Faculty Development Programme
ON
Artificial Intelligence and Data Science:
Foundations, Pedagogy, Tools and Emerging Research Trends

Ethical AI and Bias in Machine Learning: Implications for Teaching and Research

Dr. Sudhakar Mishra
Asst. Professor
Department of Artificial Intelligence
SVNIT, Surat

Learning Objectives

1. Define ethical AI and distinguish between different types of bias in machine learning.
2. Identify sources and manifestations of bias across the ML lifecycle
3. Analyze ethical implications of biased AI systems in academic teaching contexts.
4. Examine ethical considerations in research design, methodology, and publication.
5. Design and teach AI ethics as an integral part of ML curricula.

Real-life examples of unfair AI practices

1. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) race bias with reoffending rates: More likely to say black defendants were at risk of reoffending than their white counterparts
2. US healthcare algorithm underestimated black patients’ needs.
3. Uber users have complained that they pay more for rides if their smartphone battery is low, even if officially, the level of a user’s smartphone’s battery does not belong to the parameters that impact Uber’s pricing model.
4. AI avatar app produced sexualized images of women: The AI avatar app Lensa came under scrutiny for its biased outputs. While male users received diverse, professional avatars depicting them as astronauts or inventors, women often got sexualized images.
5. AI facial recognition leads to wrongful arrest of innocent man.
6. HireVue’s AI video interview platform fails to understand lesser-abled candidates

And many more….

Working Definition of Ethical AI

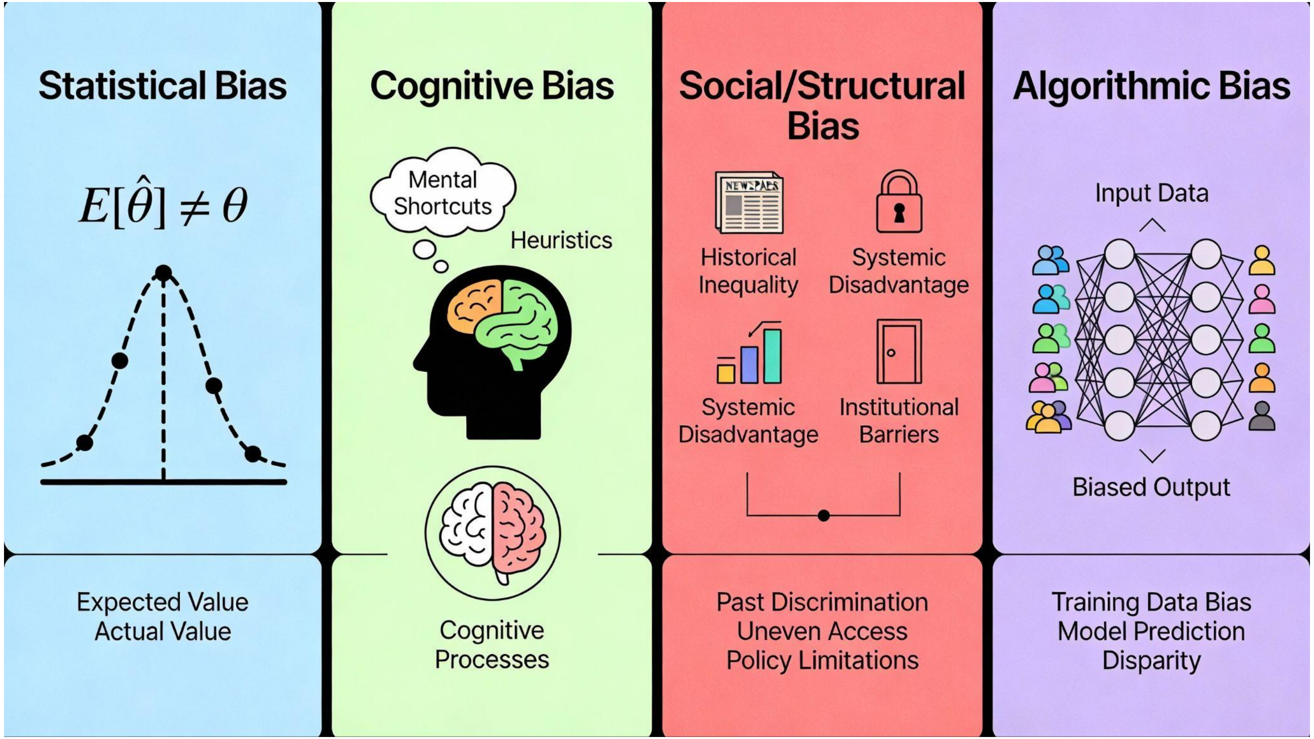
Ethical AI is the design, development, deployment, and evaluation of AI systems in ways that:

1. Minimize harm (physical, financial, psychological, reputational, structural)
2. Respect human rights, dignity, autonomy, and agency
3. Ensure fair distribution of benefits and burdens
4. Promote transparency and accountability
5. Remain responsive to affected communities and broader social contexts.

“Bias” in Machine Learning: Clarifying the Term

Bias is overloaded. We must distinguish several meanings:

1) Statistical Bias, 2) Cognitive Bias, 3) Social/Structural Bias, 4) Algorithmic Bias



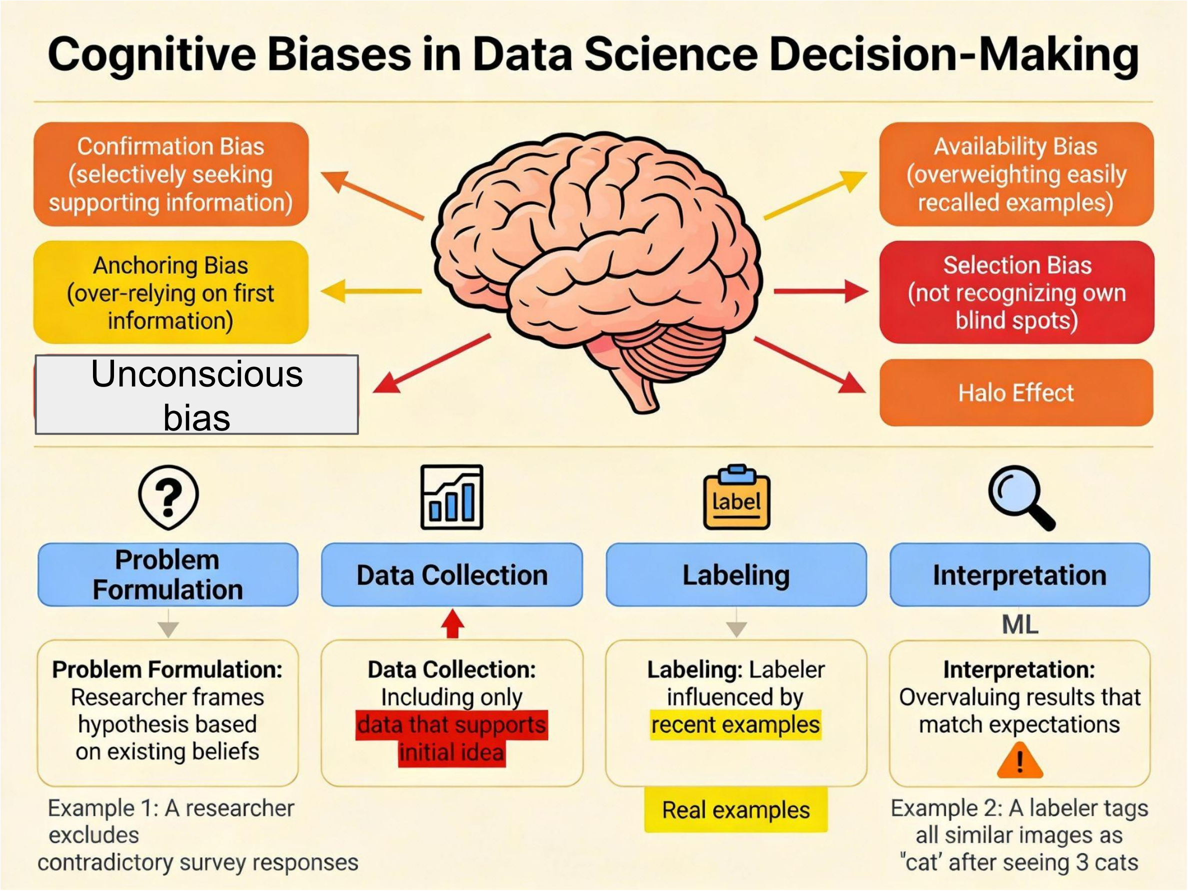
Statistical Bias

Definition: A biased estimator has an expected value that differs from the true parameter.

Formula: $bias(\hat{\theta}) = E[\hat{\theta}] - \theta$

It is not unethical unless it systematically underestimates or overestimates a critical quantity. For example, underestimating disease risk in certain populations. Like, NCD in rural population.

Cognitive Bias



Cognitive Biases in Data Science Decision-Making

Confirmation Bias (selectively seeking supporting information)

Anchoring Bias (over-relying on first information)

Unconscious bias

Availability Bias (overweighting easily recalled examples)

Selection Bias (not recognizing own blind spots)

Halo Effect

Problem Formulation: Researcher frames hypothesis based on existing beliefs

Data Collection: Including only data that supports initial idea

Labeling: Labeler influenced by recent examples

Real examples

Interpretation: Overvaluing results that match expectations

Example 1: A researcher excludes contradictory survey responses

Example 2: A labeler tags all similar images as "cat' after seeing 3 cats

Cognitive Bias

Definition: Systematic patterns in how humans process information and make decisions.

Three types of Cognitive Biases:

- Confirmation bias (seeking information confirming prior beliefs)
- Anchoring bias (relying too heavily on first piece of information)
- Availability bias (overweighting easily recalled examples)
- Unconscious bias: automatic, ingrained stereotypes and attitudes (favorable or unfavorable) that affect our perceptions and decisions without our awareness.

Example:

An emotional stimuli is assigned a particular emotion assuming that participant will feel the same emotion is a confirmation bias.

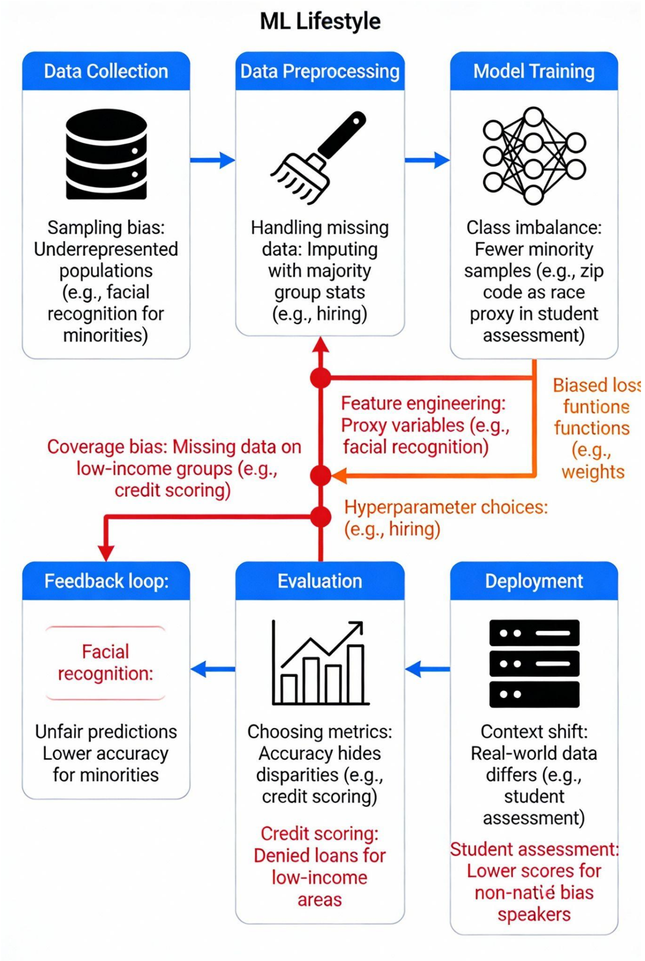
If I give a scale of 0 to 100 to rate any emotion intensity and ask participants to rate happiness less than or greater than 100, the participant will be biased more towards higher side of the scale.

Emotion which are easy to capture for example, happy and sad, basic emotion categories. There is less research on complex emotion category.

Social/Structural Bias

- Definition: Systematic, institutionalized patterns of advantage and disadvantage that map onto social categories (gender, race, caste, disability, language, regions, religion, sexual orientation, socioeconomic status, etc.)
- Origin: Historical inequalities, power imbalance, and institutional practices–not from individual prejudice.
- Example: Identical infant emotional responses are often labeled as "anger" when the infant is perceived as a boy, but "fear" when perceived as a girl.
- When ML models are trained on structurally biased data, they can amplify and legitimize that bias at scale, under the guise of “objectivity”. Example: In the last decade women were not recruited by IAF as pilot.

Algorithmic Bias



ML Lifestyle

Data Collection

Sampling bias: Underrepresented populations (e.g., facial recognition for minorities)

Data Preprocessing

Handling missing data: Imputing with majority group stats (e.g., hiring)

Model Training

Class imbalance: Fewer minority samples (e.g., zip code as race proxy in student assessment)

Feature engineering: Proxy variables (e.g., facial recognition)

Biased loss funtione functions (e.g., weights

Coverage bias: Missing data on low-income groups (e.g., credit scoring)

Hyperparameter choices: (e.g., hiring)

Feedback loop:

Facial recognition:

Unfair predictions Lower accuracy for minorities

Evaluation

Choosing metrics: Accuracy hides disparities (e.g., credit scoring)

Credit scoring: Denied loans for low-income areas

Deployment

Context shift: Real-world data differs (e.g., student assessment)

Student assessment: Lower scores for non-natiė bias speakers

Algorithmic Bias

- Definition: Systematic, unjustified, or unfair differences in how an ML model treats or performs across different groups or individuals.
- It can arise from:
  - Biased, unrepresentative, or incomplete training data
  - Proxy variables: using features that correlated with protected attributes (e.g., zip code as proxy for race)
  - Model choice: some architectures may be more sensitive to minority-class examples due to some rules or assumptions by architect of the model. For example: racialized anger bias

Comparison among different biases

Statistical bias ≠ Algorithmic bias ≠ Cognitive bias ≠ Social bias (Though, they interact)

Type	Nature	Fixability	Ethical Status
Statistical	Technical property	Immediate (swap estimators)	Not inherently unethical
Cognitive	Human mental shortcuts	Short-medium term	Not inherently, but problematic when unchecked
Social/Structural	Systemic inequality	Long-term (decades/centuries)	Fundamentally unjust
Algorithmic	Emergent from ML pipeline	Medium-long term	Violates fairness principles

Why Bias Matters “Especially” in Academia?

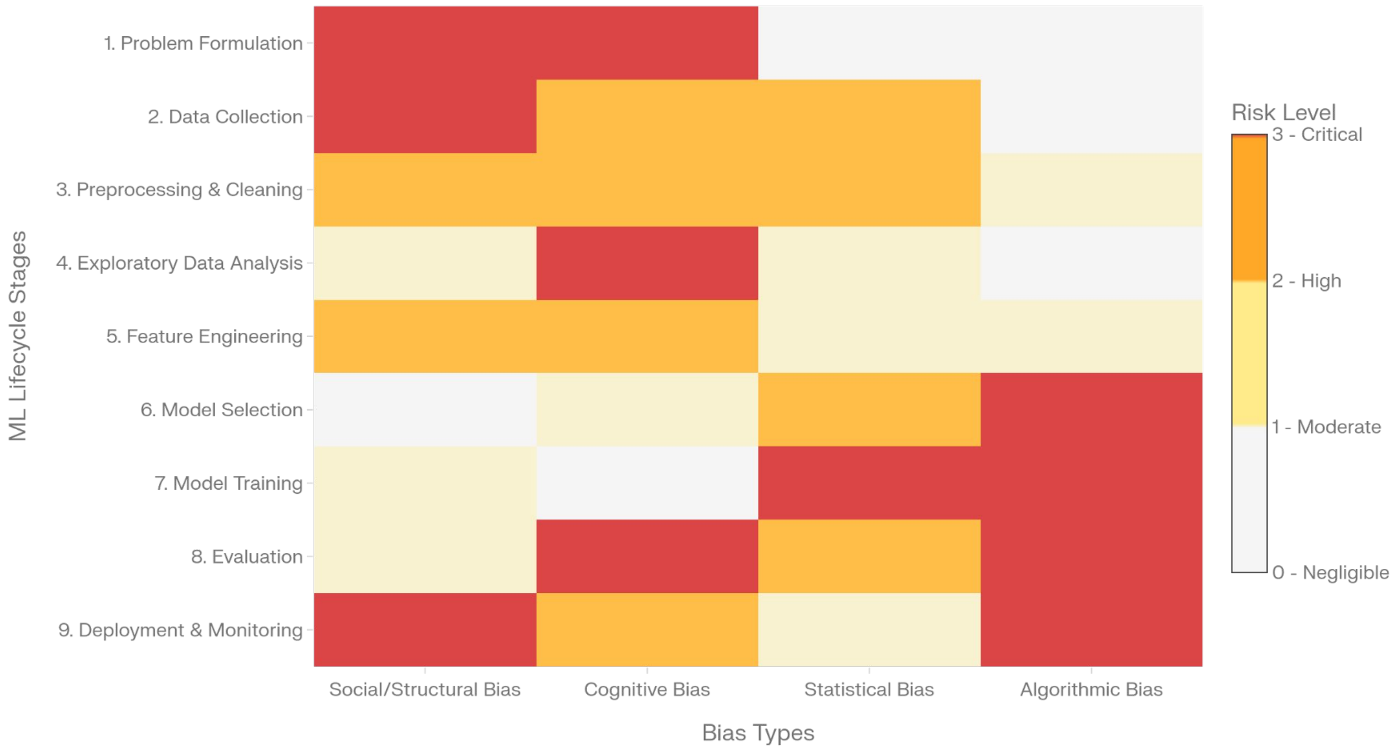
Because of Critical Roles:

1. As Designers of AI systems: Biased systems can amplify existing inequalities (e.g., students from underrepresented groups systematically flagged as “at-risk” or “low-engagement”).
2. As Teachers and Educators: Your curriculum choice affects whether students see bias mitigation as central or marginal to ML practice.
3. As Researchers: Research in academia influences industry and policy; biased research has consequences beyond your institution.

The ML Lifecycle and Where Bias Enters

Bias Risk Levels Across ML Lifecycle Stages

Critical risks concentrated in early and late stages



The ML Lifecycle and Where Bias Enters

1. **Social/Structural Bias (First Column):**
   - **Critical (Level 3)** at **Problem Formulation** (framing the problem from a dominant perspective) and **Data Collection** (historical discrimination embedded in data).
   - It re-emerges at **Deployment** (Level 3) because that is where the societal harm actually occurs.

2. **Cognitive Bias (Second Column):**
   - **Critical (Level 3)** at **Problem Formulation** (assumptions) and **EDA** (Exploratory Data Analysis).
   - *Why EDA?* This is where **Confirmation Bias** is rampant—researchers tend to "find" patterns that match their prior beliefs and stop looking once they find them.

3. **Statistical Bias (Third Column):**
   - **Low** in early stages.
   - **Critical (Level 3)** at **Model Training** because this is where mathematical estimation errors (variance-bias trade-off) occur.

4. **Algorithmic Bias (Fourth Column):**
   - **Low/None** in early stages (it doesn't exist yet).
   - **Critical (Level 3)** at **Model Selection** (choosing a model with poor inductive bias for the minority group), **Training**, and **Evaluation** (metric selection masking disparities).

The ML Lifecycle and Where Bias Enters

Stage 1: Problem Formulation

Stage 2: Data Collection

Stage 3: Data Preprocessing & Feature Engineering

Stage 4: Model Selection & Training

Stage 5: Evaluation & Validation

Stage 6: Deployment & Monitoring

The ML Lifecycle and Where Bias Enters: Stage-1

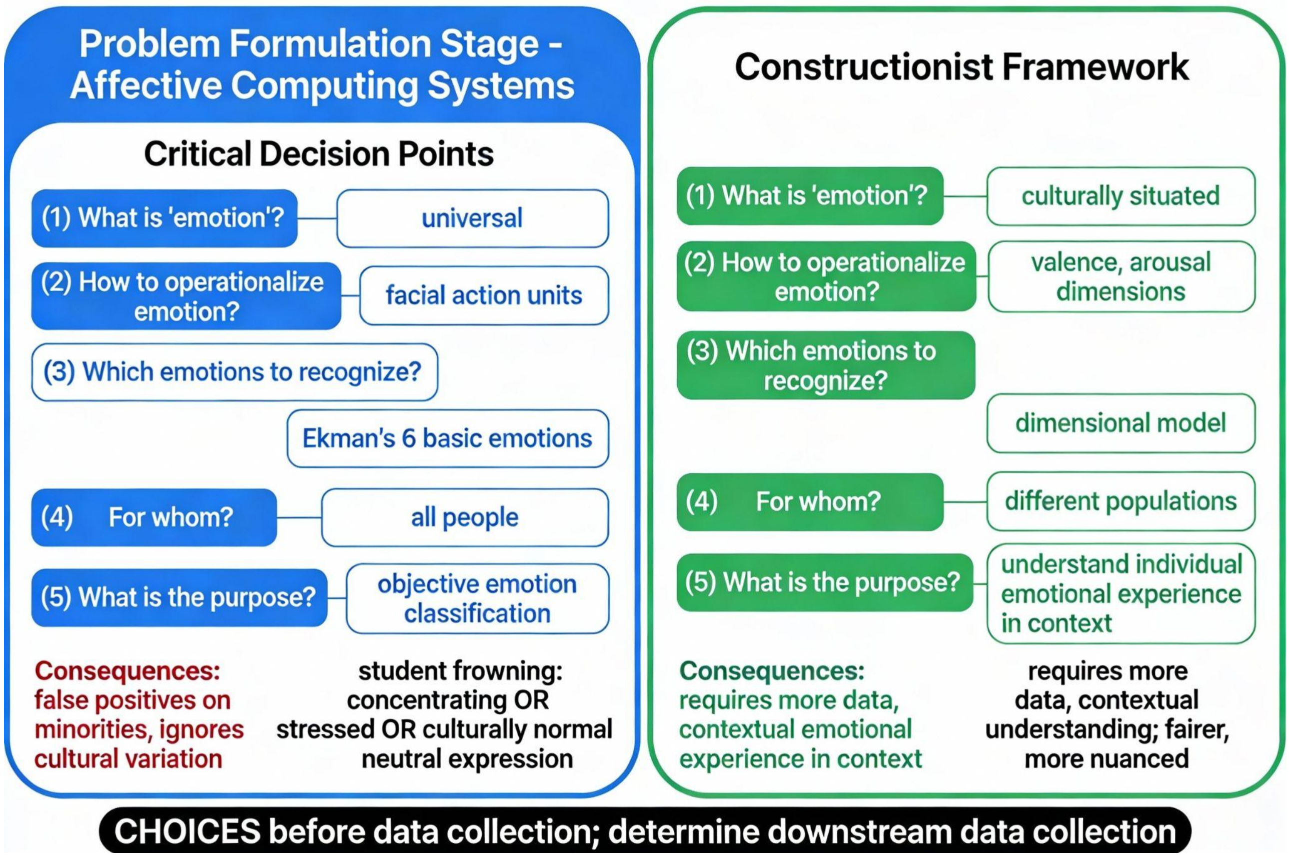
Problem Formulation: It is Construct Validity Bias / Theoretical Bias

- What is the problem we are solving? Who decided?
- For whom? Whose needs are prioritized; whose are invisible?
- What are we optimizing? Which outcomes count as “success”?

Example:

1. We will classify student faces into one of Ekman’s 6 Basic Emotions (Happy, Sad, Angry, Fear, Surprise, Disgust) + Neutral. We assume that 'Happy/Surprise' correlates with Engagement and 'Neutral/Sad' correlates with Boredom/Disengagement." Problems with following assumptions:
    a. Essentialism
    b. Cultural Bias: Considering FACS (Facial Action Coding System) developed on Western population.

The ML Lifecycle and Where Bias Enters: Stage-1



Problem Formulation Stage - Affective Computing Systems

Critical Decision Points

(1) What is 'emotion'? — universal

(2) How to operationalize emotion? — facial action units

(3) Which emotions to recognize? — Ekman's 6 basic emotions

(4) For whom? — all people

(5) What is the purpose? — objective emotion classification

Consequences: false positives on minorities, ignores cultural variation

student frowning: concentrating OR stressed OR culturally normal neutral expression

Constructionist Framework

(1) What is 'emotion'? — culturally situated

(2) How to operationalize emotion? — valence, arousal dimensions

(3) Which emotions to recognize? — dimensional model

(4) For whom? — different populations

(5) What is the purpose? — understand individual emotional experience in context

Consequences: requires more data, contextual emotional experience in context

requires more data, contextual understanding; fairer, more nuanced

CHOICES before data collection; determine downstream data collection

The ML Lifecycle and Where Bias Enters: Stage-1

How This Bias Cascades (The Consequence):

Because the problem was formulated as "Detect the 6 distinct facial shapes," the rest of the lifecycle is poisoned:

- Data Collection (Stage 2): The team collects data by asking actors to "make a happy face" or "make a bored face." This results in exaggerated, stereotypical data that lacks real-world validity.
- Annotation (Stage 3): Annotators are told to label images as "Engaged" or "Not Engaged." They unknowingly project their own cultural biases onto the students' faces (e.g., rating a Black student's resting face as "angry" or "bored" due to racial stereotypes).
- Evaluation (Stage 4): The model is tested on similar posed data and achieves 98% accuracy.
- Deployment (Stage 5): The system flags a brilliant but introverted student (or a student from a culture that values stoicism) as "disengaged" simply because they aren't smiling at the camera.

The ML Lifecycle and Where Bias Enters: Stage-1

Component	Description
Context	Automated Student Engagement Detection
Bias Type	Construct Validity Bias / Theoretical Bias
Life Cycle Stage	Problem Formulation (defining the target variable)
The Error	Operationalizing "Engagement" as "Ekman's Basic Emotions" (Smile = Engaged).
Ethical Impact	Invalidates diverse ways of learning; penalizes students with non-standard facial expressions (cultural differences, neurodivergence, or simply introversion).

The ML Lifecycle and Where Bias Enters: Data Collection

- Coverage Bias: Not all relevant populations are represented (emotion recognition trained only on “WEIRD” populations).
- Measurement Bias: How are labels assigned? Labels are human judgments, not objective ground truth.
- Temporal Bias: Data collected at one time may not generalize to another (e.g., online behaviour during COVID).
- Selection Bias: Systematic differences in who/what gets included in the dataset (e.g., hospitals collecting data from only their patient population, who may differ systematically from the general populations).

Coverage Bias (Representation Bias)

Dataset composition (typical affective computing benchmark):

- 80% White/European faces
- 70% ages 20-40
- 60% male
- 90% well-lit, front-facing
- Actors/posed expressions: 70%

Real deployment population:

- Global university students (50% Asian, 20% African)
- Ages 15-60
- Variable lighting, angles, spontaneous expressions

Result: Model achieves 85% accuracy on dataset → 45% accuracy on real diverse students

For example,

Model learns: "Anger = furrowed brows + pursed lips" (Western expression)
Real Asian student: Furrowed brows (concentration) → Misclassified as "angry/disengaged"
Real autistic student: Flat affect (focus) → Misclassified as "sad/bored"

Measurement Bias

Measurement Method	Bias Introduced	Affected Groups
Webcam (lab lighting)	Lighting bias	Dark-skinned individuals
Self-report surveys	Social desirability	Cultures valuing emotional restraint
Acted emotions	Performance bias	Spontaneous emotion expression
Single modality (face)	Modality bias	Hearing-impaired, face-masked

Real Life Example:

Collection method: Lab setting, actors asked to "show anger"
Technical issue: High-quality lab lighting optimized for light skin tones
Result: Darker-skinned actors' facial muscles less visible → poor feature extraction
Model consequence: Learns light-skinned anger expressions well, dark-skinned poorly

Stage-3: Data Preprocessing & Feature Engineering

- **Missing Data Bias:** If certain groups have systematically missing values, imputation strategies may bias the data. For example, *Anger, complex emotions: embarrassment, envy and so on.*
- **Aggregation Bias:** Combining disparate data sources with different standards or definitions. For example, DEAP dataset with recording done on two different sites.
- **Feature Construction Bias:** Creating features that inadvertently encode sensitive attributes. For example, *using zip code as a feature which can associate with the location of minority population or underprivileged population*.
- **Class Imbalance:** For example there are certain categories of emotions which are hard to elicit in the lab setting. For example, *one collects the data with dominant male participation than female due to noise in data from female participants.*

Stage-4: Model Selection & Training - Algorithmic Bias

Algorithmic bias is not caused by the algorithm itself, but by how the data science team collects and codes the training data. Specific causes include:

- Biases in training data
- Biases in algorithm design
- Biases in proxy data
- Biases in evaluation

Biases in training data

- **Flawed data** is characterized as **non-representative**, lacking information, historically biased or otherwise “bad” data.
- It leads to algorithms that produce unfair outcomes and amplify any biases in the data.
- For example, data with low representation from underrepresented communities/sections of the society. Creating an student engagement system on WEIRD population and deploying it to South Asian population.

Biases in algorithm design

- Developers might embed the algorithm with subjective rules based on their own conscious or unconscious biases.
- Algorithmic design biases occur when AI systems reflect and amplify societal prejudices, often stemming from biased training data or flawed design, leading to unfair outcomes.
- For example hiring (Amazon's gender-biased tool), criminal justice (COMPAS recidivism scores unfairly flagging minorities), and facial recognition (lower accuracy for darker skin tones).

Biases in Proxy Data

- AI systems sometimes use proxies as a stand-in for protected attributes, like race or gender.
- However, proxies can be unintentionally biased as they might have a false or accidental correlation with the sensitive attributes they were meant to replace.
- For example, if an algorithm uses postal codes as a proxy for economic status, it might unfairly disadvantage certain groups where postal codes are associated with specific racial demographics.

Biases in Evaluation

- Biases in evaluation occur when algorithm results are interpreted based on the **preconceptions of the individuals** involved, rather than the objective findings.
- Even if the algorithm is neutral and data-driven, how an individual or business applies the algorithm’s output can lead to unfair outcomes depending on how they understand the outputs.

Real-World Examples of Algorithmic Biases

Algorithmic bias can occur in any scenario or sector that uses an AI system to make decisions. For example,

- Bias in the criminal justice system
- Bias in healthcare
- Bias in recruitment
- Bias in financial services
- Bias in facial recognition systems

Stage-5: Evaluation & Validation

- **Metric selection bias:** Different fairness metrics can lead to different conclusions.
    - Fairness in AI means ensuring AI systems treat **all individuals and groups equitably**, without bias or discrimination, leading to just and impartial outcomes, which involves proactively designing against historical biases in data and algorithms to prevent harm
- **Evaluation set bias:** If your test set doesn’t represent the deployment population, evaluation is meaningless.
- **Underreporting:** Focusing on aggregate performance while not disaggregating by group.

Metric selection bias

Metric selection bias is a form of evaluation bias where:

- There are many valid fairness metrics (demographic parity, equalized odds, accuracy, equal opportunity, etc.).
- The evaluator chooses:
  - Only a subset of these metrics, and
  - Often after seeing model performance,
- Then claims “the model is fair” based on the one metric on which the model looks good. This is analogous to **p-hacking** in statistics: you don’t change the model, you change the yardstick.

In affective computing or any ML system, this means:

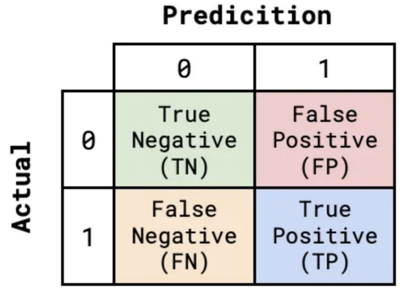
- One fairness report might say: “Our engagement detector is fair” (because demographic parity looks good).
- Another, using a different metric on the same model, might say: “The system is unfair and harms introverted / minority students” (because error rates are very different).

Both statements can be mathematically correct, but they reflect different fairness notions.

Fairness Metrics: Equal Opportunity

- The term "equal opportunity" is used for **parity in the True Positive Rate (TPR)** because it directly relates to ensuring that all qualified individuals, regardless of their sensitive attributes (like race, gender, etc.), have the same chance of receiving a positive outcome or "opportunity".

$$TPR = TP/(TP+FN)$$



		Predicition	
		0	1
Actual	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

- Remember, we are assuming a positive prediction will lead to some benefit.
- This means
  - denominator can be seen as the number of people who **should benefit** from the model.
  - The numerator is the number who **should and have benefited.**
  - So, TPR can be interpreted as the percentage of people who have **rightfully benefitted** from the model.

Fairness Metrics: Equal Opportunity

	1	0	Ratio
Race	83.9%	89.3%	1.07
Sex	81.0%	92.1%	1.14

Accuracy

$$TPR_0 = TPR_1 \quad (1)$$

$$TPR_1 - TPR_0 < \text{Cutoff} \quad (2)$$

$$\frac{TPR_0}{TPR_1} > \text{Cutoff} \quad (3)$$

	1	0	Ratio
Race	61.1%	53.3%	0.87
Sex	63.2%	44.3%	0.70

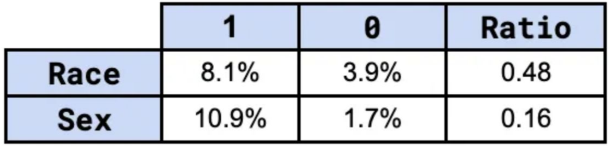
TPR Based

Fairness Metrics: Equalized Odds

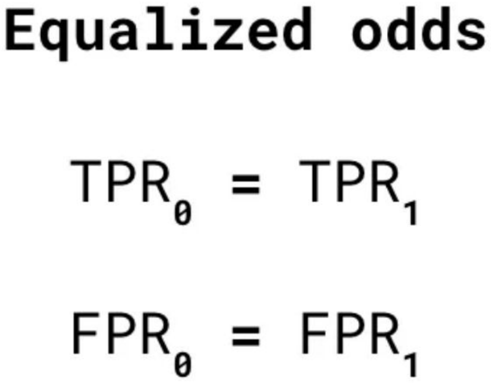
- The concept of "equalized odds" is a stricter fairness metric that requires parity in both the True Positive Rate (TPR) and the False Positive Rate (FPR) across groups. The name "equalized odds" is used because it ensures the model performs equally well (or equally poorly) for all possible actual outcomes (both positive and negative ground truths).

$$\text{FPR} = \text{FP}/(\text{FP} + \text{TN})$$

FPR is the percentage of actual negatives incorrectly predicted as positive. This can be interpreted as the percentage of people who have wrongfully benefited from the model.



	1	0	Ratio
Race	8.1%	3.9%	0.48
Sex	10.9%	1.7%	0.16



Equalized odds

$$\text{TPR}_0 = \text{TPR}_1$$

$$\text{FPR}_0 = \text{FPR}_1$$

- Table shows that a higher percentage of the privileged group has wrongfully benefited from the model.

Stage-6: Deployment & Monitoring

- Context shift: A fair model in training context may become unfair when deployed to a different population or distribution.
- Feedback loops: Biased decisions feed back into future training data, amplifying bias over time.
- Blind deployment: No ongoing monitoring of fairness metrics post-deployment

Ethical Principles for ML Research

- Principle-1: Transparency & Reproducibility
- Principle-2: Disaggregated Evaluation
- Principle 3: Diverse, Representative Sampling
- Principle 4: Stakeholder Engagement
- Principle 5: Responsible Dissemination
- Principle 6: Ongoing Monitoring & Accountability

Principle-1: Transparency & Reproducibility

- Clearly describe your sample: demographics, recruitment method, inclusion/exclusion criteria, any systematic biases in access.

- Provide data and code (with appropriate privacy protections) so others can audit and reproduce.

- Document design choices: Why this metric? Why this threshold? Why this test?

- **Why:** Enables external audit and replication by diverse researchers who might catch what you missed.

Principle-2: Disaggregated Evaluation

- Always report results separately for demographic subgroups (gender, race, age, language, disability, etc.), not just overall accuracy.

- Use multiple fairness metrics; don't just report one.

- Report failures, edge cases, and limitations, not just successes.

- **Why:** Hides bias and enables others to understand where the model works and where it fails.

Principle 3: Diverse, Representative Sampling

- If you want to generalize, sample from the population you're making claims about.

- If sampling is convenient (students, crowdworkers), be transparent about limitations and don't over-generalize.

- Actively seek diverse perspectives and participants; don't treat diversity as optional.

- **Why:** Enables the model and findings to generalize without bias.

Principle 4: Stakeholder Engagement

- Who is affected by this research and the models it produces? Talk to them.

- Involve affected communities in problem formulation, design, and interpretation.

- Seek informed consent; explain implications; allow people to opt out.

- **Why:** Communities understand contexts and potential harms that researchers might miss; inclusion is ethical in itself.

Principle 5: Responsible Dissemination

- Communicate limitations, not just strengths. What does the model not work on?

- Discuss potential harms and misuse, not just benefits.

- **Avoid sensationalism:** "AI system achieves human-level performance" (on a narrow task, in lab conditions, compared to untrained humans...).

- Consider who benefits from your research being true and how that might bias your interpretation.

- **Why:** Prevents misuse and over-interpretation of research.

Principle 6: Ongoing Monitoring & Accountability

- If your research leads to deployed systems, commit to monitoring for bias post-deployment.

- Create mechanisms for feedback and rapid response if bias is detected.

- Don't just publish and walk away.

- **Why:** Ensures that idealized, lab-validated fairness translates to real-world fairness.

Teaching Ethical AI: Pedagogical Strategies

Strategy-1: Integrate ethics throughout, not as an add-on

Strategy-2: Use diverse, real datasets

Strategy-3: Teach fairness metrics & trade-offs

Strategy-4: Conduct bias audits as assignments

Strategy-5: Teach stakeholders perspectives

Strategy-6: Make space for uncertainty & legitimate disagreement

Strategy-7: Teach responsibility, not just competence

Integrate ethics throughout, not as an add-on

- Don’t teach “ML” and then “ML Ethics” separately. Integrate ethical question into every unit.
- When teaching classification: “Which groups might this be unfair to? How would you test?”
- When teaching evaluation metrics: “Why does accuracy alone mislead? We should care about fairness too”.
- Outcome: Students internalize that ethics is “inseparable” from ML practice.

Use diverse, real datasets

- Avoid using only canonical, clean datasets (MNIST, CIFAR-10) where everyone gets the same result.
- Include messy, real data from diverse sources: datasets with known biases, historical datasets data from underrepresented communities.
- Have students audit datasets: “Who is represented? Who is missing? What does that mean for a model trained here?”
- Outcome: Students develop critical eyes for data quality, representativeness, and potential harms.

Teach fairness metrics & trade-offs

- Introduce fairness notions early: demographic parity, equalized odds, individual fairness, group fairness.
- Show that no single metric is universally “right”-different stakeholders have different interests.
- Have students compute and compare fairness metrics on real datasets: “Is the model fair? Depends on your definition and values.”
- Outcome: Students learn that fairness is a design choice, not a technical parameter to optimize blindly.

Conduc Bias Audits as Assignments

- Instead of “build a classifier on this dataset,” ask: “Build a classifier and audit it for bias. What do you find? How would you fix it?”
- Provide tools: bias detection libraries (Fairness Indicators, themis ML (python library for detecting discrimination))
- Require students to disaggregate performance by group, plot results, discuss implications.
- Outcome: Students practice the habit of questioning and auditing models.

Teach Stakeholder Perspectives

- Not just “the developer’s view” but also affected communities, policymakers, subjects, those harmed.
- Case studies: Ask students to write from the perspective of different stakeholders (student flagged as at-risk, parent, teacher, administrator).
- Outcome: Students develop empathy and understand that ML happens in contented, plural value landscape.

Make Space for Uncertainty & Legitimate Disagreement

- Don’t present “the answer” to what’s fair or ethical.
- Encourage debate: “Is this system fair?” Some say yes because it improves accuracy. Other say no because it treats groups differently. What do you think, and why?
- Model intellectual humility: “This is a hard problem. Reasonable people disagree”.
- Outcome: Students develop critical judgment instead of passive rule-following.

Teach Responsibility, Not Just Competence

- Beyond “how to build fair ML,” teach: “What is my responsibility as an ML practitioner?”
- Discuss: Documentation, transparency, consent, resources, ongoing monitoring.
- Why it matters: Even a fair model can be deployed irresponsibly: unfair models can be mitigated through responsible practices.
- Outcome: Students see ethics as part of “professional responsibility”

References

1. Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and Machine Learning. https://fairmlbook.org (free online textbook)

2. Benjamin, R. (2019). Race After Technology: Abolitionist Tools for the New Jim Crow. Polity Press.

3. Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems*, 29.

4. Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency* (pp. 77–91).

5. Dreyfus, S. E., & Dreyfus, H. L. (1980). A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition. (Referenced for stages of learning; applicable to learning ethics as a skill.)

6. European Commission. (2020). Ethics of Artificial Intelligence: Issues and Initiatives. (EPRS Research Briefing)

7. Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, 29, 3315–3323.

8. UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. (Comprehensive international framework)

9. O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown.

10. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68).

Thank you for your attention

Any Question