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 refers to systematic and unfair discrimination that occurs when AI systems produce results that are prejudiced against certain groups of people, often reflecting historical inequalities or flawed assumptions embedded in training data or system design.

Two examples:

1. **Facial Recognition Systems**: Many facial recognition technologies have shown significantly higher error rates for people with darker skin tones and women, particularly Black women. This occurs because training datasets historically contained predominantly lighter-skinned male faces, leading to systems that perform poorly on underrepresented groups.
2. **Hiring Algorithms**: Some AI recruitment tools have exhibited bias against women or certain ethnic groups. For instance, Amazon scrapped an AI recruiting tool that systematically downgraded resumes from women because it was trained on historical hiring data that reflected past discrimination in male-dominated fields.

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

Transparency refers to how openly the AI system's components, data sources, and operations are disclosed. Essentially about visibility.

Explainability describes the extent to which the internal mechanics and decision- making process of the AI can be understood by humans- about clarity.

Why both matter:

Ethical Accountability: Together they enable users, regulators, and developers to evaluate fairness, safety, and ethical implications.

Trust and Adoption: Explainable and transparent systems are more likely to be accepted and responsibly deployed.

Debugging and Improvement: Developers need both to identify flaws and enhance system performance without relying on "black box" assumptions.

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

Data Processing Requirements: AI systems must have a lawful basis for processing personal data, with explicit consent or legitimate interest clearly established. This affects how training data is collected and used.

Data Minimization: AI developers must ensure they only collect and process data that is necessary for their specific purpose, limiting the scope of data used in training and operation.

Right to Explanation: While not explicitly stated, GDPR's requirement for "meaningful information about the logic involved" in automated decision-making has been interpreted to support individuals' rights to understand AI decisions that significantly affect them.

Privacy by Design: AI systems must incorporate data protection considerations from the design phase, including techniques like differential privacy, federated learning, or data anonymization.

Data Subject Rights: Individuals have rights to access, rectify, erase, and port their data, which creates technical challenges for AI systems that have already incorporated this data into trained models.

Impact Assessments: High-risk AI processing requires Data Protection Impact Assessments (DPIAs), adding regulatory overhead to AI development projects.

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.*

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 (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.

Primary Source: Historical Training Data The bias originated from training the AI system on Amazon's historical hiring data from a 10-year period, which predominantly reflected hiring patterns in male-dominated tech roles. The training data contained:

- Resumes from mostly male hires (reflecting past discrimination and industry demographics)
- Historical hiring decisions that favored traditionally male-associated qualifications
- Language patterns that correlated with gender (e.g., words like "executed" or "captured" more common in male resumes)

Secondary Source: Feature Engineering The model learned to penalize resumes containing words like "women's" (as in "women's chess club captain") and downgraded graduates from all-women's colleges, indicating the algorithm identified gender-correlated features as negative predictors of job performance.

Model Design Issue: Proxy Discrimination The system wasn't designed to recognize and mitigate proxy variables—features that correlate with protected characteristics but aren't directly discriminatory themselves.

2. Three Proposed Fixes

Fix 1: Diversified and Balanced Training Data

- Collect training data from multiple time periods and companies with diverse hiring practices
- Artificially balance the dataset to include equal representation of successful hires across gender lines
- Use synthetic data generation to create balanced examples while preserving realistic resume patterns
- Remove or de-bias historical hiring decisions that reflected past discrimination

Fix 2: Fairness-Aware Machine Learning Techniques

- Implement demographic parity constraints during model training to ensure equal positive prediction rates across groups
- Use adversarial debiasing techniques that train the model to be unable to predict protected characteristics from its internal representations
- Apply post-processing calibration to ensure equal opportunity (equal true positive rates) across gender groups

Fix 3: Feature Auditing and Bias Detection

- Conduct systematic feature importance analysis to identify gender-correlated features
- Implement automated bias detection systems that flag when certain words or patterns disproportionately affect one group
- Use techniques like LIME (Local Interpretable Model-agnostic Explanations) to understand individual predictions and identify discriminatory patterns
- Create gender-neutral feature engineering that focuses on skills and qualifications rather than language patterns

Fairness Metrics for Evaluation

- **Demographic Parity:** Measures whether selection rates are equal across different gender groups.
- **Equal Opportunity Difference:** Checks whether qualified candidates from each group have equal chances of being selected.
- **False Positive / False Negative Rates by Group:** Evaluates whether error rates differ across genders — this is crucial for understanding unintended harm.

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.

Ethical Risks

1. **Wrongful Arrests & False Identifications:** Misidentifications have led to documented cases where innocent individuals — disproportionately from minority communities — were arrested or detained. This undermines trust in law enforcement and violates due process.
2. **Privacy Violations:** Continuous surveillance using facial recognition can infringe on citizens' right to privacy, especially if deployed without consent or transparency in public spaces.
3. **Bias Amplification:** Existing societal biases (racial, gender, socioeconomic) can be exacerbated when AI systems inherit skewed data patterns. This creates a feedback loop of unequal treatment.
4. **Chilling Effect on Freedoms:** Ubiquitous surveillance may deter free expression, peaceful protest, and public engagement — essential in democratic societies.

Recommended Policies for Responsible Deployment

1. **Rigorous Bias Audits:** Require independent audits to assess accuracy across demographic groups before deployment. Disparities should be addressed as a precondition.
2. **Human Oversight Protocols:** Facial recognition matches should never be treated as definitive — always subject to manual review by trained professionals with checks against wrongful action.
3. **Clear Consent & Transparency Standards:** Inform the public when and where systems are used. Build opt-out provisions where feasible and publish algorithmic performance data.
4. **Limit Use to Critical Contexts:** Restrict application to high-stakes scenarios (e.g., missing persons, terrorism prevention) and prohibit mass surveillance or routine use for low-level infractions.
5. **Legal Safeguards & Accountability:** Enact laws that define acceptable use, prohibit abuse, and enable citizens to challenge misuse — similar to how GDPR empowers EU citizens over their data.

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.

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 AI Reflection: DG Walls and Interiors Website Project

Project Overview

Project: Development of an AI-enhanced website for DG Walls and Interiors, a specialized interior design business focusing on accent walls, custom woodwork, and comprehensive interior design services.

Scope: Creating a modern, responsive website that incorporates AI-powered features such as:

- Virtual room visualization tools
- Automated design recommendations
- Customer inquiry chatbot
- Portfolio management system
- Quote generation system

Ethical AI Principles Framework

1. Transparency and Explainability

Challenge: The AI recommendation system for interior design choices could be perceived as a "black box" to clients.

Implementation Strategy:

- Clearly label all AI-generated content and recommendations
- Provide explanations for why certain design elements or color schemes are suggested
- Include disclaimers about AI limitations in creative decision-making
- Offer users the ability to understand how their preferences influence recommendations

Example: When the AI suggests a particular accent wall design, the system will explain: "This recommendation is based on your preference for modern aesthetics, the room's natural lighting conditions, and popular design trends in similar spaces."

2. Privacy and Data Protection

Challenge: The website will collect sensitive personal information including home layouts, financial information for quotes, and design preferences.

Implementation Strategy:

- Implement GDPR-compliant data collection practices
- Use minimal data collection principles - only gather what's necessary
- Provide clear privacy policies in plain language
- Offer users control over their data (deletion, modification, portability)
- Ensure secure data storage and transmission
- Anonymize analytics data

Specific Measures:

- Encrypted storage for client photos and room measurements
- Automatic deletion of unused uploaded images after 30 days
- Opt-in consent for marketing communications
- Clear separation between functional data and marketing data

3. Fairness and Non-Discrimination

Challenge: AI design recommendations might inadvertently reflect biases in training data, potentially excluding certain cultural aesthetics or economic backgrounds.

Implementation Strategy:

- Diversify training data to include various cultural design traditions
- Regularly audit AI recommendations for bias
- Ensure accessibility across different economic segments

- Include diverse design styles in the recommendation engine
- Test with diverse user groups during development

Monitoring Plan:

- Monthly analysis of recommendation patterns across different demographic groups
- Feedback collection system to identify potential bias
- Regular updates to the AI model with more inclusive training data

4. Accountability and Responsibility

Challenge: Determining liability when AI provides design recommendations that don't meet client expectations or cause dissatisfaction.

Implementation Strategy:

- Establish clear human oversight for all AI-generated recommendations
- Maintain detailed logs of AI decision-making processes
- Implement human review checkpoints for significant recommendations
- Create clear escalation procedures for AI-related issues
- Provide human alternative for all AI-powered features

Accountability Measures:

- Designated AI ethics officer (myself) responsible for monitoring system behavior
- Regular audits of AI recommendations against client satisfaction
- Clear documentation of AI limitations and capabilities
- Human designer final approval for all major recommendations

5. Beneficial Impact and Social Good

Challenge: Ensuring the AI enhances rather than replaces human creativity and expertise in interior design.

Implementation Strategy:

- Position AI as a tool to augment human designers' capabilities
- Focus on democratizing good design principles rather than replacing professional expertise
- Include educational content about design principles
- Support local artisans and sustainable design practices through recommendations
- Promote accessibility in design choices

Positive Impact Goals:

- Make quality interior design guidance accessible to broader economic segments
- Reduce waste by helping clients make better-informed decisions
- Support local craftsmanship through supplier recommendations
- Promote sustainable and environmentally conscious design choices

Implementation Safeguards

Technical Safeguards

- **Fail-safe mechanisms**: If AI systems fail, users can still access all website functionality through traditional means
- **Regular testing**: Continuous monitoring of AI performance and accuracy
- **Version control**: Ability to quickly revert to previous AI model versions if issues arise
- **Performance monitoring**: Real-time tracking of AI response accuracy and user satisfaction

Ethical Safeguards

- **Ethics review board**: Quarterly reviews of AI system performance and ethical compliance
- **User feedback integration**: Regular incorporation of user feedback into ethical considerations
- **Transparency reports**: Annual reports on AI system performance and ethical considerations
- **Continuous education**: Ongoing training on AI ethics for all team members

Monitoring and Evaluation

Key Performance Indicators (KPIs)

- **User satisfaction scores** with AI recommendations
- **Fairness metrics** across different demographic groups
- **Transparency scores** based on user understanding of AI recommendations
- **Privacy compliance** audit results
- **Human oversight effectiveness** measurements

Review Schedule

- **Weekly**: Technical performance monitoring
- **Monthly**: Bias and fairness analysis
- **Quarterly**: Comprehensive ethics review
- **Annually**: Full system audit and ethics assessment

Continuous Improvement Plan

Feedback Integration

- Implement user feedback loops for AI recommendations
- Regular surveys on user trust and satisfaction with AI features
- Professional designer feedback on AI recommendation quality
- Client outcome tracking to measure real-world effectiveness

Adaptation Strategy

- Regular updates to AI models based on new data and feedback
- Incorporation of emerging ethical AI best practices
- Adaptation to changing regulatory requirements
- Evolution of features based on user needs and ethical considerations

---

Bonus Task (Extra 10%)

- **Policy Proposal**: Draft a 1-page guideline for *ethical AI use in healthcare*. Include:
    - Patient consent protocols.
    - Bias mitigation strategies.
    - Transparency requirements.

Ethical AI Use in Healthcare Guidelines

Patient Consent Protocols

Informed Consent Requirements: Healthcare providers must obtain explicit consent before using AI systems in patient care. Patients must be informed about AI technologies being used, data processing methods, potential risks and limitations, and their right to opt-out without compromising care quality. Consent must be documented and regularly reviewed during extended treatments.

Emergency and Vulnerable Populations: In emergencies, AI use is limited to life-saving interventions with consent sought immediately when possible. Special procedures apply to minors, cognitively impaired patients, and those with language barriers, requiring simplified forms and additional safeguards.

Bias Mitigation Strategies

Data Diversity: AI training datasets must include representative samples across all demographics including age, gender, race, ethnicity, socioeconomic status, and geography. Regular audits must identify and address representation gaps to prevent discriminatory outcomes.

Algorithmic Fairness: All AI systems must undergo rigorous bias testing before deployment and during routine evaluations. Performance metrics must be disaggregated by demographic groups to identify disparities. Systems showing significant bias must be retrained or discontinued until fairness standards are met.

Continuous Monitoring: Healthcare organizations must establish protocols tracking AI performance across patient populations. Regular retraining cycles must incorporate new data and address identified biases. Staff must receive training on recognizing and mitigating AI bias in clinical decisions.

Transparency Requirements

System Explainability: AI systems must provide clear, understandable explanations for recommendations and decisions. Healthcare providers must access and interpret AI reasoning to make informed clinical judgments. Complex decisions must be translatable into plain language for patient communication.

Performance Disclosure: Organizations must publicly report AI system performance metrics including accuracy rates, error types, and demographic variations. Regular transparency reports must detail system usage, outcomes, and improvements.

Documentation: Comprehensive records must be maintained for all AI-assisted clinical decisions, including system inputs, outputs, and final decisions. Audit trails must enable retrospective analysis and quality improvement.

Implementation and Governance

Oversight: Healthcare organizations must establish AI ethics committees with diverse representation including clinicians, ethicists, patient advocates, and technical experts to review implementations and ensure compliance.

Training: All staff using AI systems must complete mandatory training on ethical AI use, bias recognition, and patient communication. Regular competency assessments must ensure maintained knowledge and skills.

Compliance: Organizations must conduct regular audits and implement corrective actions for violations. Patient feedback mechanisms must address ethical concerns about AI use in their care.

Accountability: Non-compliance may result in regulatory sanctions including suspension of AI privileges. Individual practitioners misusing AI systems may face professional disciplinary action.