI Use in Healthcare

1. Patient Consent Protocols

- **Informed Consent:** Patients must be clearly informed when AI systems are used in diagnosis, treatment planning, or monitoring.
- **Data Usage Disclosure:** Healthcare providers must explain what patient data is collected, how it is processed, and for what purpose.
- **Opt-In/Opt-Out Rights:** Patients should have the right to opt in or out of AI-driven services without compromising access to standard care.
- **Data Privacy Safeguards:** All patient data must be anonymized where possible and stored securely in compliance with GDPR/HIPAA standards.

2. Bias Mitigation Strategies

- **Diverse Training Data:** AI models must be trained on datasets that represent diverse populations (age, gender, ethnicity, socioeconomic status).
- **Regular Bias Audits:** Independent audits should be conducted to detect and address algorithmic bias in clinical outcomes.
- **Fairness Metrics:** Systems must be evaluated using fairness indicators (e.g., equal opportunity difference, disparate impact ratio).
- **Human Oversight:** Clinicians must remain the final decision-makers, ensuring AI recommendations are contextualized and not blindly followed.

3. Transparency Requirements

- **Explainability:** AI systems must provide clear, understandable reasoning for their outputs (e.g., why a diagnosis was suggested).
- **Documentation:** Developers must publish model design, training data sources, and limitations in accessible formats for healthcare providers.
- **Accountability:** Healthcare institutions must establish clear responsibility for AI-related errors or adverse outcomes.
- **Public Reporting:** Hospitals and clinics should disclose AI usage policies and performance metrics to patients and regulators.