📚 **Final PDF Structure: PLP Academy AI Ethics Assignment**

Title: COMPAS Bias Audit and Ethical AI Reflection
Author: Leonard Phokane
Date: July 2025

Part 1: Theoretical Understanding

📘 1. Short Answer Questions

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

Definition: Algorithmic bias refers to systematic errors within AI systems that result in unfair, discriminatory, or unequal treatment of individuals or groups—often stemming from skewed training data, biased design choices, or flawed assumptions.

Examples:

- **Hiring Tools Bias:** Amazon's former recruitment AI downgraded resumes with the word "women's," unfairly penalizing female candidates due to biased historical hiring data.
- **Facial Recognition Errors:** Systems like those used in law enforcement misidentify minorities at significantly higher rates, leading to wrongful arrests and unequal surveillance.

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

Transparency is about openness—making details about AI systems (e.g., data sources, algorithms, model architecture) accessible and clear to stakeholders.

Explainability is about interpretation—enabling users to understand *why* and *how* AI made a specific decision or prediction.

Why They Matter:

- **Trust & Accountability:** Stakeholders can hold developers accountable and identify problematic behaviors.
- **Regulation & Governance:** Transparent systems are easier to regulate, and explainable decisions help comply with legal standards.
- **User Empowerment:** Clear AI behavior reduces fear, fosters confidence, and allows informed engagement.

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

GDPR ensures AI systems handle personal data responsibly by enforcing:

- **Consent and Control:** Individuals must explicitly opt into data collection and can withdraw consent.
- **Data Minimization & Purpose Limitation:** AI systems should only collect data necessary for specific tasks.
- **Right to Explanation:** Individuals have the right to understand how automated decisions affect them.
- **Accountability & Impact Assessments:** Developers must assess and mitigate risks to data subjects before deployment.

This encourages ethical design, elevates user rights, and challenges developers to build systems that prioritize human dignity.

🧭 2. 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.

Part 2: Case Study Analysis

📂 Case 1: Biased Hiring Tool – Amazon's AI Recruiting System

🔍 Source of Bias

- **Training Data:** The model was trained on resumes submitted over a 10-year period—most from male applicants—leading to gendered bias patterns.
- **Feature Selection:** The system downgraded resumes mentioning terms like "women's" or female-associated activities, reflecting biases in historical hiring trends.
- **Reinforcement Bias:** The model used success metrics that replicated past hiring decisions, amplifying biased outcomes.

🛠️ Proposed Fixes

1. **Rebalance Training Data:** Use a gender-balanced dataset with inclusive features that don't penalize based on identity markers.
2. **Bias Mitigation Algorithms:** Apply pre-processing techniques from AI Fairness 360 to neutralize harmful correlations**.**
3. **Human-AI Collaboration:** Reintroduce human oversight in final decision-making to audit algorithmic recommendations.

📏 Fairness Evaluation Metrics

- **Disparate Impact Ratio:** Compare hiring rates across gender groups to identify disparity.
- **Equal Opportunity Difference**: Examine true positive rates to ensure equal chances of being shortlisted.
- **Bias Confusion Matrix Analysis**: Analyze false positives/negatives across demographic splits.

🕵️ Case 2: Facial Recognition in Policing

⚠️ Ethical Risks

- **Wrongful Arrests**: Higher misidentification rates for minorities could lead to unfair detentions.
- **Privacy Violations:** Surveillance without consent undermines civil liberti**es.**
- **Erosion of Trust:** Communities may grow distrustful of law enforcement and government AI use.

🛡️ Responsible Deployment Policies

1. **Mandatory Bias Audits:** Use tools like AI Fairness 360 before deployment to assess accuracy across demographics.
2. **Regulated Usage Protocols:** Limit use to confirmed criminal investigations; prohibit real-time public surveillance.
3. **Transparency & Public Oversight:** Publish model performance reports and allow third-party audits.
4. **Human-in-the-Loop Systems:** Always require human validation before action is taken based on AI recognition.

Part 3: Practical Audit

✅ Part 3: Practical Audit (25%)

📝 Summary Report (300 Words)

Title: Racial Bias Audit in COMPAS Recidivism Predictions

- Our audit leveraged IBM's AI Fairness 360 toolkit to examine racial disparities in the COMPAS dataset. This dataset, widely used in recidivism prediction, has been flagged for potential racial bias—particularly its treatment of African-American defendants.
- Using binary fairness metrics, we found a **disparate impact ratio below 0.8**, suggesting significant bias against unprivileged groups. Moreover, the **false positive rate for African-Americans** was markedly higher than for Caucasians, implying they are unfairly labeled as high-risk more often. This pattern poses ethical risks—exacerbating over-policing and reinforcing systemic injustice.
- To mitigate bias, we recommend implementing **preprocessing techniques** like reweighing and resampling. These methods adjust distributions without altering outcome labels, promoting equity while retaining model integrity. Additionally, retraining models using fairness-aware algorithms can help recalibrate outcomes.
- We also suggest integrating fairness dashboards and continuous bias audits post-deployment. Tracking metrics such as **equal opportunity difference** and **disparate impact ratio** can highlight persistent issues, while transparent reporting ensures accountability.
- Ethical AI isn't just about avoiding harm—it's about actively building trust. Through conscientious dataset handling, model design, and stakeholder feedback loops, we can design systems that serve justice, not distort it.

⭐ Bonus Task (Extra 10%)

Policy Proposal: Ethical AI Use in Healthcare

🧭 Ethical AI Use in Healthcare: Guideline Proposal

1. 📝 Patient Consent Protocols

- **Informed Consen**t: Patients must be notified when AI is used in diagnosis, triage, or treatment planning. Consent should be explicit and documented before deployment.
- **Opt-out Optio**n: AI-assisted decisions must remain optional. Patients must be allowed to request human-only reviews and override algorithmic suggestions.
- **Accessible Explanatio**n: Use plain language and visual summaries to describe how AI systems operate and what data they process.

- **Data Scope Disclosure:** Patients must be informed of:
    - What personal data is used (e.g., EHRs, imaging)
    - Where and how data is stored
    - Whether data contributes to model training or improvement

2. ⚖️ Bias Mitigation Strategies

- **Fairness Audit**s: Regularly evaluate models for disparate impacts across race, gender, age, disability status, and socioeconomic markers.
- **Representative Dataset**s: Source training data from diverse populations to reflect real-world heterogeneity and avoid systemic exclusion.
- **Algorithmic Rebalancin**g: Apply techniques like reweighing, adversarial debiasing, and equal opportunity adjustments where disparities are detected.
- **Clinical Oversight Panel**s: Establish multidisciplinary teams—including ethicists, clinicians, and data scientists—to review AI outcomes for fairness.

3. 🔍 Transparency Requirements

- **Model Explainabilit**y: Every deployed AI system must provide interpretable outputs that clinicians and patients can understand and challenge.
- **Decision Log**s: Maintain audit trails of AI-assisted decisions, including model versioning and inputs used.
- **Stakeholder Access**: Researchers, patients, and regulators should have access to:
    - High-level system documentation
    - Performance evaluations
    - Limitations or known failure modes
- **Public Disclosur**e: AI tools used in clinical settings must be registered and disclosed in national medical technology registries.

Conclusion: Ethical AI in healthcare demands that we protect patient autonomy, prevent harm through bias, and ensure trust through transparency. These guidelines are designed to foster patient-centered, accountable, and inclusive AI systems.