UNIT - III

BIAS AND FAIRNESS IN AI SYSTEMS

Bias in AI :- Bias in AI means that an ai system gives unfair, unequal or undiscriminatory results.

This happens when the AI Learns from data that contains human mistakes, stereotypes or imbalances.

What is Bias in AI ?

Bias in AI occurs when an Ai system favours one group over another or produces results that are not fair.

Since Ai learns from data, if the data is baised, the AI will also behave in a baised way.

Ex: If a facial recognition system is trained mostly or images of one skin tone, it may not work well for others.

Causes of Bias in AI :-

① **Baised Training Data :-**

Data may Contain stereotypes or may not represent all groups equally.

② **Imbalanced data :-**

If some groups are missing or under represented in the dataset.

③ **Human Bias :-** Developer's assumptions and decisions Can un-intentionally introduce bias.

④ **Historial Bias :-** past inequalities get stored in data and used by AI.

⑤ **measurement or Labeling Error :-**

InCorrect or inConsistent data labeling leads to biased learning.

Types of Bias in AI :-

* Data Bias → Bias present in the training data.
* Algorithmic Bias → Bias caused by the model design
* Sample Bias → when the sample does not represent the whole population.
* prejudice Bias → Reflects human stereotypes
* measurement Bias → Incorrect data collection

Effects of Bias in AI :-

* Unfair hiring or job selection
* wrong medical decisions
* Inaccurate facial recognition
* Inequality in loan approvals
* Discrimination in criminal justice systems

How to Reduce Bias in AI :-

* Use diverse and balanced datasets
* Conduct regular adults of AI system
* Ensure human oversight in decision making.
* Use transparent algorithms
* Apply ethical guidelines during AI development.

Ethics in AI :-

Ethics in Artificial intelligence refers to the set of moral principles, guidelines and values that ensure AI systems are developed and used in a responsible, safe and fair manner. The goal is to make AI benefits society without causing harm.

Why Ethics in AI is important :-

* AI effects may decisions (hiring, healthcare, banking, education)
* Unethical AI can lead to bias, privacy loss, unemployment, misuse or even harm.

* ensure trust, safety and accountability.

Key ethical principles in AI :-

① Fairness :- AI should treat everyone equally and avoid discrimination based on gender, race, caste, religion, age etc.

② Transparency :- The working of AI systems should be clear and explainable. Users should know how and why decisions are made.

③ Accountability :- Developers and organizations must take responsibility for AI's actions and decisions.

④ privacy :- AI must protect personal data and follow proper data security rules.
No misuse or unauthorized access to user information.

⑤ Safety :- AI systems should not cause harm they must be tested to ensure safe operation in real environments.

⑥ Reliability :- AI should work accurately and consistently across different situations.

⑦ Human oversight :- Humans must be able to control, stop or override AI decisions when needed.

⑧ Inclusiveness :- AI must be built in a way that benefits all groups of society.

Examples of Ethical Issues in AI :-

* A hiring AI rejecting women more than → Bias problem.
* Face recognition misidentifying darker skin tones → Fairness issue.
* Apps collecting and selling personal data → privacy issue.
* AI-generate deepfakes used to spread misinformation → misuse.

Considerations :-

These are the important things we need to keep in mind while designing or using AI.

1. Avoid Discrimination :- AI should not favour or harm any group Based on : → Gender,
→ Religion,
→ Age
→ socio-economic status

Ex: Hiring AI should not prefer men over women

2. protect privacy :- AI should handle personal data safely.

It should not misuse, leak, & collect unnecessary data.

Ex: Face recognition apps should not store photo without permission.

3. Align AI with Human Values :-

AI must follows human morals, cultural values, and basic Ethics.

It should do what security Consider "right"

Ex: A chatbots should not give harmful advic

4. Prevent misuse of AI :- AI can be used for :

* Deepfakes * Spying * Hacking * fraud.

So, Systems must be protected from bad use.

Ex: Deepfakes videos should not be used to damage someone's reputations.

5. Ensure Transparency :-

users should understand how AI works and why it made a decision.

Ex: If AI rejects a loan, it must Explain why.

6. Accessibility and fair use :-

AI should be accessible to Everyone, not only rich people or big Companies

Ex: Education AI tools should be available to all students.

⇒ Fairness in AI :- Fairness in AI means making sure that AI systems treat all people equality, without discrimination, bias or unfair decisions.

AI should not favour or human anyone based on

* Gender * Race * Religion * Age * Caste
* Economic status * Language * Disability
* other personal characteristics.

Why Fairness in AI is Important?

1. Avoids discrimination :- AI decisions affects jobs, loans, education, healthcare and more. Biased AI can harm people.

2. Builds trust :- when AI is fair, people trust the system.

3. Improves accuracy :- A fair AI system performances better for all groups.

4. protects human rights :-

Everyone deserves equal treatment.

How to Ensure Fairness in AI ?

1. Use diverse and balanced datasets include data from all groups fairly.
2. Regularly test for bias check if the model treats any group unfairly.
3. Transparent algorithms clearly explain how decisions are made.
4. Human oversight experts must monitor and correct AI decisions.
5. Ethical guidelines and law follow standards to ensure rights and fairness.

Examples of Fairness Issues :

* A job recruitment AI prefers men over women
* A loan approval AI rejects poorer communities unfairly.
* A face recognition system works better for some skin tones.

Fairness metrics :-

1. Demographic parity :- AI should give equal outcomes for all groups.

Ex: If 60% of men get a loan, then about 60% of women should also get a loan.

2. Equal Opportunity :- AI should have equal true positive rates for all groups.

Ex: If a qualified person applies, the model should approve them at the same rate across groups.

3. Callibration :- For people with the same predicted score, the real outcome should be similar across groups.

Ex: If a model give a 70% chance of succes, then all groups should actually succeed around 70% of the time.

4. Treatment Equality :- The ratio of false negatives to false positives should be similar across groups.

5. Individual Fairness :- The ratio of false negatives to false positives should be similar across groups.

6. Predictive Parity :- The precision (positive predictive value) should be same across groups.

Bias mitigation strategies :-

1. pre-processing Techniques :-

a) Data Balancing / Re-sampling :-

* over sampling minority groups
* undersampling majority groups Ensures equal representation

b) Data cleaning :- Remove biased, incorrect, or incomplete data.

c) Data Augementation :-

Create more samples for under-represented groups.

d) Reweighting :- Assign higher weights to minority group samples to the algorithm treats them equally.

e) Removing Sensitive Attributes :-

Remove features like gender, caste, race - if appropriate.

2. In processing Techniques :-

Fairness is built inside the algorithm during model training.

a) Fairness - aware Algorithms :- use models that include fairness constraints

b) Adding Regularization Terms :-

Add penality in the loss function when the model becomes unfair.

c) Adversarial Debiasing :- Train a second model to detect bias. main model learns to hide or eliminate that bias.

d) Constraint optimization :- set fairness togels (e.g., equal opportunity) as part of training.

⇒ Transparency in AI Systems :-

Transparency in AI means clearly showing how an AI system works, how it make decisions and why it gives a particular output.

⇒ It makes the internal process understandable to users.

⇒ It makes working visible, understandable and open, so users know how and why an AI made a decision.

⇒ Why Transparency is important :-

1. Builds Trust :-

People trusts AI more when they understand how decisions are made.

ex: patient trusting AI-Based diagnostic system in hospital.

2. Support accountability :

Transparency allows Developers and companies to explain and justify decision, ensuring responsibility for AI outcomes.

Ex: If a loan is wrongly rejected, the bank must explain the AI's reasoning.

3. Enables Auditing & fairness checks & oversight :

Regulators and auditors can verify the AI model whether it is following rules, ethics and fairness standards.

⇒ It also detects unintended bias or errors in decision making processes.

Ex: Government audit an AI hiring system to ensure no gender Bias.

4. Help Improve AI models :

Clear understanding how the model works helps find mistakes, identify weakness and areas for improvement.

Ex: Developers fix error after seeing why the AI misclassified images.

⇒ Techniques to Achieve Transparency :-

1. Explainable AI [XAI] :-

model gives human - understandable (or) interpretable Explaination for prediction.

Here this model makes AI prediction clear and understandable to human.

Techniques includes :-

feature importance
Heatmaps for images
Rule based Explaination

Ex:- An AI X-Rays system highlights the exact area of the lungs that suggests pneumonia.

2. Documentation :- maintain clear records of

→ model design
→ Data sources
→ assumption and limitations

microsoft and google publish model card showing

→ what dataset was used
→ Accuracy
→ Bias risks
→ Limitations

Doctors (or) auditors can review these documents

3. Open Algorithms and Code :-

making AI models, code, algorithms and datasets open for public review and testing [observation (or) examinations]

Ex: open source AI projects on Github allow

* Researchers
* students
* auditors

to verify how the model works and detect issues.

4. Decision Logging :-

Recording every input, output and reasoning path the AI used during decision making

Ex :- A hiring AI logs : → Candidate score
↔ skill analysis

→ why the candidate was selected or rejected

⇒ If someone challanges the results, the log explains the decision.

Examples of Transparency AI System :-

1. Credit Scoring

AI Explains why a loan was rejected

→ low income
→ Hights existing debt
→ Low Credit Score

2) Hiring System :- AI Explains why candidate was rejected :
→ Low Grades
→ Less Communication skills
→ Age Exceeded d over age.

Challenges to Transpaency :-

1. Complex Models :- Deep learning models have millions of parameters, making it difficult to understand how decision are made.

2. Accuracy vs interpretability Trade off :- High accuracy models are often less interpretable and highly interepretable models are sometimes less accurate.

3. propriety System :- Companies keep their algorithm secret to protect intellectual property, reducing transparency.

4. Data privacy Concerns :- Explaining decision must not reveal sensitive or personal infdmation

5. Lack of standardization :-

No ceniversal standard & rules for how to document & explain AI decision.

→ Best practices of Transparent AI :-

→ use Explainable AI (XAI) techniques & feature importance, heat maps, rule-based methods

→ maintain documentation :-

Datasheets, model cards, system design notes.

→ Use auditable decision logs :-

Every AI action should be trackable

→ Combine Transparency with privacy techniques

Explain decision, without exposing personal data.

→ Communicate AI decision Clearly :-

Explain model output in non-technical language.

⇒ Accountability :-

Accountability in AI means the people & organizations who create and use AI must take responsibility for the AI's decision, action and impact.

→ It ensures some is answerable when the AI makes a mistake, behaves unfairly & causes harm.

Ex: Self driving Car :-

If an autonomous Car causes an accident the car manufacturer & software company is responsible.

HealthCare AI :- If AI misdiagonsis a patient the hospital and AI developers must answer why it happened.

⇒ Why Accountability is important :-

Accountability is important because it makes sure someone is responsible when AI makes a mistake causes harm or behave unfairly.

1. Clarifies Responsibility :-

Determines who is liable if AI causes and of harm.

Ex: A self driving car hits a person → responsibility is checked between manufacturer, developer, and operator.

2. promotes Ethical AI Deployment :-

Encourages Companies to use AI safely and follow fairness guidelines.

Ex: Hiring AI must select candidate fairly, if not company must fix it.

2. Supports Legal Compliance :-

Ensures AI follows laws GDPR (General data protection Regulation), EU ai act (European union Artificial intelligence act etc.

Ex: Banks using AI for loan approval must follow data protection rules.

4. Enhance Trust :- Users trust AI systems more when someone is accountable for decisions.

Ex: patient trust hospital AI tool if hospital are answerable for mistakes.

5. Protect Organization for big losses :-

Accountability reduces like legal fines & reputation damages.

⇒ Strategies to Ensure Accountability :-

1. Assigning AI Governance roles :-

Organization appoints responsible people/ teams. Ex: Companies hire "AI Ethics officers & model Auditors.

2. Logging & monitoring AI Decision :-

Tracks all AI inputs, output and decision paths.

Ex: loan approval AI stores the reason why a customer was rejected.

3. Auditing AI models Regularly :-

perform internal & external checks

Ex: A company reviews to face recognition AI to detect towards darker skin tones

4. Documenting AI Development & Deployment :-

maintain clear documentation of data sources, model assumption and limitations

Ex: A medical AI keep documentation of training data and accuracy before deploying in hospital.

challenges :-

1- Complexity of AI system :-

Deep learning model are black box, hard to trace why they made a decision.

Ex: A neural network rejects a loan, but the bank can't pinpoint the exact rule.

2. shared Responsibility :-

many parties are involved (developers, companies, regulators).

3. Autonomous system :- AI system act independently, creating continuous in assigning responsibility.

4. Dynamic learning system :- AI learns and changes over time, new behaviour may not match diginal testing.

Security :- Security in AI means protecting AI systems from attacks, misuse, & unauthorized access. so that the system work safely and without manipulations.

security = protecting AI data + model + system from hackers & wrong changes.

why security is important :-

1. protecting sensitive Data :-

→ AI uses personal, financial & medical data
→ security ensures this data doesn't get leaked & stolen.

2. maintaining model inlegrity :-

→ Attackers may try to change the AI model to give wrong result.
→ security keeps the model accurate and safe.

3. Ensuring system Availability :-

→ AI systems should work all the time

→ Attacks like DDOS (Distributed Denial of service) can shut them down.

4. preventing Adversarial Exploits :-

→ small changes in input can fool AI system

→ security protects the system from such tricks.

* Common threats to AI Security :-

1. Adversarial Attacks :-

⇒ Hackers change input slightly to confuse the AI.

2. Data poisoning :-

⇒ Bad data is added to training dataset, so model give wrong data (or) results.

3. model Theft (or) model Inversion :-

⇒ Hackers steal the AI model & reconstruct sensitive data from it.

4. Unauthorized Access :-

Accessing & changing the system without permission.

5. Evasion attack :- Try to bypass AI detection.

* strategies to improve Security :-

1. Robust training :-

⇒ Train AI with many examples so it becomes strong against attacks

2. Validating & Testing :-

⇒ check the model regularly and test how it behave under attacks

3. Data Security :-

→ Encrypt the data so nobody can read it without permission.

(or)

→ use Encryption to keep data safe so no one can read or steal it.

4. Access Control :-

⇒ Give permission only to trusted people to use or change the AI system.

5. Continuous Monitoring :-

⇒ Keep watching the AI by ~~unusual~~ unusual or suspicious activities.

6. Model Hardening :-

⇒ use special methods like adversial training to make AI stronger aganist attacks.

7. Regular Security Audits :-

⇒ security experts check for weaknesses and fix them.

⇒ Challenges :-

* AI needs Continuous update :-

Attackers keep finding new tricks, so AI security must always be updated this is challenging

* Lack of security knowledge :-

Not all developers understand AI security fully. This makes system weaker and easy to attack.

⇒ Privacy :- Privacy in AI means protecting peoples personal data when AI is collecting, storing & using it.

AI should not reveal a persons identity or sensitive details.

The main goal is to keep data safe, secure and anonymous.

* why privacy is important :-

→ protect people from identity theft, misuse & discrimination

→ Built trust in AI systems.

* Techniques :-

1. Data Anonymization :- Remove personal details

Data pseudonymization :- Replace real data with fake/temporary IDs.

2. Federated Learning :-

→ Data stays on user's device.

→ only the model updates are send to the server.

→ The original data never leaves the device.

3. Differential privacy :-

→ Add random noise to hide individual information

→ Even if someone tries to attack, they cannot identify a person.

⇒ Inclusivity :- Inclusivity in AI means designing AI system that works fairly for all types of people. - different ages, genders, languages, cultures, disabilities and backgrounds.

AI should not favour one group and should understand everyone properly.

Importance :-

1. Reduce Bias and Ensure Equitable access
2. Improves accuracy – AI works better for all types of users.

Techniques:-

1. Use Diverse and Representative Datasets

Train AI with data that includes different genders, ages, languages, accounts skintones, and backgrounds

→ This Ensure AI works well for everyone

2. Regular testing on Different user Groups:-

→ Test the AI on multiple demoglaphic groups and checks if someone of someglroup is getting wlong results.

→ fix system whenever bias appear.

⇒ Robustness :-

Robustness in AI means the ability of an AI system to work correctly even in difficult, unexpected or noisy situations.

⇒ Robust AI should not easily get confused & give wrong results.

Importance :-

1. Ensure safe operations in critical areas [self-driving cars, healthcare]
2. Reduce errors caused by noise, bad data, & Environmental changes
3. Protect AI from malicious inputs (attacks)

Techniques :-

1. Train with Adversarial / Difficult Examples

Give the AI challenging, noise, on purposely confusing inputs during training.

This makes the model stronger.

2. Add Error - Handling and fall back mechani

Add backup systems so AI still works wh
something goes wrong.

⇒ Sustainability :-

Sustainability in AI means developi
AI systems in a way that uses less energy, les
computing power, and has low Environmental
impact.

Goal → Make AI Energy — Efficient and
Environment friendly.

Importance :-

1. AI training uses huge Energy (Gpu power,
data centers)

2. Reduces Carbon footprint and Environmental
damage.

3. Encourages responsible and green AI
development.

Techniques :-

1. Optimize Models for lower Energy use :-

Use efficient algorithms and models that require less compulations and less power.

2. Use Green / Renewable - Energy Data Centers :-

Train AI in data centers powered by solar, wind, or hydro energy instead of coal based power

⇒ Reliability :- Reliability in AI means the AI system should give correct, consistent, and dependable result every time, not just sometimes.

→ It should work properly even if you run it again and again.

Importance :-

1. users can trust the system.
2. AI gives same ~~content~~ correct output under similar conditions
3. Reduces failures in critical fields (healthcare, banking, transport)

Techniques :-

1. Regular Testing & Validation :-

Test the AI with many different datasets to check if it always gives stable and correct results.

→ Helps identify errors early.

2. Monitoring & updating the model :-

- Continuously monitor AI performance and updates when accuracy decreases.

→ keep the model reliable overtime.