

Responsible AI

Principles and Practices

Mahsa Dibaji
Graduate Research Assistant
Electrical and Software Engineering Department

March 4th, 2024

Outline

- What is Responsible AI?
- Responsible AI principles
- Algorithmic Bias in ML

Learning Goals

- Learn what is responsible AI and what are the different aspects of it.
- Get familiar with how we can identify biases, evaluate fairness, and mitigate biases to improve fairness of AI models.

What is Responsible AI?

- Responsible Artificial Intelligence (Responsible AI) is an approach to developing, assessing, and deploying AI systems in a safe, trustworthy, and ethical way. AI systems are the product of many decisions made by those who develop and deploy them.
- Responsible AI aims to minimize potential harm to society and environment while pursuing technology innovation.
- Five main principles:
 - Reliability and Safety
 - Privacy and Security
 - Transparency
 - Accountability
 - Fairness

Reliability and Safety

- Reliability in AI emphasizes consistent performance and resilience against potential adversarial threats.
- AI systems should consistently deliver reliable results across varied scenarios.
- Recommendations:
 - Test your model for subgroups in the data: are results consistent?
 - Consider if anyone might have incentive to make the system misbehave, and what would be the consequences?
 - Develop an approach to combat threats: Test the performance of your systems in the adversarial setting

Privacy and Security

- ML models often use sensitive data.
- The sensitive data must be protected against breaches and unauthorized access.
- Especially important in sensitive domains such as healthcare and banking.
- Recommendations:
 - Ethical data acquisition is important (participants consent and control over data).
 - Proper data anonymization and limiting the use of sensitive data.
 - Federated Learning.
 - Consider the privacy impact of how the models were constructed and may be accessed.

Transparency

- In critical applications that impact people's lives, it becomes especially important for people to understand how AI decisions were made.
- Interpretability is an important part!
 - It helps with designing, developing, and debugging models, making sure they work correctly and as intended.
 - It can happen before, during, and after designing and developing the model.

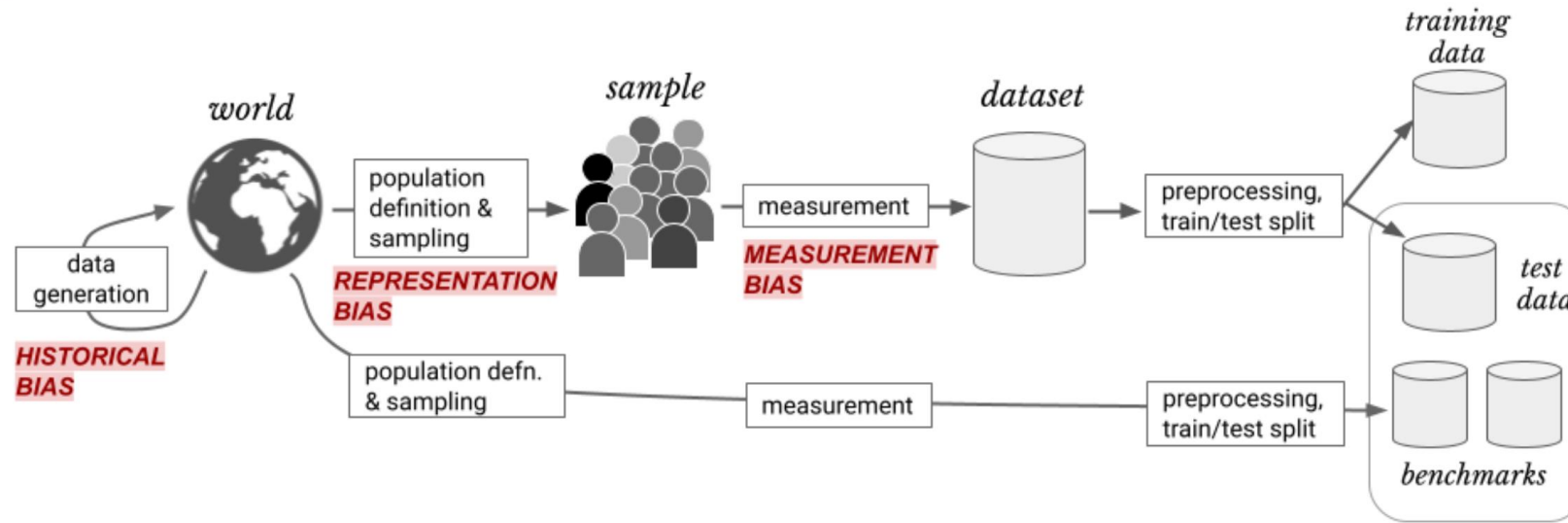
Accountability

- Who would be responsible if AI makes mistakes?
- Humans should have meaningful control over autonomous systems
- Especially critical in sensitive areas like healthcare

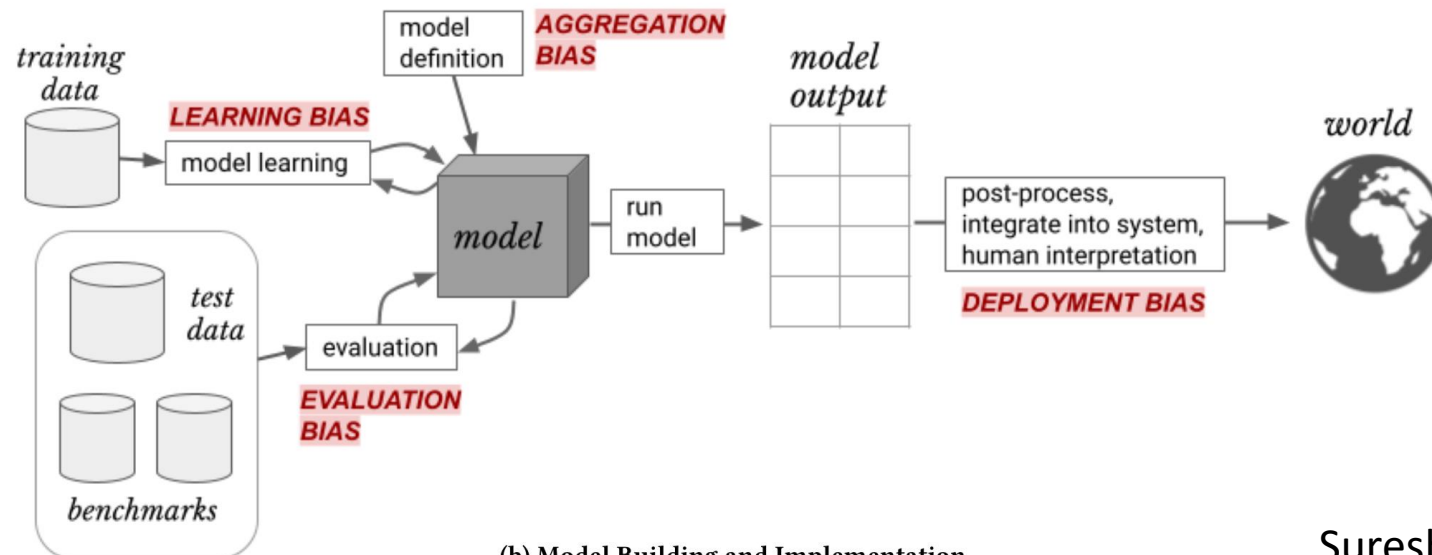
Fairness

- **Fairness** in machine learning: absence of any favoritism towards an individual or group based on their intrinsic/acquired traits for decision-making.
- Decisions made by computers after a machine-learning process may be considered unfair if they were based on variables considered sensitive such as sex, gender, ethnicity, socio-economic status.

Algorithmic Bias



(a) Data Generation



(b) Model Building and Implementation

Suresh et al., 2019.



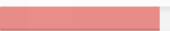















1. Historical Bias

- Bias that already exists in the world for historically disadvantaged or excluded groups.
- Word embeddings trained on Google news articles show and perpetuate gender-based stereotypes!

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai

2. Representation Bias

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0% 	79.2% 	100% 	98.3% 	20.8% 
 FACE++	99.3% 	65.5% 	99.2% 	94.0% 	33.8% 
 IBM	88.0% 	65.3% 	99.7% 	92.9% 	34.4% 



Representation bias occurs when the development sample underrepresents some part of the population, and subsequently fails to generalize well for a subset of the use population.

3. Measurement Bias

- Measurement bias occurs when choosing, collecting, or computing features and labels to use in a prediction problem.
- Typically, a feature or label is a proxy (a concrete measurement) chosen to approximate some construct (an idea or concept) that is not directly encoded or observable.
- Example:
 - Risk assessment with Northpointe's COMPAS.
 - Proxies: "rearrest" to measure "recidivism" or "arrest" to measure "crime".
 - These mapping are different in different demographics: blacks are more policed than whites.

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

4. Aggregation Bias

- Aggregation bias arises when a one-size-fits-all model is used for data in which there are underlying groups or types of examples that should be considered differently.
- Example:
 - Healthcare: For diagnosing and monitoring diabetes, models have historically used levels of Hemoglobin A1C (HbA1c) to make their predictions.
 - Different levels across ethnicities

Racial differences in performance of HbA1c for the classification of diabetes and prediabetes among US adults of non-Hispanic black and white race

[Christopher N. Ford](#),¹ [R. Whitney Leet](#),^{1,2} [Lauren Daniels](#),¹ [Mary K. Rhee](#),³ [Sandra L. Jackson](#),⁴ [Peter W. F. Wilson](#),⁵
[Lawrence S. Phillips](#),³ and [Lisa R. Staimez](#)¹

5. Learning Bias

- Modeling choices might amplify performance disparities across different samples in data
 - Objective/Loss function: prioritizing one objective could damage another objective.
- Example:
 - Compact models amplify performance disparities on underrepresented data since the model learns to preserve most frequent features.

Characterising Bias in Compressed Models

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, Emily Denton

6. Evaluation Bias

- Evaluation bias happens when the benchmark data doesn't represent the whole target population.
- Benchmarks are used for comparing models to each other. Often leading to *general statements* about model 'good' performance.
- What if the benchmark has historical, representation biases?
- Metrics are important! Aggregated performance metrics (e.g. accuracy) hide subgroup underperformances!

7. Deployment Bias

- Deployment bias occurs when the model is used different from the way it was intended to be used.
- Results from some models need to be confirmed by a human first. However, they users might end-up trusting the results 100%!
- Think of the risk assessment tools again:
 - Intended use: predicting the likelihood of a person committing a crime in future.
 - Wrong use: deciding the length of sentence based on this.

Group Fairness Definitions

1. Demographic Parity

- The probability of a positive outcome should be the same for both protected and unprotected groups.

$$\text{Positive Rate } (A=0) = \text{Positive Rate } (A=1)$$

- Disparate Impact** is similar, but it tries to minimize the difference between positive outcomes of the groups (doesn't have to be zero, usually uses a 80% threshold).

2. Treatment equality

- Ratio of False Negative and False Positives is the same for both protected class subgroups.

3. Equal Odds

- Ensures that protected and unprotected groups have equal True Positive and False Positive rates.

4. Equal Opportunity / True Positive Rate

- Ensures that protected and unprotected groups have equal True Positive rates.

5. False Positive Rate

6. Positive Predictive Value (Predictive Parity)

$$PPV = \frac{TP}{TP + FP}$$

7. Negative Predictive Value

$$NPV = \frac{TN}{TN + FN}$$

Individual Fairness Definitions

- **Fairness Through Unawareness:**

- “An algorithm is fair as long as any protected attributes A are not explicitly used in the decision-making process”

- **Fairness Through Awareness:**

- “An algorithm is fair if it gives similar predictions to similar individuals”
- Similarity w.r.t a similarity metric (distance?)

- **Counterfactual Fairness:**

- A decision is fair towards an individual if it is the same in both the actual world and a counterfactual world where the individual belonged to a different demographic group.

Binary Classifier Evaluation

		Predicted condition			
		Predicted Positive (PP)	Predicted Negative (PN)	Informedness, bookmaker informedness (BM) $= \text{TPR} + \text{TNR} - 1$	Prevalence threshold (PT) $= \frac{\sqrt{\text{TPR} \times \text{FPR}} - \text{FPR}}{\text{TPR} - \text{FPR}}$
Actual condition	Total population $= P + N$				
	Positive (P) ^[a]	True positive (TP), hit ^[b]	False negative (FN), type II error, miss, underestimation ^[c]	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{\text{TP}}{P} = 1 - \text{FNR}$	False negative rate (FNR), miss rate $= \frac{\text{FN}}{P} = 1 - \text{TPR}$
	Negative (N) ^[d]	False positive (FP), type I error, false alarm, overestimation ^[e]	True negative (TN), correct rejection ^[f]	False positive rate (FPR), probability of false alarm, fall-out $= \frac{\text{FP}}{N} = 1 - \text{TNR}$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{\text{TN}}{N} = 1 - \text{FPR}$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{\text{TP}}{\text{PP}} = 1 - \text{FDR}$	False omission rate (FOR) $= \frac{\text{FN}}{\text{PN}} = 1 - \text{NPV}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Negative likelihood ratio (LR-) $= \frac{\text{FNR}}{\text{TNR}}$
	Accuracy (ACC) $= \frac{\text{TP} + \text{TN}}{P + N}$	False discovery rate (FDR) $= \frac{\text{FP}}{\text{PP}} = 1 - \text{PPV}$	Negative predictive value (NPV) $= \frac{\text{TN}}{\text{PN}} = 1 - \text{FOR}$	Markedness (MK), deltaP (Δp) $= \text{PPV} + \text{NPV} - 1$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR-}}$
	Balanced accuracy (BA) $= \frac{\text{TPR} + \text{TNR}}{2}$	F_1 score $= \frac{2 \text{PPV} \times \text{TPR}}{\text{PPV} + \text{TPR}} = \frac{2 \text{TP}}{2 \text{TP} + \text{FP} + \text{FN}}$	Fowlkes–Mallows index (FM) $= \sqrt{\text{PPV} \times \text{TPR}}$	Matthews correlation coefficient (MCC) $= \frac{\sqrt{\text{TPR} \times \text{TNR} \times \text{PPV} \times \text{NPV}}}{\sqrt{\text{FNR} \times \text{FPR} \times \text{FOR} \times \text{FDR}}}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}$

So many more ways to assess fairness!

Different stakeholders want different things from binary classifiers!

The impossibility theorem of fairness

- **Chouldechova, 2017:** if Predictive Parity, False Positive rate, and False Negative rate are equal between 2 groups, then the prevalence is also the same between the groups (contradiction!).
- Proves easily by simple algebra (define metrics in terms of FP, TP, FN, TN).
- You can pick any three binary classifier metrics to achieve this.
- Take-away message: different metrics matter to different stakeholders, and there is no way we can achieve all at the same time!

Bias Mitigation Techniques



Data

- Reweighting
- Optimized Pre-processing
- Disparate Impact Remover

Bias Mitigation Techniques



Data

- Reweighting
- Optimized Pre-processing
- Disparate Impact Remover



Classifier

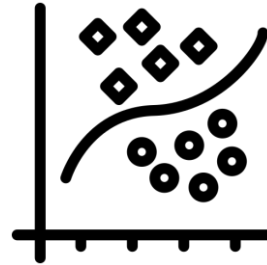
- Prejudice Remover
- Adversarial Debiasing

Bias Mitigation Techniques



Data

- Reweighting
- Optimized Pre-processing
- Disparate Impact Remover



Classifier

- Prejudice Remover
- Adversarial Debiasing



Predictions

- Equalized Odds Post-processing
- Calibrated Odds Post-processing
- Reject Option Classification

Open Challenges

- **Sustainable AI**
 - Carbon footprint, energy consumption, ethical issues of resource allocation.
 - Balance between technological innovation and environmental sustainability.
- **Governance**
 - When and how do we need a certain level of governance? Ranging from research groups, to overarching governmental frameworks.
- **Data Collection**
 - Lack of data from underrepresented groups due to different reasons.
 - Data acquisition processes often don't capture information correctly from diverse individuals.
- **Generalization**
 - One model for all, or one model for every individual?
 - transparent communication of limitations, for which population the model is designed for.
- **Job Security**
 - AI technologies that enhance human capabilities, instead of replacing them.

Questions to ask about AI

- Should we even be doing this?
- What bias is in the data?
- Can the code and data be audited?
- What are error-rates for different subgroups?
- What is the accuracy of a simple rule-based alternative?
- What processes are in place to handle appeals or mistakes? Can we catch them before a disaster?
- How diverse is the team that built it?

Summary

- Responsible AI focuses on developing AI systems that are safe, trustworthy, and ethical, aiming to minimize potential harm to society and the environment.
- RAI is built on principles of Reliability and Safety, Privacy and Security, Transparency, Accountability, and Fairness.
- Algorithmic bias in machine learning can lead to unfair outcomes, and fairness involves ensuring decisions do not discriminate based on sensitive attributes.
- Mitigating bias requires techniques at different stages of the AI lifecycle, including data preprocessing, model training, and post-processing of predictions.

Additional Resources and Tools

- <https://ethics.fast.ai/>
- [21 fairness definitions and their politics](#)
- <https://aif360.res.ibm.com/>
- <https://dssg.github.io/aequitas/>