WEEK 7 ASSIGNMENT
PART 1 - THEORETICAL UNDERSTANDING (30%)

1. Short Answer Questions

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

  Algorithmic bias is systematic and unfair discrimination in AI systems due to flawed data, design choices or unintended learning patterns that favor or disadvantage certain groups of people.

  Examples include:

  a) Hiring algorithms- AI recruitment tools trained on historical hiring data may favor male candidates for technical roles if past hiring was biased.
  b) Facial Recognition: Some facial analysis systems have higher error rates for darker-skinned individuals due to underrepresentation in training data.

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

  Transparency refers to openness about how an AI system is developed, including data sources, model architecture, and decision-making processes while explainability is the ability to interpret and justify an AI's decisions in human-understandable terms (e.g., feature importance in a loan approval model).

  They are both important because: Transparency builds trust and accountability (e.g., ensuring compliance with regulations). Explainability helps users, regulators, and developers debug biases, ensure fairness, and justify decisions in crucial situations such as medical diagnosis with AI.

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

  GDPR imposes strict requirements on AI systems in the EU, including:

  a) Right to explanation where users can demand explanations for automated decisions such as loan rejections.

b) Data minimization where it limits unnecessary data collection therefore affecting training datasets.
c) Bias mitigation- the GDPR requires fairness in automated processing to avoid discriminatory outcomes.
d) Accountability- GDPR insists that developers must document AI decision-making processes for audits.

2. Ethical Principles Matching

Match the following principles to their definitions:

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

PART TWO - 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).
    a) Training data- Historical hiring data favored male candidates (e.g., more men in technical roles at Amazon in the past).
    b) Model design- The AI learned to associate male-associated keywords such as "strong" with higher rankings, while penalizing female-associated terms.
    c) Feedback loop: The model reinforced existing biases by downgrading resumes that differed from male-dominated profiles.

2. Propose three fixes to make the tool fairer.
    a) Debiased training data - Use synthetic data augmentation to balance gender representation. Remove gender proxies such as names, gendered phrases, unrelated extracurriculars.
    b) Fairness-aware model design - Apply adversarial debiasing, where a secondary model penalizes gender-based predictions. Use reweighting techniques to prioritize underrepresented groups.
    c) Human-in-the-loop validation - Implement human oversight for flagged decisions such as low-scoring female candidates. Regular auditing of the model outputs for disparities using fairness metrics should be done.

3. Suggest metrics to evaluate fairness post-correction.
    a) Demographic parity - the percentage of male vs. female candidates shortlisted should be proportional.
    b) Equal opportunity - True positive rates (selection rates for qualified candidates) should be equal across genders.
    c) Predictive parity - Accuracy (e.g., hiring success rate) should not differ significantly by gender.
    d) Disparate impact ratio: (Selection rate for women / Selection rate for men) should be close to 1 (a value < 0.8 may indicate bias).

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).
    a) Wrongful arrests and injustice - False positives disproportionately target minorities, leading to unjust detentions.This includes cases like Robert Williams who was wrongly arrested due to faulty facial recognition. Reinforcement of systemic bias makes marginalized communities to be wrongfully over-policed.
    b) Privacy violations - Mass surveillance erodes civil liberties, especially in public spaces. Privacy is also violated due to the lack of consent when scraping images from social media or CCTV for training data.

c) Chilling effects on the society - Fear of surveillance may deter peaceful protests or free movement for targeted groups.
d) Lack of accountability - Police may over-rely on AI which leads them to ignoring exonerating evidence during trials leading to wrongful judgement of targeted groups.

2. Recommend policies for responsible deployment.
    a) Legal and regulatory safeguards - Ban use in sensitive contexts such that facial recognition should be prohibited for predictive policing or real-time surveillance without judicial oversight. GDPR-like protections should strictly enforce rules like requiring explicit consent for biometric data collection.
    b) Bias mitigation and transparency - Mandatory audits for racial/gender bias should be conducted by humans. Public disclosure of accuracy rates across demographics should also be transparent. Third-party testing before deployment is also important to realise where there is bias and correct it.
    c) Operational restrictions - The facial recognition policing should not be used as sole evidence. The policing department should require corroborating proof such as fingerprints and witness testimony for arrests.
    d) Human review - Officers must verify matches manually.
    e) Community engagement and participation - There should be more than enough public consultations before deployment in a city. Clear appeals process for misidentified individuals and wrongful arrests should be made possible by the public

Part 4: Ethical Reflection - Personal Project Analysis

Project Overview: AI-Powered Student Learning Analytics Platform

Project Description

For my capstone project, I am developing an AI-powered learning analytics platform that analyzes student engagement patterns, learning preferences, and performance data to provide personalized educational recommendations. The system uses machine learning algorithms to predict student outcomes and suggest interventions for at-risk learners.

Ethical AI Principles Implementation

1. Justice and Fairness

Challenge: Ensuring equitable outcomes across diverse student populations **Implementation**:

- **Bias Detection**: Implement regular audits using fairness metrics (demographic parity, equalized odds) to identify disparities across gender, ethnicity, socioeconomic status, and learning disabilities
- **Inclusive Training Data**: Ensure training datasets represent diverse learning styles, cultural backgrounds, and academic abilities
- **Algorithmic Fairness**: Use techniques like adversarial debiasing and fair representation learning to minimize discriminatory outcomes
- **Accessibility**: Design interfaces that accommodate students with disabilities, including screen readers, keyboard navigation, and adjustable text sizing

2. Non-Maleficence (Do No Harm)

Challenge: Preventing psychological harm and educational stigmatization **Implementation**:

- **Positive Framing**: Present recommendations as growth opportunities rather than deficit-based assessments
- **Mental Health Safeguards**: Include triggers to alert counselors when students show signs of distress or declining engagement
- **Gradual Implementation**: Roll out features incrementally to monitor impact and adjust based on feedback
- **Stress Prevention**: Avoid creating pressure through constant monitoring by implementing "digital wellness" periods

3. Autonomy and Consent

Challenge: Respecting student agency while providing helpful guidance **Implementation**:

- **Informed Consent**: Provide clear, age-appropriate explanations of data usage and AI decision-making processes
- **Opt-out Mechanisms**: Allow students and parents to decline participation or specific features without penalty

- **Data Control**: Enable users to view, modify, and delete their data through intuitive privacy dashboards
- **Transparent Recommendations**: Clearly indicate when suggestions come from AI vs. human educators

4. Transparency and Explainability

Challenge: Making AI decisions understandable to educators, students, and parents
Implementation:

- **Explainable AI**: Use interpretable models (decision trees, LIME explanations) to show how recommendations are generated
- **Decision Audit Trails**: Maintain logs of all AI-driven decisions for review and accountability
- **Regular Reporting**: Provide semester reports showing how AI insights contributed to student progress
- **Human-in-the-Loop**: Require educator approval for significant interventions or course changes

5. Privacy and Data Protection

Challenge: Protecting sensitive educational and personal data **Implementation**:

- **Data Minimization**: Collect only necessary data for specific educational purposes
- **Encryption**: Implement end-to-end encryption for all data transmission and storage
- **FERPA Compliance**: Ensure full compliance with Family Educational Rights and Privacy Act
- **Anonymization**: Use differential privacy techniques to protect individual student identities in aggregate analyses

6. Sustainability and Long-term Impact

Challenge: Creating lasting positive educational outcomes **Implementation**:

- **Educator Empowerment**: Train teachers to understand and effectively use AI insights rather than replacing human judgment
- **Continuous Improvement**: Implement feedback loops to refine algorithms based on long-term student outcomes
- **Environmental Responsibility**: Optimize computational efficiency to minimize carbon footprint
- **Knowledge Transfer**: Document and share ethical AI practices with the broader educational technology community

Risk Assessment and Mitigation

High-Risk Scenarios:

1. **Algorithmic Bias**: Students from underrepresented groups receive lower-quality recommendations
    - *Mitigation*: Regular bias audits, diverse testing groups, fairness-aware ML techniques
2. **Privacy Breach**: Unauthorized access to sensitive student data
    - *Mitigation*: Zero-trust security model, regular penetration testing, incident response plan
3. **Over-reliance on AI**: Educators stop using professional judgment in favor of AI recommendations
    - *Mitigation*: Training programs, AI-as-advisor positioning, human oversight requirements
4. **Labeling Effects**: Students internalize AI predictions about their abilities
    - *Mitigation*: Growth mindset framing, dynamic reassessment, positive reinforcement focus

Implementation Timeline

Phase 1 (Months 1-3): Ethics framework development, stakeholder consultation, privacy infrastructure **Phase 2 (Months 4-6)**: Bias testing, fairness metric implementation, transparency features **Phase 3 (Months 7-9)**: Pilot testing with diverse student groups, feedback integration **Phase 4 (Months 10-12)**: Full deployment with continuous monitoring and ethical review

Success Metrics

Ethical Compliance Indicators:

- **Fairness**: <5% performance disparity across demographic groups
- **Privacy**: Zero data breaches, 100% consent compliance
- **Transparency**: 90% of users understand AI recommendations
- **Autonomy**: 100% opt-out functionality availability
- **Beneficence**: Measurable improvement in student outcomes without negative side effects

Ongoing Monitoring:

- Monthly bias audits using IBM AI Fairness 360
- Quarterly stakeholder feedback sessions
- Annual third-party ethical AI assessment
- Continuous privacy impact assessments

Ethical AI Guidelines for Healthcare Implementation(Bonus Task)

A Comprehensive Policy Framework for Responsible Medical AI

Executive Summary

This guideline establishes mandatory ethical standards for AI deployment in healthcare settings, prioritizing patient safety, privacy, and equitable care while fostering innovation that benefits all populations.

1. Patient Consent Protocols

1.1 Informed Consent Requirements

Mandatory Standards:

- **Clear Communication**: All AI system usage must be explained in plain language, specifying how AI assists in diagnosis, treatment, or care planning
- **Specific Purpose Disclosure**: Patients must understand exactly what AI applications will access their data (e.g., diagnostic imaging, drug interaction checking, risk assessment)
- **Opt-out Rights**: Patients retain the right to decline AI-assisted care without penalty or reduced quality of service
- **Ongoing Consent**: Consent must be reconfirmed annually and whenever AI systems are significantly updated

1.2 Vulnerable Population Protections

- **Pediatric Patients**: Require both parental consent and age-appropriate assent for children 7+ years

- **Cognitive Impairment**: Establish legal guardian consent protocols with patient advocacy involvement
- **Emergency Situations**: Implement streamlined consent processes that prioritize life-saving care while documenting AI usage

1.3 Data Usage Transparency

- **Data Scope**: Clearly specify what data types the AI system will access (medical history, lab results, imaging, genetic information)
- **Data Retention**: Inform patients how long their data will be stored and used for AI training
- **Third-party Sharing**: Explicitly disclose any data sharing with AI vendors, research institutions, or other healthcare providers

2. Bias Mitigation Strategies

2.1 Development Phase Requirements

Mandatory Practices:

- **Diverse Training Data**: AI systems must be trained on datasets representing diverse demographics, socioeconomic backgrounds, and medical conditions
- **Bias Testing**: Conduct pre-deployment testing across gender, race, age, and socioeconomic groups using standardized fairness metrics
- **Clinical Validation**: Require validation studies in diverse patient populations before deployment

2.2 Ongoing Monitoring Framework

- **Performance Audits**: Quarterly assessments of AI system performance across demographic groups
- **Disparity Reporting**: Monthly reports identifying any performance disparities exceeding 5% between groups
- **Corrective Actions**: Mandatory system updates within 90 days when bias is detected

2.3 Healthcare Equity Measures

- **Access Monitoring**: Track AI-assisted care access across different communities and insurance types
- **Outcome Tracking**: Monitor health outcomes to ensure AI doesn't exacerbate existing healthcare disparities

- **Community Engagement**: Include diverse community representatives in AI governance committees

3. Transparency Requirements

3.1 Clinical Decision Support Transparency

Mandatory Disclosures:

- **AI Involvement**: Healthcare providers must inform patients when AI contributes to their care decisions
- **Confidence Levels**: Display AI confidence scores and uncertainty measures to clinicians
- **Human Override**: Clearly document when clinicians override AI recommendations and the rationale

3.2 Algorithm Explainability Standards

- **Interpretable Results**: AI systems must provide explanations for their recommendations in clinical terms
- **Decision Pathways**: Maintain audit trails showing how AI systems reach specific conclusions
- **Model Documentation**: Comprehensive documentation of AI model architecture, training data, and limitations

3.3 Patient Access to Information

- **AI Usage Reports**: Patients can request reports detailing how AI was used in their care
- **Decision Explanations**: Provide patient-friendly explanations of AI-assisted decisions
- **Appeal Process**: Establish mechanisms for patients to question or appeal AI-influenced care decisions

4. Implementation Framework

4.1 Governance Structure

Required Roles:

- **AI Ethics Officer**: Designated leader responsible for ethical AI compliance
- **Clinical AI Committee**: Multidisciplinary team including clinicians, ethicists, and patient advocates

- **Patient Advisory Board**: Patient representatives involved in AI policy development and review

4.2 Training and Certification

- **Mandatory Education**: All healthcare providers using AI systems must complete annual ethics training
- **Competency Assessment**: Regular testing of provider understanding of AI capabilities and limitations
- **Continuous Learning**: Updates on new ethical considerations and best practices

4.3 Quality Assurance

- **Pre-deployment Review**: Comprehensive ethical assessment before any AI system goes live
- **Continuous Monitoring**: Real-time tracking of AI system performance and ethical compliance
- **Incident Reporting**: Mandatory reporting of AI-related adverse events or ethical concerns

5. Compliance and Enforcement

5.1 Regulatory Alignment

- **HIPAA Compliance**: Ensure all AI systems meet healthcare privacy requirements
- **FDA Coordination**: Align with FDA guidelines for medical device AI applications
- **International Standards**: Adopt WHO and EU AI ethics principles for global compatibility

5.2 Accountability Measures

- **Legal Liability**: Clear assignment of responsibility for AI-assisted care decisions
- **Malpractice Coverage**: Ensure professional liability insurance covers AI-related incidents
- **Disciplinary Actions**: Defined consequences for non-compliance with ethical AI guidelines

5.3 Regular Review and Updates

- **Annual Policy Review**: Yearly assessment and updating of ethical AI guidelines
- **Stakeholder Feedback**: Regular input from patients, providers, and ethics experts

- **Technology Adaptation**: Guidelines must evolve with advancing AI capabilities

6. Emergency and Special Circumstances

6.1 Crisis Response Protocols

- **Pandemic Adaptations**: Flexible consent processes during public health emergencies
- **Resource Allocation**: Ethical AI use in triage and resource distribution decisions
- **Cross-border Care**: Guidelines for AI use in telemedicine across jurisdictions

6.2 Research and Innovation Balance

- **Ethical Innovation**: Encourage AI advancement while maintaining patient protection
- **Research Consent**: Separate consent processes for AI research vs. clinical care
- **Data Sharing**: Ethical frameworks for sharing de-identified data to improve AI systems

Implementation Timeline

Phase 1 (Months 1-3): Policy development, stakeholder consultation, governance structure establishment **Phase 2 (Months 4-6)**: Staff training, system audits, consent process implementation **Phase 3 (Months 7-9)**: Pilot testing, feedback collection, policy refinement **Phase 4 (Months 10-12)**: Full implementation, monitoring systems activation, compliance assessment

Success Metrics

- **Patient Satisfaction**: 90% approval rating for AI transparency and consent processes
- **Equity Measures**: <3% performance disparity across demographic groups
- **Compliance Rate**: 100% adherence to consent and transparency requirements
- **Safety Record**: Zero AI-related adverse events due to ethical violations

This policy framework ensures that AI in healthcare serves humanity's best interests while advancing medical innovation responsibly and equitably.