AI for Software Engineering Assignment

WEEK 7 (AI Ethics PART 1)

DATE : November 2025

INSTITUTION : PLP Academy

Group members

- **Obuye emmanuel chukwuemeke**
- **Eunice fagbemide**
- **antonia othetheaso**
- **mark ireri daizy**
- **jepchumba kiplagat**

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 an AI system produces outcomes that disproportionately favor or disadvantage certain individuals or groups. This bias often arises from the **data used to train the model**, the **design choices made by developers**, or the **hidden assumptions embedded in the algorithm** itself.

One major cause of algorithmic bias is **biased or incomplete data**. If the training data does not accurately represent all groups—for example, if it contains more information about one gender, race, age group, or region—the AI system will learn patterns that reinforce these imbalances. As a result, the algorithm may treat underrepresented groups unfairly. Another factor is **historical bias**, where data reflects past inequalities. When AI learns from such data, it can unintentionally repeat or amplify those inequalities.

Design decisions also play a role. Sometimes developers may unintentionally embed their own assumptions or overlook important factors, leading to biased rules or predictions. In other cases, **optimization choices**—such as maximizing accuracy without considering fairness—can cause an AI system to favor majority groups because they dominate the dataset.

Algorithmic bias becomes especially dangerous when AI is used in sensitive areas like recruitment, policing, lending, healthcare, or education. Biased decisions in these fields can deny people opportunities, misjudge their risks, or provide lower-quality services. Because AI systems often appear objective or neutral, biased outcomes can go unnoticed and persist for long periods.

To reduce algorithmic bias, developers and organizations must adopt **ethical AI practices**, such as diversifying datasets, performing fairness audits, testing models across different groups, and involving diverse teams in the design process. Transparency and accountability are also essential to ensure AI systems serve all users fairly and responsibly.

Examples:

1. **Biased Hiring Algorithms:**
   An AI recruitment tool trained on historical company data may favor male candidates over female candidates because past hiring patterns were biased.

2. **Facial Recognition Errors:**
   Many facial recognition systems perform significantly better on light-skinned faces than dark-skinned faces due to underrepresentation of diverse groups in training datasets.

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

- **Transparency** refers to how openly the internal workings, design choices, data sources, and decision-making steps of an AI system are disclosed. It is about *visibility* into how the system operates.

- **Explainability** is the ability of an AI system to clearly communicate *why* and *how* it made a particular decision or prediction in a way humans can understand.

Why both are important:

- They build **trust** between users and AI systems.

- They help detect and mitigate **bias or errors**.

- They support **accountability**, allowing stakeholders to question or challenge AI decisions.

- They are essential for **regulatory compliance**, especially in high-risk applications like healthcare or finance.

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

GDPR affects AI development in the following ways:

- **Data Protection Requirements:** AI developers must obtain clear consent to collect and use personal data, and must ensure data is processed lawfully and securely.

- **Right to Explanation:** Users can request explanations about automated decisions that significantly affect them, pushing developers to build more explainable systems.

- **Data Minimization:** AI systems must only use data that is necessary for their purpose, limiting unnecessary or excessive data collection.

- **Prohibition of Harmful Automated Profiling:** Certain types of automated decision-making that negatively impact individuals are restricted or require human oversight.

- **Accountability:** Organizations must demonstrate compliance through documentation, audits, and impact assessments (like DPIAs).

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.