GROUP TASK -3

ML Ethics & Bias Case Study : Groups analyze a case where machine learning caused bias or unintended consequence (like biased hiring tools or unfair credit scoring) and propose solutions or guidelines to and applicable algorithms.

1. Introduction:-

Machine Learning systems are increasingly used in high-stakes decisions such as hiring, credit scoring, healthcare, and criminal justice. While these systems aim to improve efficiency and objectivity, they can unintentionally reproduce social bias.
This case study analyzes the biased hiring tool developed by Amazon and proposes technical, organizational, and regulatory solutions to reduce bias in ML systems.

2. Case Overview:-

Around 2014, Amazon built an AI-based recruiting system to automatically screen resumes and rank job applicants.

Objective:
Automate hiring and identify top talent using machine learning.

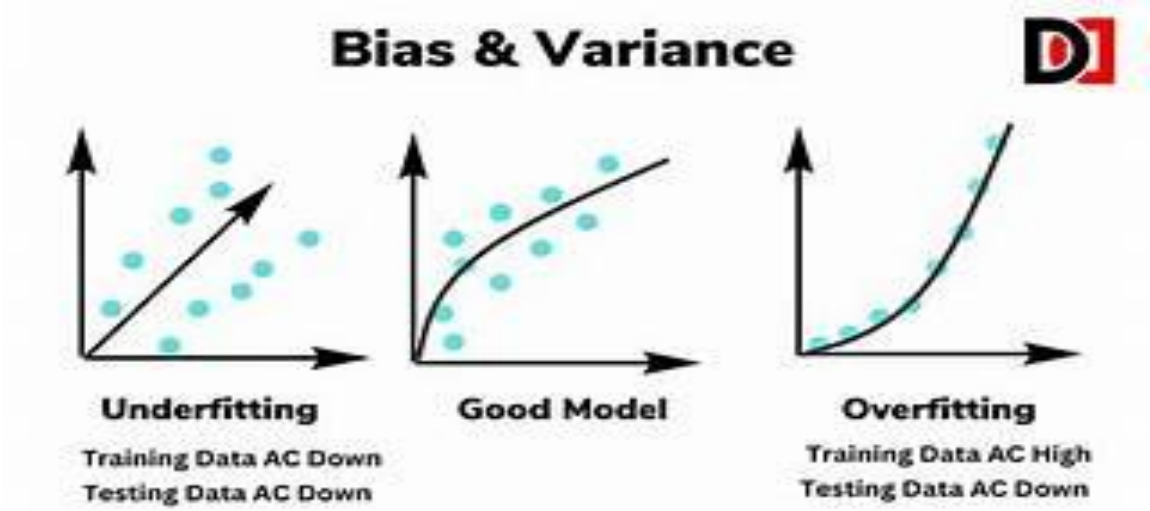
What Happened:
The model was trained on historical hiring data (mostly male candidates in tech roles). As a result, the system learned patterns that favored male applicants.

Observed Bias:

- Penalized resumes containing the word "women's".
- Downgraded candidates from women's colleges.
- Favored resumes with language statistically common in male applicants.

The tool was eventually discontinued because it showed gender bias.

3. Why Did Bias Occur?



3.1 Historical Data Bias
The training data reflected past hiring patterns where men were overrepresented.
The model learned and replicated this imbalance.

3.2 Proxy Variables
Even after removing gender as a feature, the model inferred gender through indirect indicators such as:

- College names
- Word usage
- Activities

3.3 Optimization for Accuracy Only
The system focused on predictive accuracy, not fairness.

3.4 Lack of Bias Auditing
There was insufficient fairness testing before deployment.

4. Ethical Issues Identified:-

- Gender discrimination
- Violation of equal opportunity principles
- Lack of transparency
- Over-reliance on automation
- Accountability challenges

Proposed Solutions to Improve the Model
Below are practical and implementable solutions suitable for ML systems.

A. Technical Solutions
1. Use Fairness Metrics Alongside Accuracy
Instead of optimizing only accuracy, evaluate models using fairness metrics such as:

- **Demographic Parity**
- **Equal Opportunity**
- **Equalized Odds**

The model should not significantly favor one group over another.

2. Balanced and Representative Training Data

- Ensure diversity in the dataset
- Perform statistical analysis to detect underrepresented groups
- Use data re-sampling or re-weighting techniques

Example:
If males = 80% and females = 20%
→ Apply rebalancing techniques before training.

3. Remove Proxy Bias
Even if sensitive attributes (like gender) are removed:

- Perform feature correlation analysis
- Detect indirect gender indicators
- Use adversarial debiasing techniques
- Apply fairness-aware feature selection

4. Human-in-the-Loop Approach
AI should assist, not fully replace human decision-makers.
Implement:

- Manual review of shortlisted candidates
- Override mechanism
- Continuous feedback loop

5. Bias Testing Before Deployment
Before production:

- Run fairness audits
- Conduct subgroup performance analysis
- Perform simulation testing on different demographic groups

6. Explainable AI (XAI)
Use interpretable models or tools such as:

- SHAP
- LIME
- Feature importance analysis

This ensures:

- Transparency
- Accountability
- Easier bias detection

B. Organizational Solution:
1. Diverse Development Teams
Include:

- Different genders
- Different cultural backgrounds
- Ethics experts
- Legal advisors

Diverse teams are more likely to identify hidden bias.

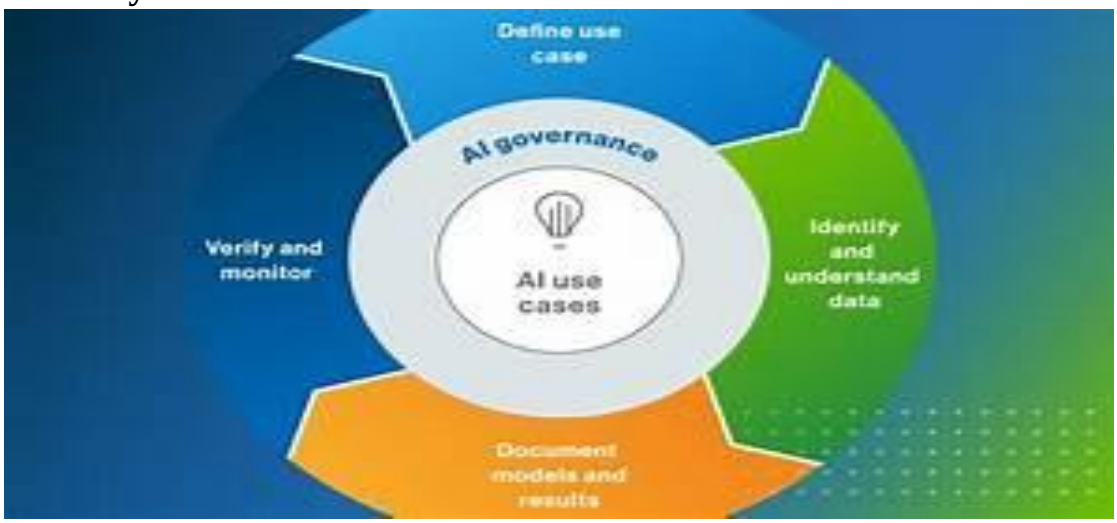
2. Internal AI Ethics Committee
Organizations should:

- Evaluate risks before deployment.
- Conduct impact assessments.
- Monitor long-term fairness.

3. Regular Audits

- Third-party audits
- Annual bias review
- Continuous performance monitoring

C. Policy & Governance Recommendations



1. Algorithmic Impact Assessment (AIA)

Before deploying ML systems in hiring:

- Assess social risks
- Evaluate discrimination potential
- Document decision-making logic

2. Right to Explanation

Applicants should have:

- Access to reasons for rejection
- Ability to request human review

3. Compliance with Anti-Discrimination Laws

AI systems must comply with:

- Equal Employment Opportunity laws
- Fair hiring standards

5. General Guidelines for Ethical ML Development:-

1. Audit dataset for imbalance.
2. Use fairness metrics in evaluation.
3. Monitor model after deployment.
4. Maintain transparency and documentation.

6. Key Learnings:-

1. ML models learn from historical data — including past discrimination.
2. Removing sensitive attributes does not automatically remove bias.
3. Fairness must be intentionally designed and measured.

7. Conclusion:-

The case study on machine learning ethics and bias clearly demonstrates that ML systems are not neutral; they often reflect and amplify the biases present in historical data, design choices. For example, biased risk assessment tools such as COMPAS have shown how lack of transparency and biased training data can disproportionately affect certain racial groups. This highlights the ethical responsibility of developers and organizations to go beyond accuracy and prioritize fairness, accountability, transparency, and explainability (FATE) in ML systems.

The study reinforces that ethical ML is not a one-time technical fix but a continuous socio-technical process involving:

- Diverse and representative datasets.
- Fairness-aware model design.
- Regular bias audits and monitoring.
- Clear documentation and transparency.
- Human oversight in decision-making.