Algorithmic Fairness in Machine Learning

DS 4400 Guest Lecture

Samantha Dies

3/25/2026

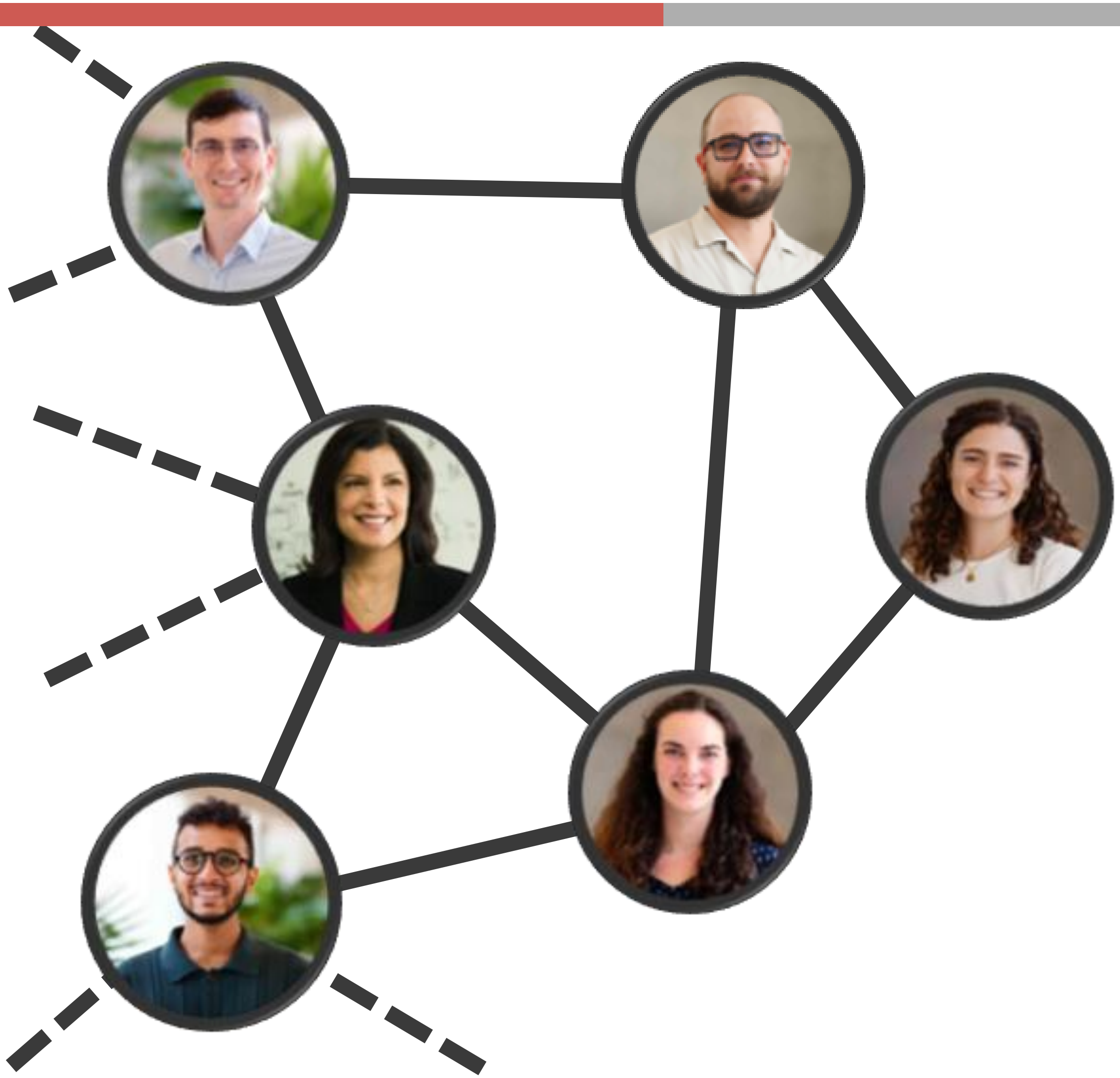
About Me

- 4th year Computer Science PhD candidate at Northeastern University
- Advised by Prof. Tina Eliassi-Rad
- Email: dies.s@northeastern.edu

LVX VERITAS VIRTVS

network science institute

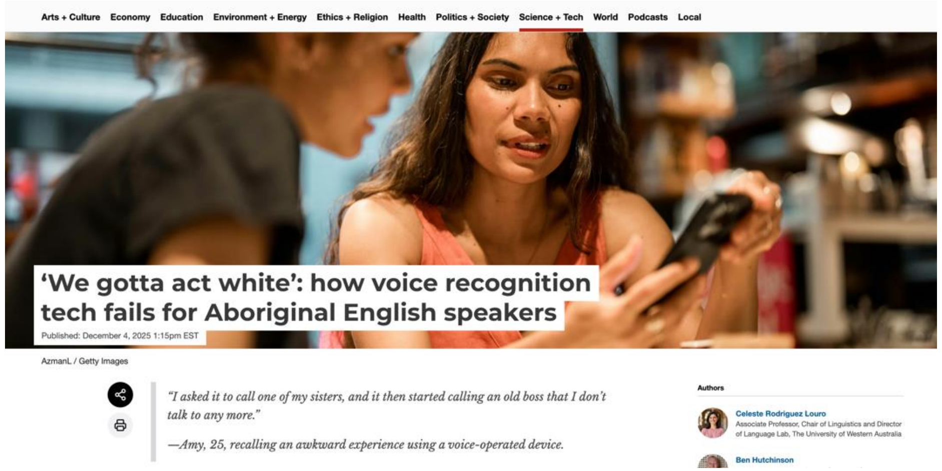
What I do



- I use network science and machine learning to uncover hidden signals in social and AI systems.
- **My goal:** understand how these signals create inequality, and how to measure/mitigate it.

Why do we care?

Why do we care?



‘We gotta act white’: how voice recognition tech fails for Aboriginal English speakers

Published: December 4, 2025 1:15pm EST

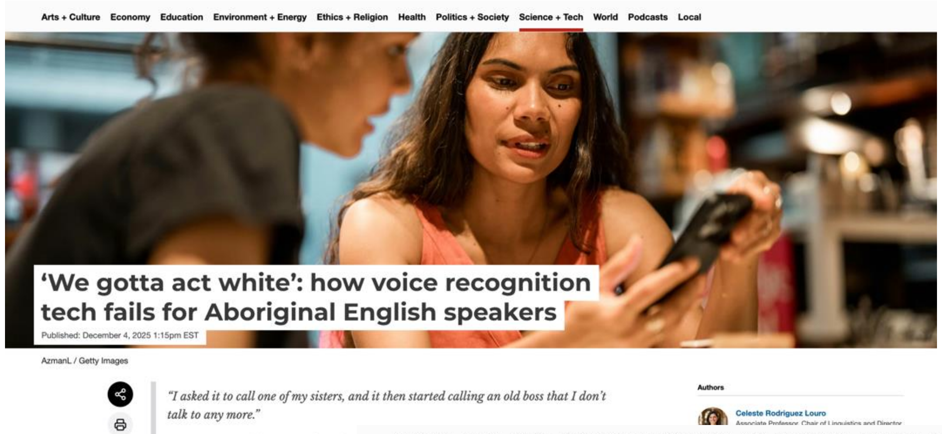
AzmanL / Getty Images

> *“I asked it to call one of my sisters, and it then started calling an old boss that I don’t talk to any more.”*
>
> *—Amy, 25, recalling an awkward experience using a voice-operated device.*

Celeste Rodriguez Louro
Associate Professor, Chair of Linguistics and Director of Language Lab, The University of Western Australia

Ben Hutchinson

Why do we care?



‘We gotta act white’: how voice recognition tech fails for Aboriginal English speakers

Published: December 4, 2025 1:15pm EST

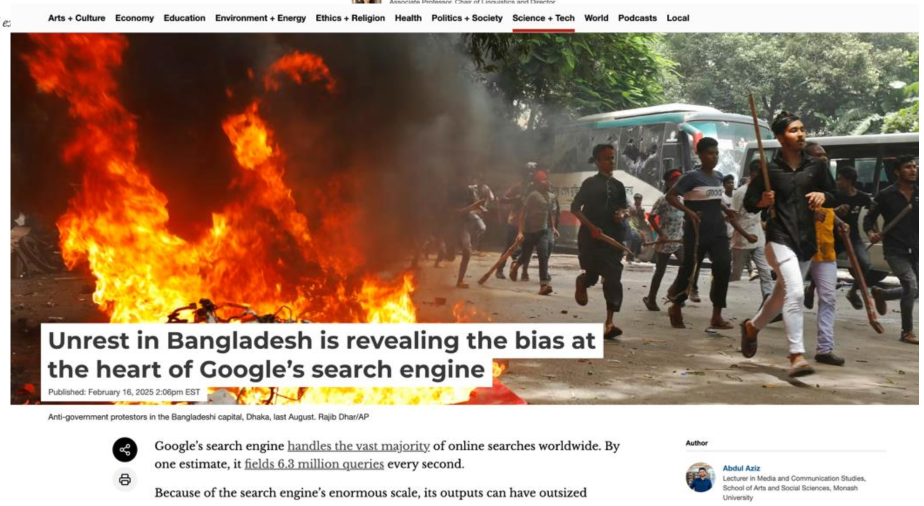
AzmanL / Getty Images

> “I asked it to call one of my sisters, and it then started calling an old boss that I don’t talk to any more.”
>
> —Amy, 25, recalling an awkward e

Authors

Celeste Rodriguez Louro

Associate Professor, Chair of Linguistics and Director



Unrest in Bangladesh is revealing the bias at the heart of Google’s search engine

Published: February 16, 2025 2:06pm EST

Anti-government protestors in the Bangladeshi capital, Dhaka, last August. Rajib Dhar/AP

Google’s search engine handles the vast majority of online searches worldwide. By one estimate, it fields 6.3 million queries every second.

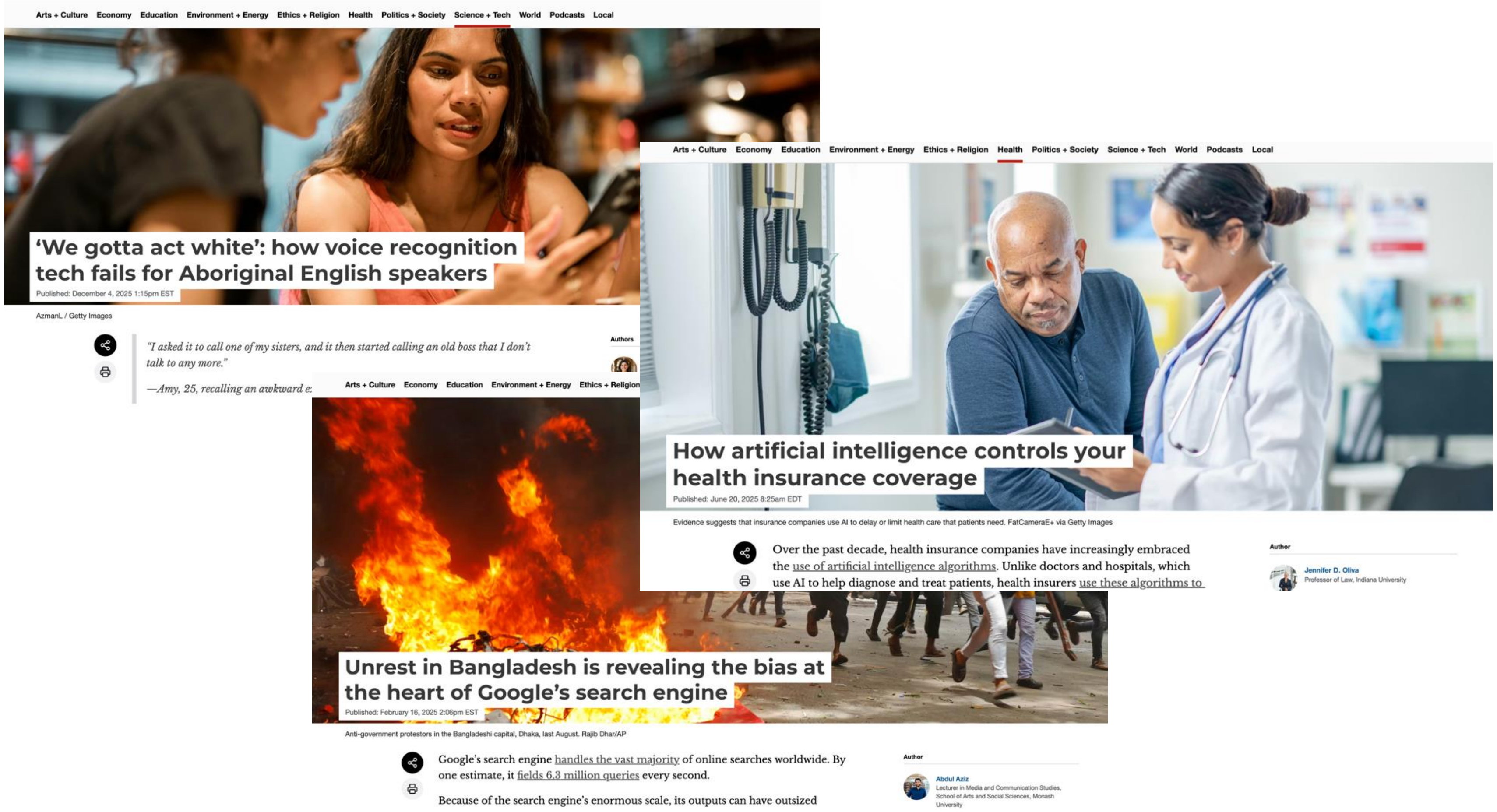
Because of the search engine’s enormous scale, its outputs can have outsized

Author

Abdul Aziz

Lecturer in Media and Communication Studies, School of Arts and Social Sciences, Monash University

Why do we care?


Arts + Culture Economy Education Environment + Energy Ethics + Religion Health Politics + Society Science + Tech World Podcasts Local

‘We gotta act white’: how voice recognition tech fails for Aboriginal English speakers

Published: December 4, 2025 1:15pm EST

AzmanL / Getty Images

> “I asked it to call one of my sisters, and it then started calling an old boss that I don’t talk to any more.”
>
> —Amy, 25, recalling an awkward e

Authors

Arts + Culture Economy Education Environment + Energy Ethics + Religion Health Politics + Society Science + Tech World Podcasts Local

How artificial intelligence controls your health insurance coverage

Published: June 20, 2025 8:25am EDT

Evidence suggests that insurance companies use AI to delay or limit health care that patients need. FatCameraE+ via Getty Images

Over the past decade, health insurance companies have increasingly embraced the use of artificial intelligence algorithms. Unlike doctors and hospitals, which use AI to help diagnose and treat patients, health insurers use these algorithms to

Author

Jennifer D. Oliva
Professor of Law, Indiana University

Arts + Culture Economy Education Environment + Energy Ethics + Religion

Unrest in Bangladesh is revealing the bias at the heart of Google’s search engine

Published: February 16, 2025 2:06pm EST

Anti-government protestors in the Bangladeshi capital, Dhaka, last August. Rajib Dhar/AP

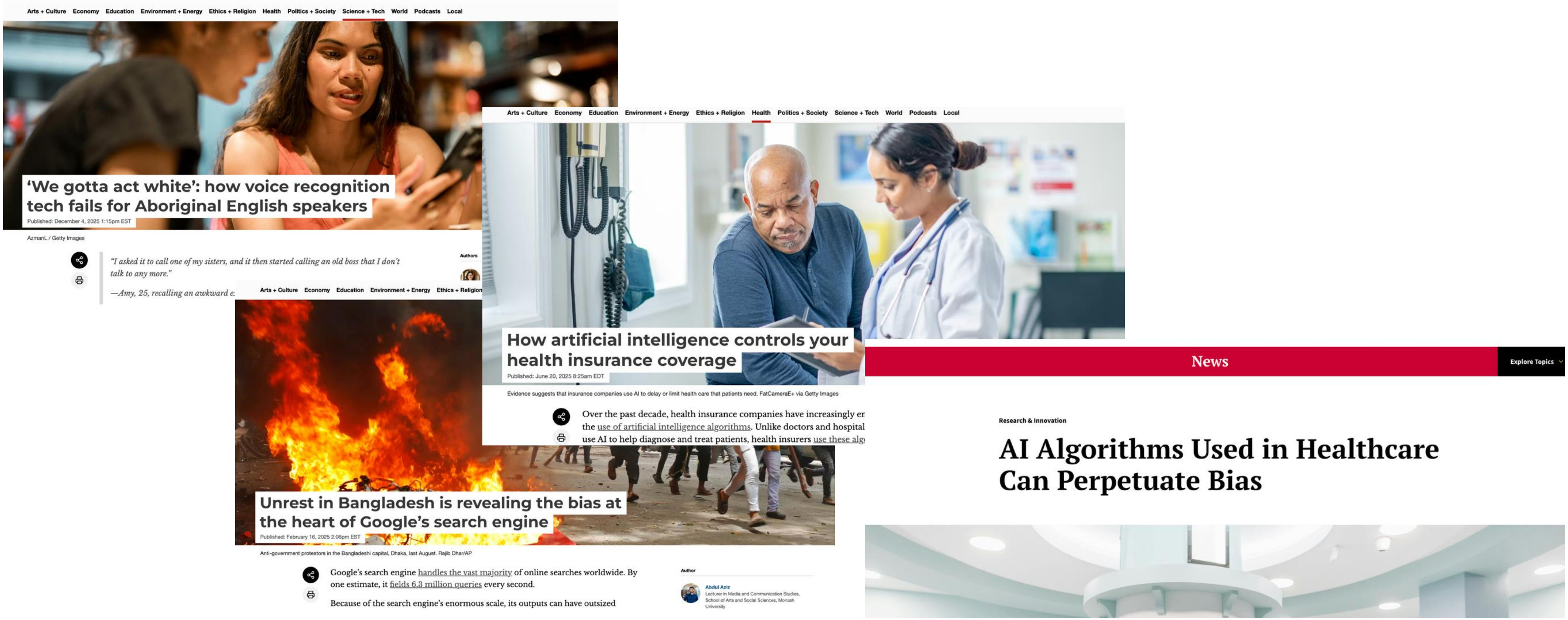
Google’s search engine handles the vast majority of online searches worldwide. By one estimate, it fields 6.3 million queries every second.

Because of the search engine’s enormous scale, its outputs can have outsized

Author

Abdul Aziz
Lecturer in Media and Communication Studies, School of Arts and Social Sciences, Monash University

Why do we care?



'We gotta act white': how voice recognition tech fails for Aboriginal English speakers

Published: December 4, 2025 1:15pm EST

AzmanL / Getty Images

"I asked it to call one of my sisters, and it then started calling an old boss that I don't talk to any more."

—Amy, 25, recalling an awkward e

Authors

Unrest in Bangladesh is revealing the bias at the heart of Google's search engine

Published: February 16, 2025 2:06pm EST

Anti-government protestors in the Bangladeshi capital, Dhaka, last August. Rajib Dhar/AP

Google's search engine handles the vast majority of online searches worldwide. By one estimate, it fields 6.3 million queries every second.

Because of the search engine's enormous scale, its outputs can have outsized

Author

Abdul Aziz
Lecturer in Media and Communication Studies, School of Arts and Social Sciences, Monash University

How artificial intelligence controls your health insurance coverage

Published: June 20, 2025 8:25am EDT

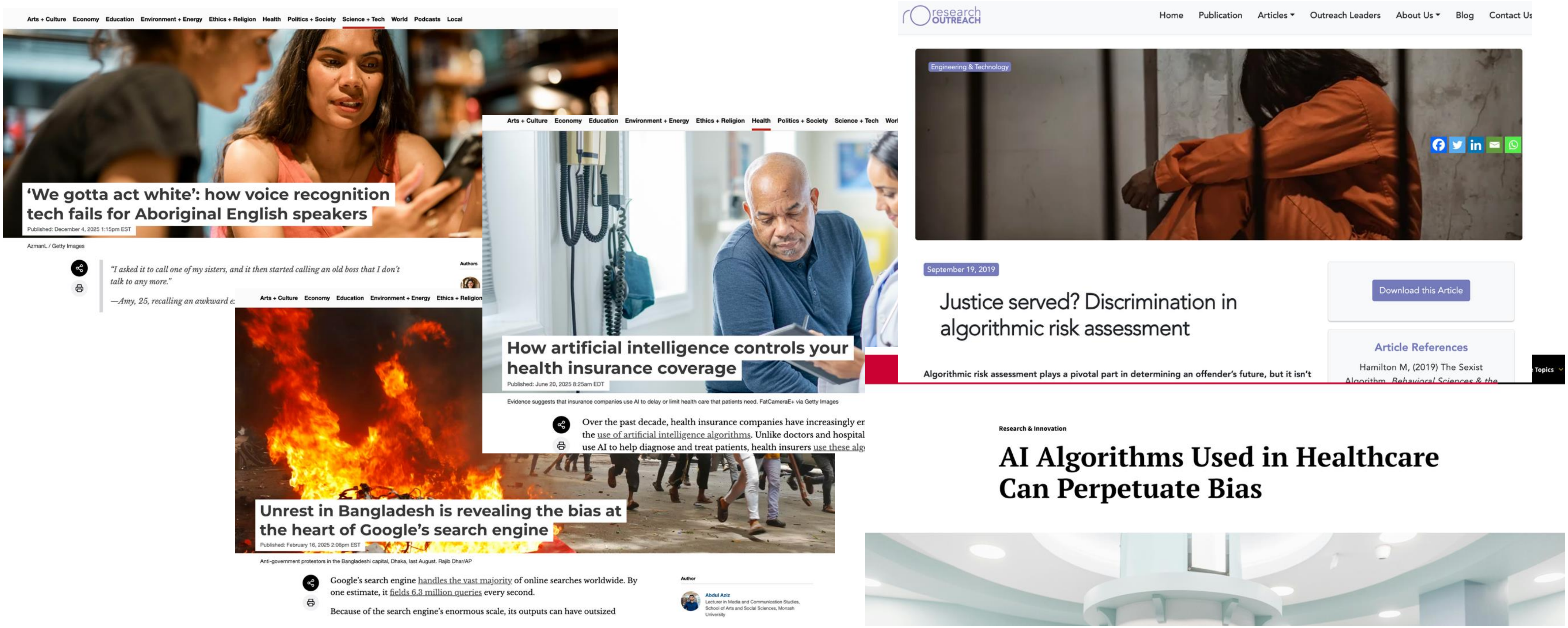
Evidence suggests that insurance companies use AI to delay or limit health care that patients need. FatCameraE+ via Getty Images

Over the past decade, health insurance companies have increasingly er
the use of artificial intelligence algorithms. Unlike doctors and hospital
use AI to help diagnose and treat patients, health insurers use these alg

Research & Innovation

AI Algorithms Used in Healthcare Can Perpetuate Bias

Why do we care?


Arts + Culture Economy Education Environment + Energy Ethics + Religion Health Politics + Society Science + Tech World Podcasts Local
research OUTREACH
Home Publication Articles Outreach Leaders About Us Blog Contact Us
Engineering & Technology
Download this Article
Topics
Authors
Author

‘We gotta act white’: how voice recognition tech fails for Aboriginal English speakers

Published: December 4, 2025 1:15pm EST

AzmanL / Getty Images

> “I asked it to call one of my sisters, and it then started calling an old boss that I don’t talk to any more.”
>
> —Amy, 25, recalling an awkward e

How artificial intelligence controls your health insurance coverage

Published: June 20, 2025 8:25am EDT

Evidence suggests that insurance companies use AI to delay or limit health care that patients need. FatCameraE+ via Getty Images

Over the past decade, health insurance companies have increasingly er the use of artificial intelligence algorithms. Unlike doctors and hospital use AI to help diagnose and treat patients, health insurers use these alg

Unrest in Bangladesh is revealing the bias at the heart of Google’s search engine

Published: February 16, 2025 2:06pm EST

Anti-government protestors in the Bangladeshi capital, Dhaka, last August. Rajib Dhar/AP

Google’s search engine handles the vast majority of online searches worldwide. By one estimate, it fields 6.3 million queries every second.

Because of the search engine’s enormous scale, its outputs can have outsized

Abdul Aziz
Lecturer in Media and Communication Studies, School of Arts and Social Sciences, Monash University

September 19, 2019

Justice served? Discrimination in algorithmic risk assessment

Algorithmic risk assessment plays a pivotal part in determining an offender’s future, but it isn’t

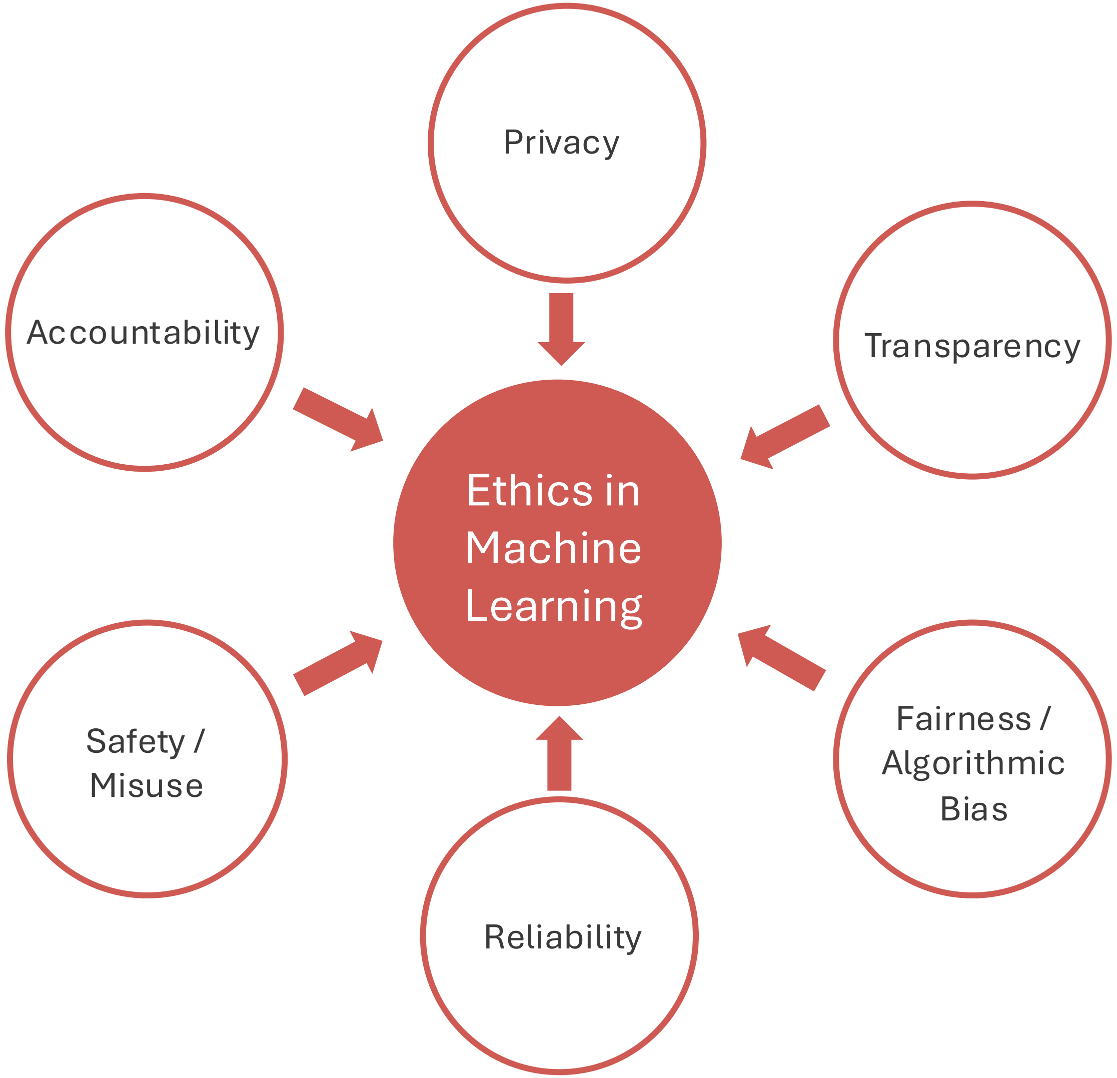
Article References

Hamilton M, (2019) The Sexist Algorithm. *Behavioral Sciences & the*

Research & Innovation

AI Algorithms Used in Healthcare Can Perpetuate Bias

Ethics in Computer and Data Science



Privacy

Accountability

Transparency

Ethics in Machine Learning

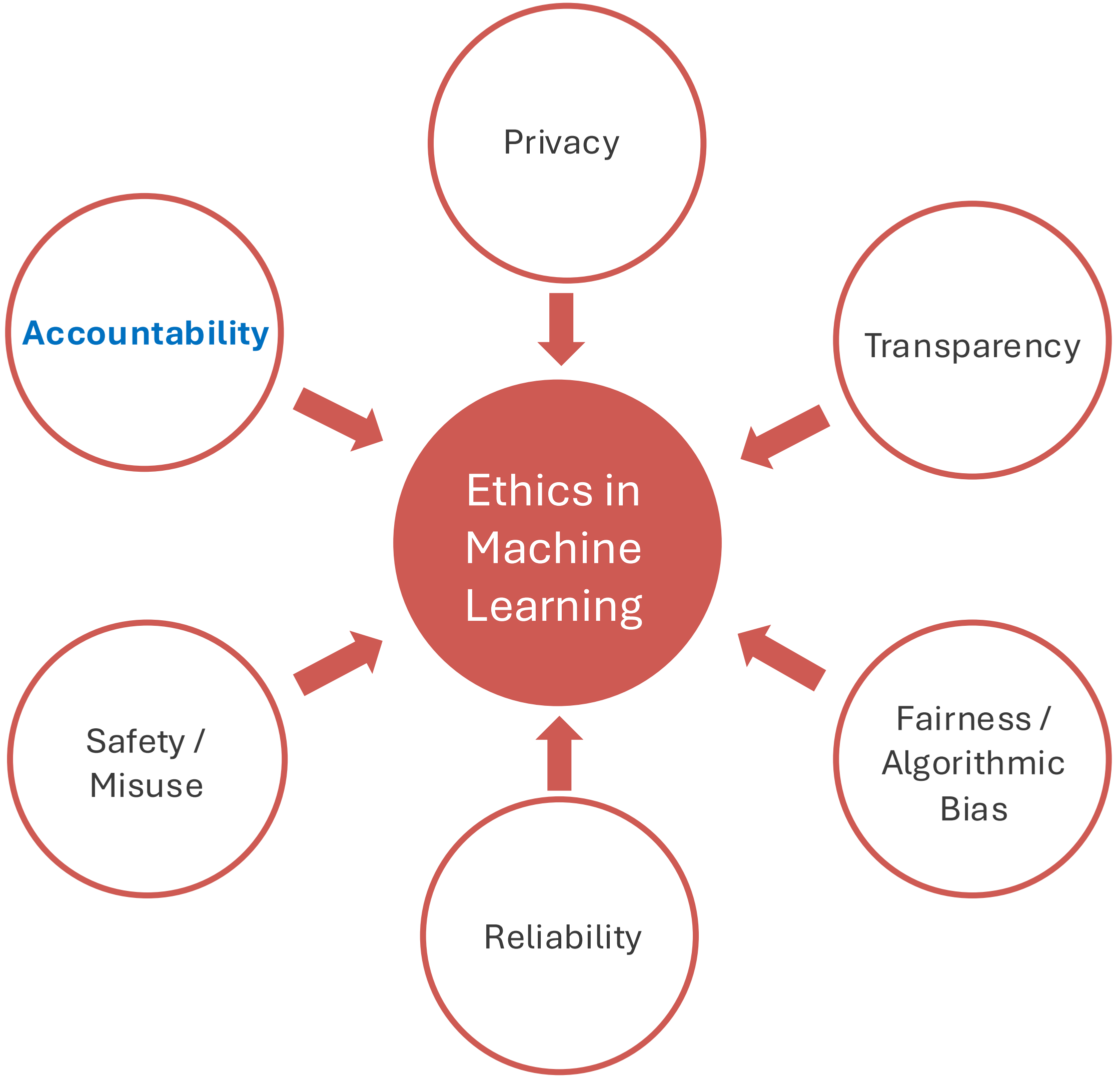
Safety / Misuse

Fairness / Algorithmic Bias

Reliability

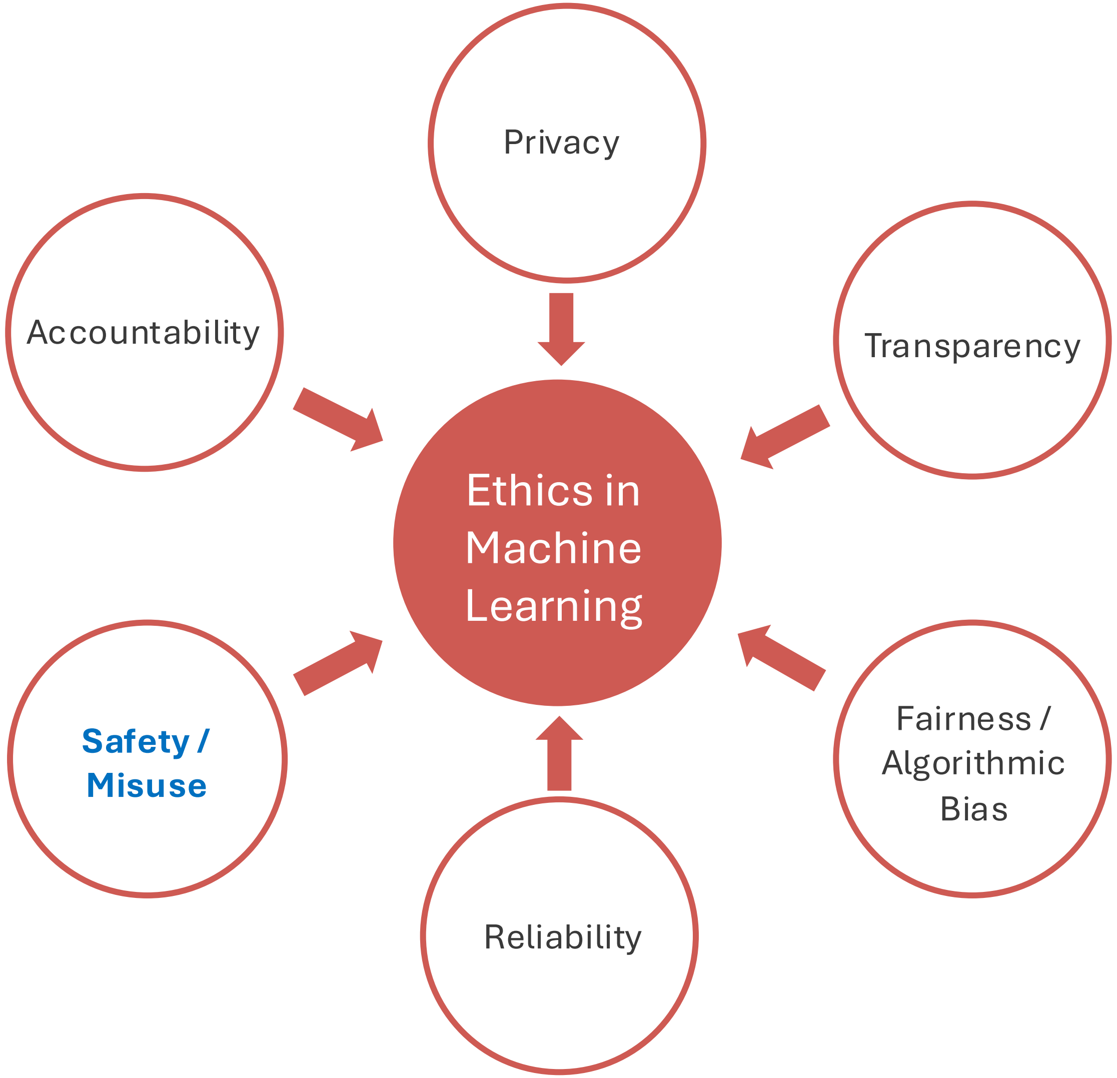
Ethics in Computer and Data Science

Who's responsible for model behavior?



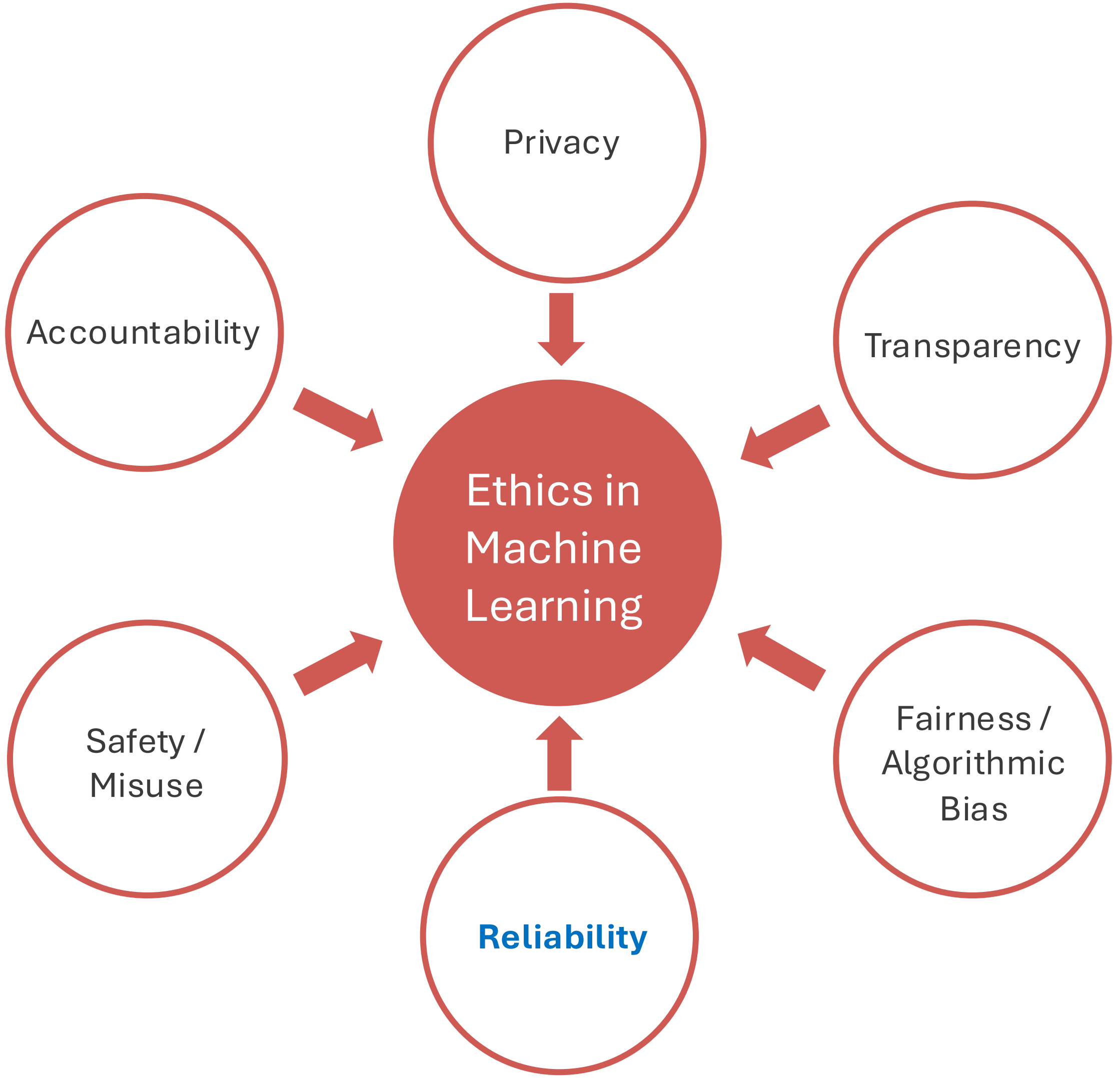
Ethics in Computer and Data Science

Could the system cause harm?

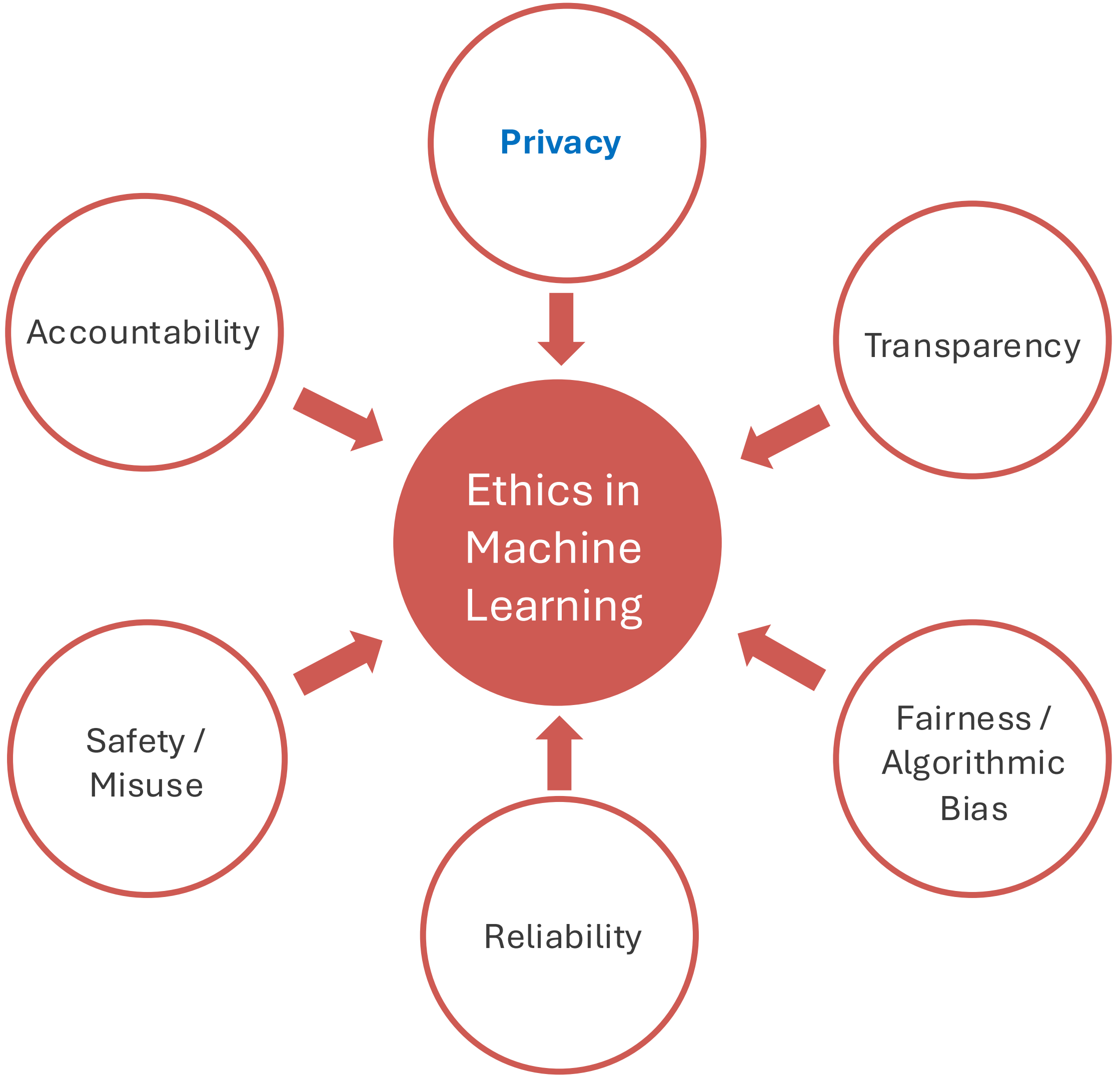


Ethics in Computer and Data Science

Will the system work the way we think during deployment?

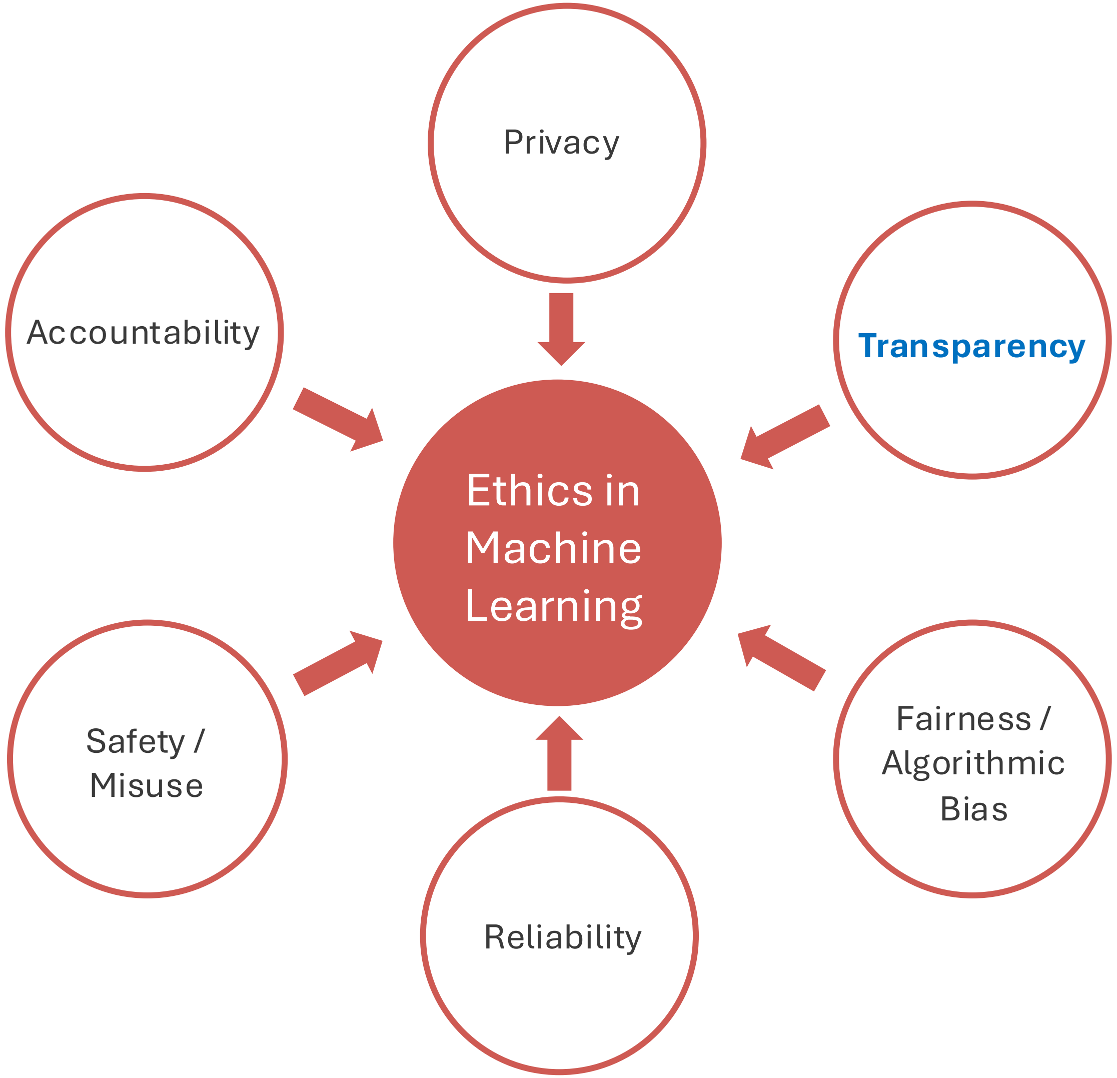


Ethics in Computer and Data Science



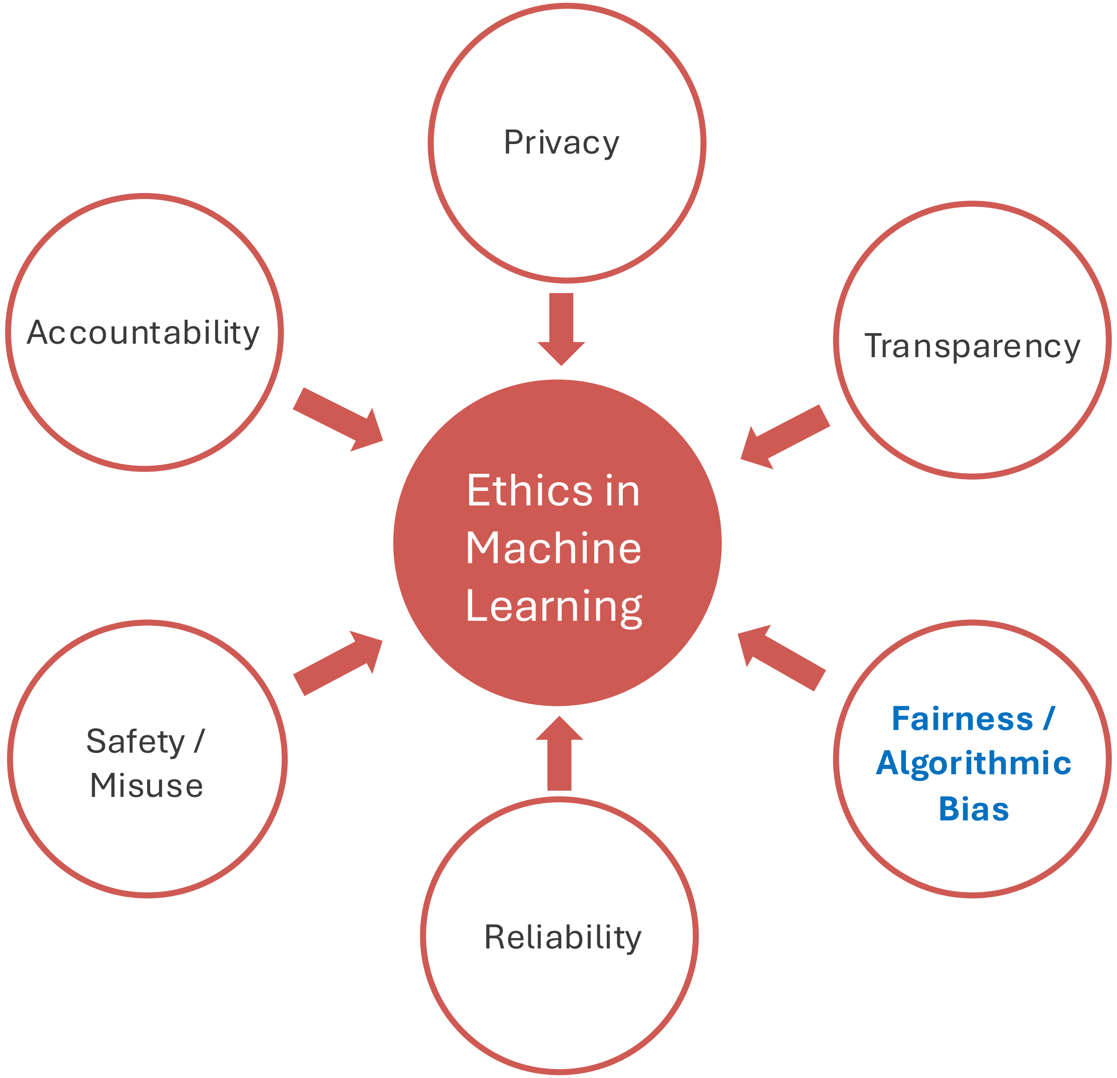
How do we protect personal data?

Ethics in Computer and Data Science



Do we understand how the model makes decisions?

Ethics in Computer and Data Science



Does the model treat different groups differently?

What is Algorithmic Bias?

Systematic differences in model behavior across protected groups

What is Algorithmic Bias?

Different error rates, accuracy, etc.

Systematic differences in model behavior across protected groups

What is Algorithmic Bias?

Systematic differences in model behavior across protected groups

Gender, race, age, etc.

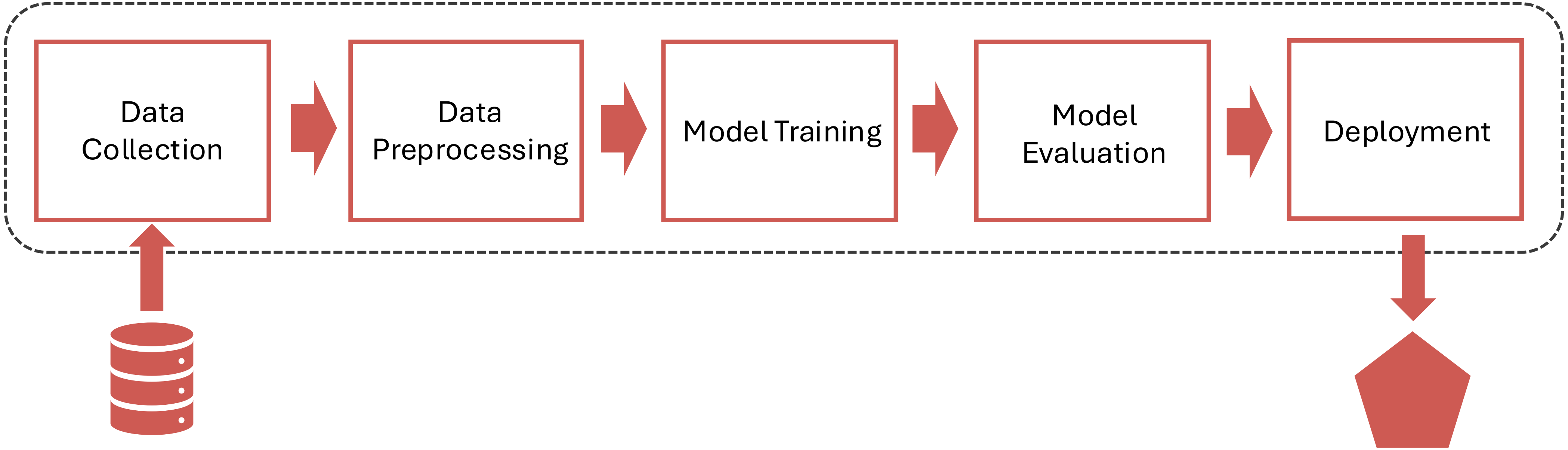
What is Algorithmic Bias?

Causal? Correlational? What about mediating variables?

Systematic differences in model behavior across protected groups

Sources of Bias

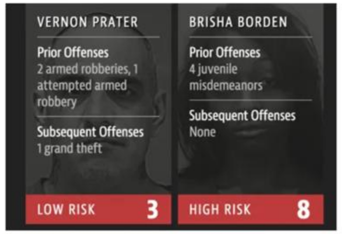
Machine Learning Pipeline



Bias can be introduced at *every stage*, and obscured by evaluation metrics

Example - Recidivism

“The tendency of a convicted criminal to reoffend”



VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

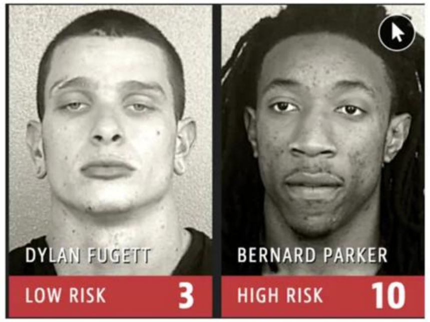
LOW RISK 3

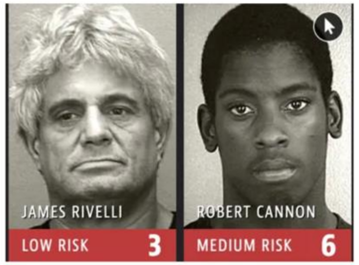
BRISHA BORDEN

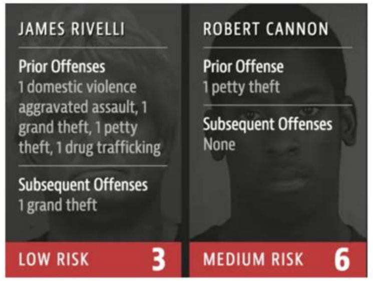
Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8







JAMES RIVELLI

Prior Offenses
1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking

Subsequent Offenses
1 grand theft

LOW RISK 3

ROBERT CANNON

Prior Offense
1 petty theft

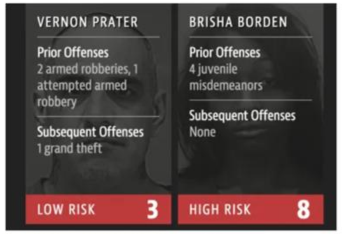
Subsequent Offenses
None

MEDIUM RISK 6

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
https://s3.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf

Example - Recidivism

"The tendency of a convicted criminal to reoffend"



VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

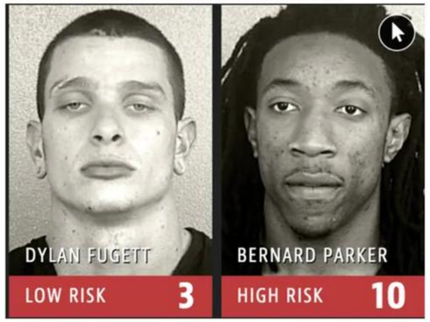
LOW RISK 3

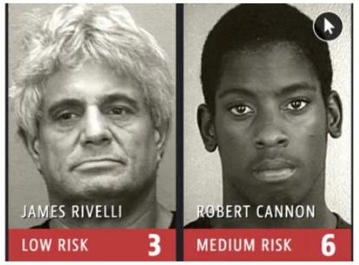
BRISHA BORDEN

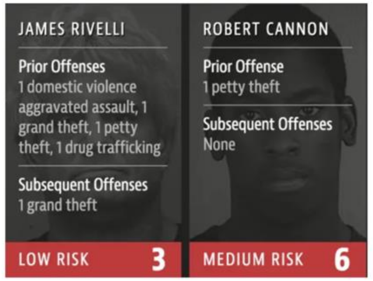
Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8







JAMES RIVELLI

Prior Offenses
1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking

Subsequent Offenses
1 grand theft

LOW RISK 3

ROBERT CANNON

Prior Offense
1 petty theft

Subsequent Offenses
None

MEDIUM RISK 6

COMPAS Software (2016):

- Used in courts to predict recidivism risk and inform parole decisions
- Racially biased according to some metrics, but fair according to others
- Currently in use in states including New York, California, and Florida

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
https://s3.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf

Today's Plan

1. Discuss fairness
2. Define fairness metrics mathematically
3. Fairness code example
4. Practice choosing fairness metrics

Activity #1: Think/Pair/Share (~10 min)

Instructions:

1. Think individually
2. Discuss in groups of 3-4
3. Each group will share one idea/question

Prompts:

- Where have you seen algorithmic bias discussed in the past, if at all?
- Have you noticed any bias or unfairness in the technologies you use (e.g., Instagram, tiktok, ChatGPT, etc.)?
- What key words, assumptions, and questions come to mind when you think about algorithmic bias and fairness?

Algorithmic Bias is Complicated

- People are affected differently by different types of bias
- Different stakeholders care about different aspects of bias/fairness
- Measuring (and mitigating!) fairness is nontrivial

Algorithmic Bias is Complicated

- People are affected differently by different types of bias
- Different stakeholders care about different aspects of bias/fairness
- Measuring (and mitigating!) fairness is nontrivial

But… it's *extremely* important

Algorithmic Bias is Complicated

- People are affected differently by different types of bias
- Different stakeholders care about different aspects of bias/fairness
- Measuring (and mitigating!) fairness is nontrivial

But… it’s *extremely* important

- Ethically: we don’t want our models to inadvertently cause harm
- Legally: we could get in a lot of trouble if we do

How do we measure fairness?

- What are some of the different metrics?
- How do we calculate them?
- What do they measure?
- What are the differences between them?

Confusion Matrix Reminder

		Predicted Class	
		+	–
Actual Class	+	True Positive (TP)	False Negative (FN)
	–	False Positive (FP)	True Negative (TN)

Confusion Matrix Reminder

		Predicted Class	
		+	–
Actual Class	+	True Positive (TP)	False Negative (FN)
	–	False Positive (FP)	True Negative (TN)

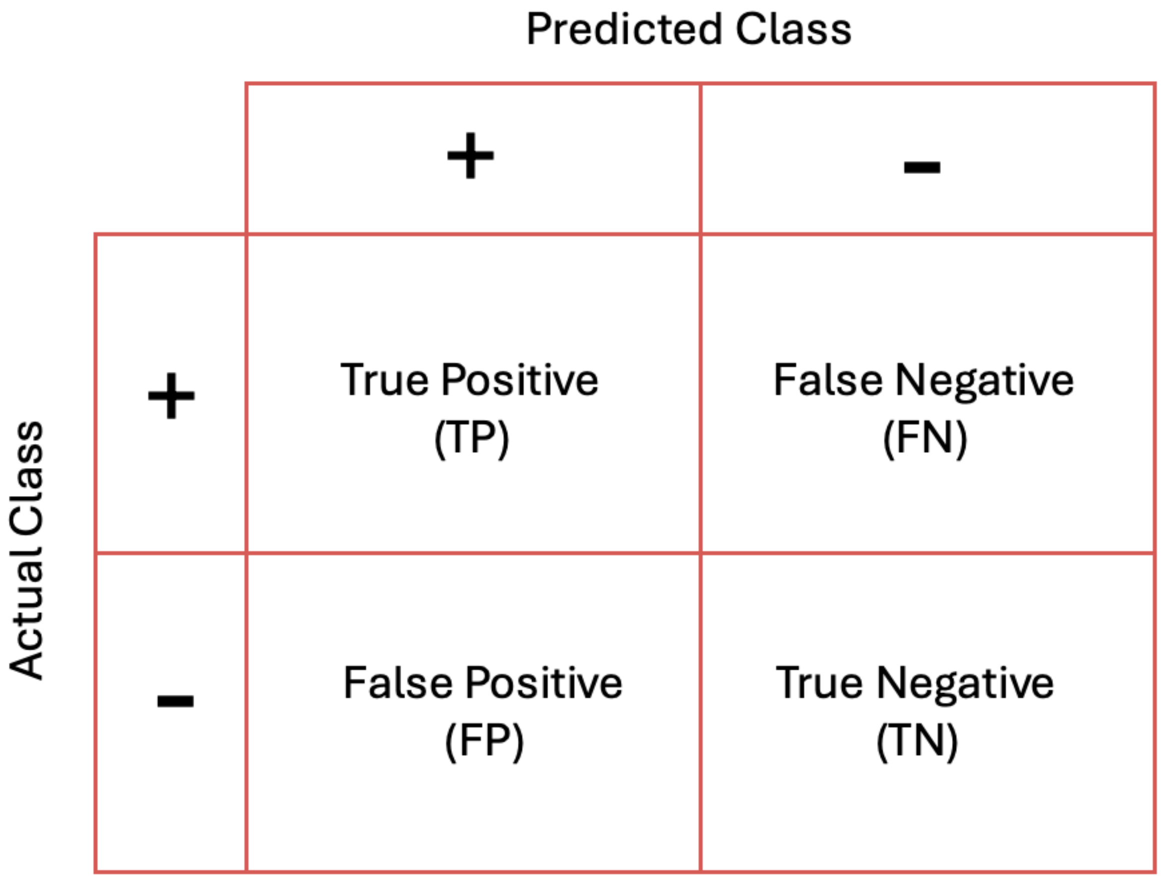
Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$

Precision: $\frac{TP}{TP+FP}$

Recall: $\frac{TP}{TP+FN}$

Traditional Evaluation

Calculate chosen metric using the entire test set



		Predicted Class	
		+	–
Actual Class	+	True Positive (TP)	False Negative (FN)
	–	False Positive (FP)	True Negative (TN)

Parity-based Fairness

Compare the chosen metric across different subgroups

		Predicted Class	
		+	-
Actual Class	+	True Positive (TP)	False Negative (FN)
	-	False Positive (FP)	True Negative (TN)

vs.

		Predicted Class	
		+	-
Actual Class	+	True Positive (TP)	False Negative (FN)
	-	False Positive (FP)	True Negative (TN)

e.g., is accuracy the same for men and women in the test set?

Parity-based Fairness

		Predicted Class	
		+	–
Actual Class	+	True Positive (TP)	False Negative (FN)
	–	False Positive (FP)	True Negative (TN)

For two subgroups A and B of the test set, a model is said to be fair according to a given metric if

$$Metric(A) \approx Metric(B)$$

Parity-based Fairness

		Predicted Class	
		+	–
Actual Class	+	True Positive (TP)	False Negative (FN)
	–	False Positive (FP)	True Negative (TN)

For two subgroups A and B of the test set, a model is said to be fair according to a given metric if

$$Metric(A) \approx Metric(B)$$

Overall Accuracy Parity:

$$\left| \frac{TP_A + TN_A}{TP_A + TN_A + FP_A + FN_A} - \frac{TP_B + TN_B}{TP_B + TN_B + FP_B + FN_B} \right|$$

Recall/TPR Parity:

$$\left| \frac{TP_A}{TP_A + FN_A} - \frac{TP_B}{TP_B + FN_B} \right|$$

Precision/Predictive Parity:

$$\left| \frac{TP_A}{TP_A + FP_A} - \frac{TP_B}{TP_B + FP_B} \right|$$

Parity-based Fairness

		Predicted Class	
		+	–
Actual Class	+	True Positive (TP)	False Negative (FN)
	–	False Positive (FP)	True Negative (TN)

Other parity-based measures:

- False Negative Rate Parity
- False Discovery Rate Parity
- True Negative Rate (Specificity) Parity
- False Positive Rate Parity
- Negative Predictive Value Parity
- False Omission Rate Parity

How do we Choose a Fairness Metric?

		Predicted Class	
		+	–
Actual Class	+	True Positive (TP)	False Negative (FN)
	–	False Positive (FP)	True Negative (TN)

How do we Choose a Fairness Metric?

		Predicted Class	
		+	–
Actual Class	+	True Positive (TP)	False Negative (FN)
	–	False Positive (FP)	True Negative (TN)

Recall/TPR Parity:

$$\left| \frac{TP_A}{TP_A + FN_A} - \frac{TP_B}{TP_B + FN_B} \right|$$

Precision/Predictive Parity:

$$\left| \frac{TP_A}{TP_A + FP_A} - \frac{TP_B}{TP_B + FP_B} \right|$$

How do we Choose a Fairness Metric?

		Predicted Class	
		+	–
Actual Class	+	True Positive (TP)	False Negative (FN)
	–	False Positive (FP)	True Negative (TN)

Recall/TPR Parity:

$$\left| \frac{TP_A}{TP_A + FN_A} - \frac{TP_B}{TP_B + FN_B} \right|$$

Precision/Predictive Parity:

$$\left| \frac{TP_A}{TP_A + FP_A} - \frac{TP_B}{TP_B + FP_B} \right|$$

How do we Choose a Fairness Metric?

		Predicted Class	
		+	–
Actual Class	+	True Positive (TP)	False Negative (FN)
	–	False Positive (FP)	True Negative (TN)

Recall/TPR Parity:

$$\left|\frac{TP_A}{TP_A+FN_A}-\frac{TP_B}{TP_B+FN_B}\right|$$

Precision/Predictive Parity:

$$\left|\frac{TP_A}{TP_A+FP_A}-\frac{TP_B}{TP_B+FP_B}\right|$$

How do we Choose a Fairness Metric?

		Predicted Class	
		+	–
Actual Class	+	True Positive (TP)	False Negative (FN)
	–	False Positive (FP)	True Negative (TN)

Conditions on a positive ground-truth label

Recall/TPR Parity:

$$\left| \frac{TP_A}{TP_A + FN_A} - \frac{TP_B}{TP_B + FN_B} \right|$$

Precision/Predictive Parity:

$$\left| \frac{TP_A}{TP_A + FP_A} - \frac{TP_B}{TP_B + FP_B} \right|$$

Conditions on a positive model prediction

How do we Choose a Fairness Metric?

1. Decide on the most critical classification outcome (numerator)
    - True positive
    - True negative
    - False positive
    - False negative

2. Decide on the conditioning factor (denominator)
    - What actually happened
      +: TP + FN
      - : TN + FP
    - What the model predicted would happen
      +: TP + FP
      - : TN + FN

How do we Choose a Fairness Metric?

1. Decide on the most critical classification outcome (numerator)

   - True positive
   - True negative
   - False positive
   - False negative

Medicine: FN means we failed to detect a disease
Recidivism: FP means someone sits in jail longer unnecessarily

2. Decide on the conditioning factor (denominator)

   - What actually happened
     +: TP + FN
     - : TN + FP
   - What the model predicted would happen
     +: TP + FP
     - : TN + FN

How do we Choose a Fairness Metric?

1. Decide on the most critical classification outcome (numerator)
   - True positive
   - True negative
   - False positive
   - False negative

Medicine: FN means we failed to detect a disease
Recidivism: FP means someone sits in jail longer unnecessarily

2. Decide on the conditioning factor (denominator)
   - What actually happened
     +: TP + FN
     - : TN + FP
   - What the model predicted would happen
     +: TP + FP
     - : TN + FN

Actual ≈ Moral
Model ≈ Legal

Demo

https://colab.research.google.com/drive/1c8UqKrCCnpUXgMHecmBRf-wkdN1i5ntF?usp=sharing

(also on the course website)

Activity #2: Peer Explanation (~5 min)

Instructions:

1. Answer a poll individually
2. Discuss response with neighbors
3. Vote again

Scenario #1: Cancer Detection

The CDC has an ML model which attempts to predict whether a patient has lung cancer.

The worst-case scenario is that women are more likely than men to be predicted to be cancer free when they actually have lung cancer.

Which fairness metric is most appropriate?

a. False Omission Rate Parity

$$\frac{FN_A}{TN_A - FN_A} - \frac{FN_B}{TN_B - FN_B}$$

b. Recall Parity

$$\frac{TP_A}{TP_A - FN_A} - \frac{TP_B}{TP_B - FN_B}$$

c. False Negative Rate Parity

$$\frac{FN_A}{TP_A - FN_A} - \frac{FN_B}{TP_B - FN_B}$$

		Predicted Class	
		+	-
Actual Class	+	True Positive (TP)	False Negative (FN)
	-	False Positive (FP)	True Negative (TN)

PollEv.com/samanthadies671

Scenario #1: Cancer Detection

The CDC has an ML model which attempts to predict whether a patient has lung cancer.

The worst-case scenario is that women are more likely than men to be predicted to be cancer free when they actually have lung cancer.

Which fairness metric is most appropriate?

a. False Omission Rate Parity

$$\frac{FN_A}{TN_A - FN_A} - \frac{FN_B}{TN_B - FN_B}$$

b. Recall Parity

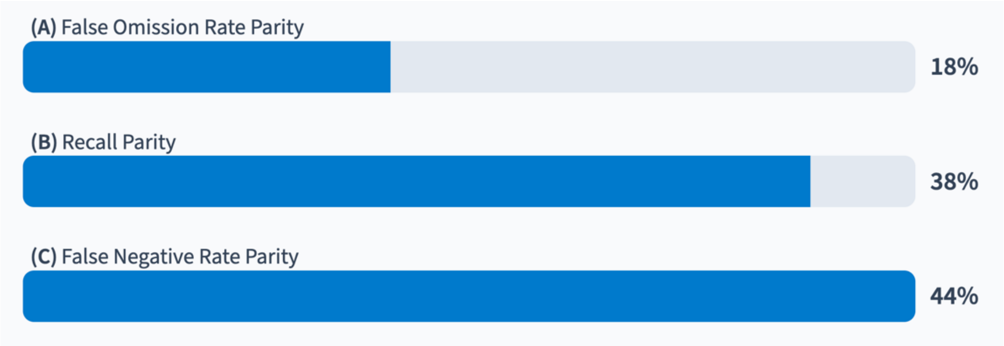
$$\frac{TP_A}{TP_A - FN_A} - \frac{TP_B}{TP_B - FN_B}$$

c. False Negative Rate Parity

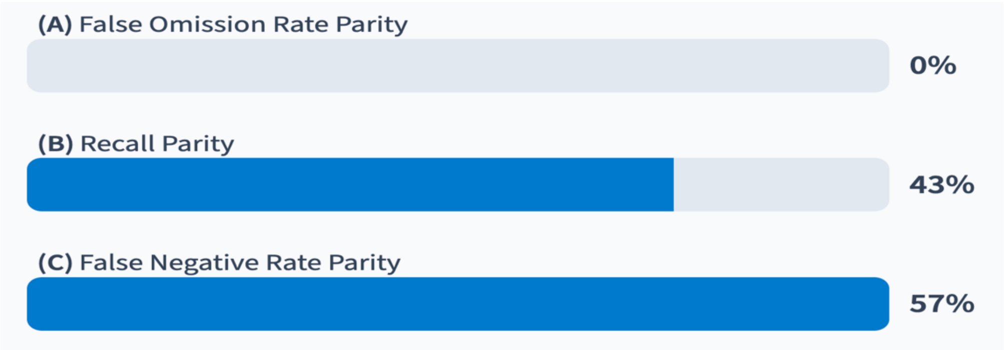
$$\frac{FN_A}{TP_A - FN_A} - \frac{FN_B}{TP_B - FN_B}$$

Sam's pick
(but it's arguable)

Before:

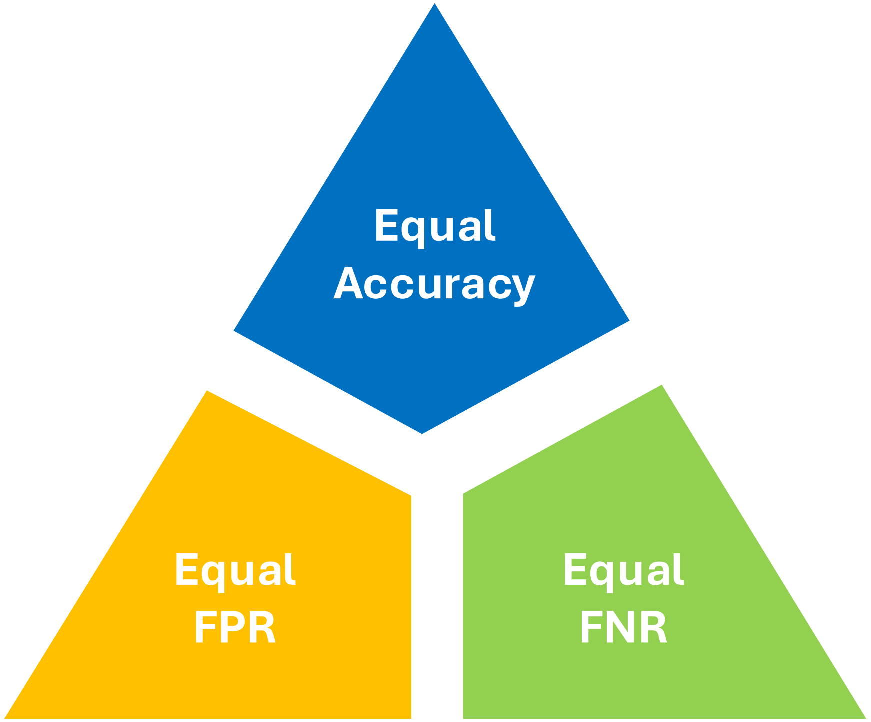


After:



Fairness is Hard!

- Many fairness metrics exist
- The different metrics capture different types of biases
- The metrics don’t always agree
- Sometimes you can’t satisfy them all



Can’t have all three
(unless you have a perfect classifier!)

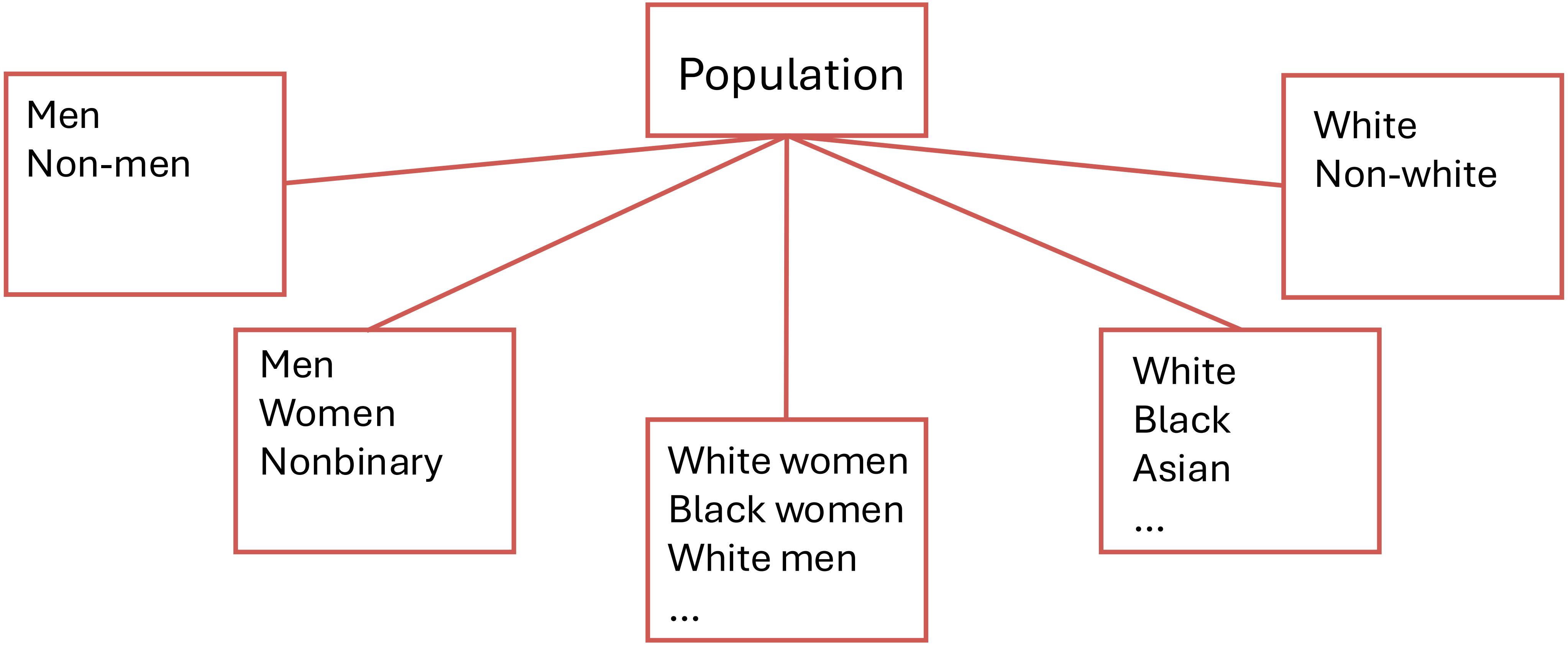
Context Matters

The right fairness metric depends on context and the ML task

Domain	Most Harmful Error
Recidivism	False high-risk (FP)
Medicine	Overlooked disease (FN)
Job Hiring	Depends on goals

Subgroup Selection Matters

Different population partitions can lead to different fairness outcomes



Population

Men
Non-men

Men
Women
Nonbinary

White women
Black women
White men
…

White
Black
Asian
…

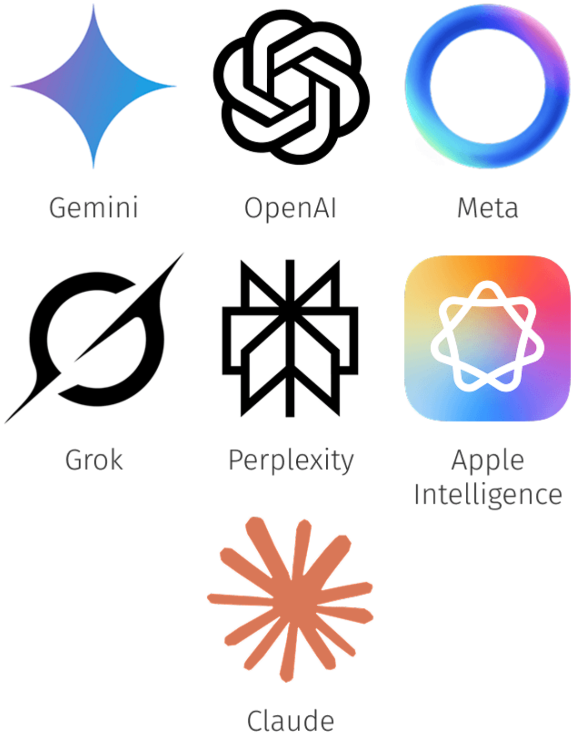
White
Non-white

Real-world Systems are Complicated

Today’s Lecture	Real Systems
Simple, white-box models	Black-box models
Binary labels	Imperfect labels
One protected group	Multiple protected attributes
Parity-based fairness metrics	Even more fairness metric options

Modern AI / LLMs are Even Trickier

- Answers depend on context
- Models can generate new, unpredictable content
- Bias may appear in subtle ways
- Hard to define what “fair” means



Final Thoughts

Fairness is not automatic

We must choose:

- What errors matter
- Which groups to compare
- Which metrics to use

These choices reflect goals and values

Final Thoughts

Fairness is not automatic

We must choose:

- What errors matter
- Which groups to compare
- Which metrics to use

Fairness is a design decision, not just a calculation

These choices reflect goals and values

Additional Resources

Python packages for measuring and addressing bias in ML models

- AIF360: https://aif360.readthedocs.io/en/stable/
- Dalex: https://dalex.drwhy.ai/python-dalex-fairness.html
- Fairlearn: https://fairlearn.org/v0.8/auto_examples/index.html
- Responsibly: https://docs.responsibly.ai/_modules/responsibly/fairness/metrics/visualization.html
- Google What-If Tool: https://pair-code.github.io/what-if-tool/
  - Jupyter extension for analyzing models

Useful links

- Wikipedia - Fairness: https://en.wikipedia.org/wiki/Fairness_(machine_learning)
- Fairness and Machine Learning (fairmlbook): https://fairmlbook.org/index.html