Data, Decisions, and Dilemmas

Evaluating Ethics and Bias in Healthcare AI

Kaitlyn Wade, PhD Candidate
Department of Computer Science
Western University

Agenda

What Does it Mean for an AI to be Fair?

Metrics and Their Trade-offs

Sources of Bias

Philosophy & Aligning AI with Human Values

Case Study

Tips for Designing Fair AI Models

Defining Fairness

Terminology

Bias: Systematic error that disadvantages certain groups

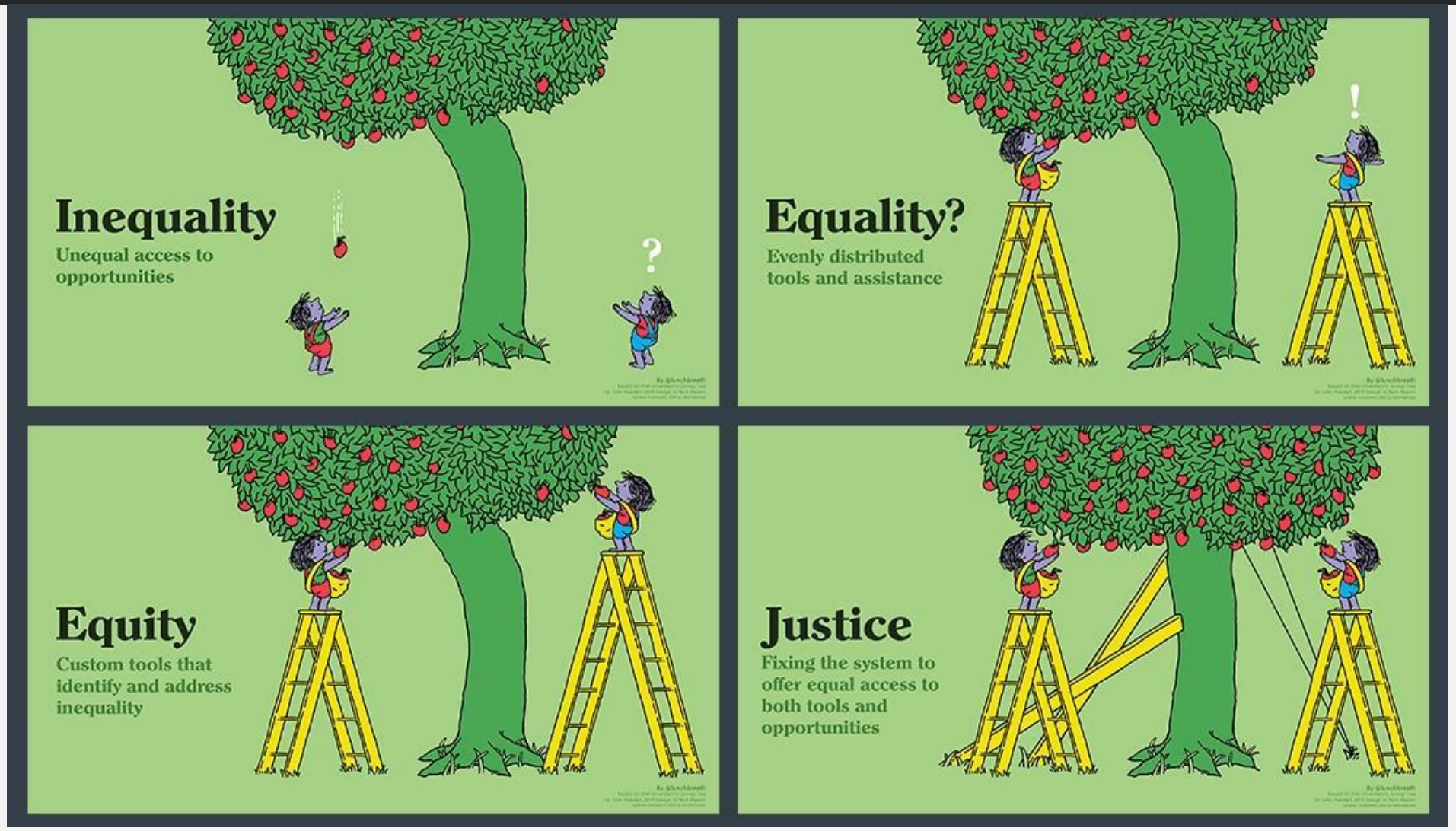
Fairness: The absence of unjust or unequal treatment

Equity: Adjusting for differences to achieve equal outcomes

Equality: Treating everyone the same regardless of circumstance

Ethics: The study of what is right, good, or just

Terminology



A Real-World Scenario

You're designing an AI tool to help doctors make referrals for diagnostic testing.

What would a "fair" AI system look like?

- Follow the same process for everyone, regardless of how sick they are?
- Same health outcomes for everyone?
- Access be prioritized for those who need it most?
- Who should decide what "fair" means?

What does it mean for an AI model to be “fair”?

Fairness Is Complicated

- There are many valid definitions of “fairness” and they can conflict
- Design choices determine which definitions of fairness are built into our AI models
- Fairness is not only a technical property, but it’s also an ethical one
- Different stakeholders prioritize different definitions

The COMPAS Case Study

- **C**orrectional **O**ffender **M**anagement **P**rofiling for **A**lternative **S**anctions

The COMPAS Case Study

The Good

- Data-driven decision support → efficiency
- “Impartial” and “unbiased”
- Calibrated with respect to race

The Bad

- Black defendants more likely to be falsely flagged high-risk
- White defendants more likely to be incorrectly flagged low-risk

The Ugly

- Calibration doesn’t guarantee fairness
- Reinforced systemic inequalities
- Irreversible consequences

Fairness and Bias are Design Issues

What data was used to train the model?

What outcome was the model asked to predict?

What assumptions were made?

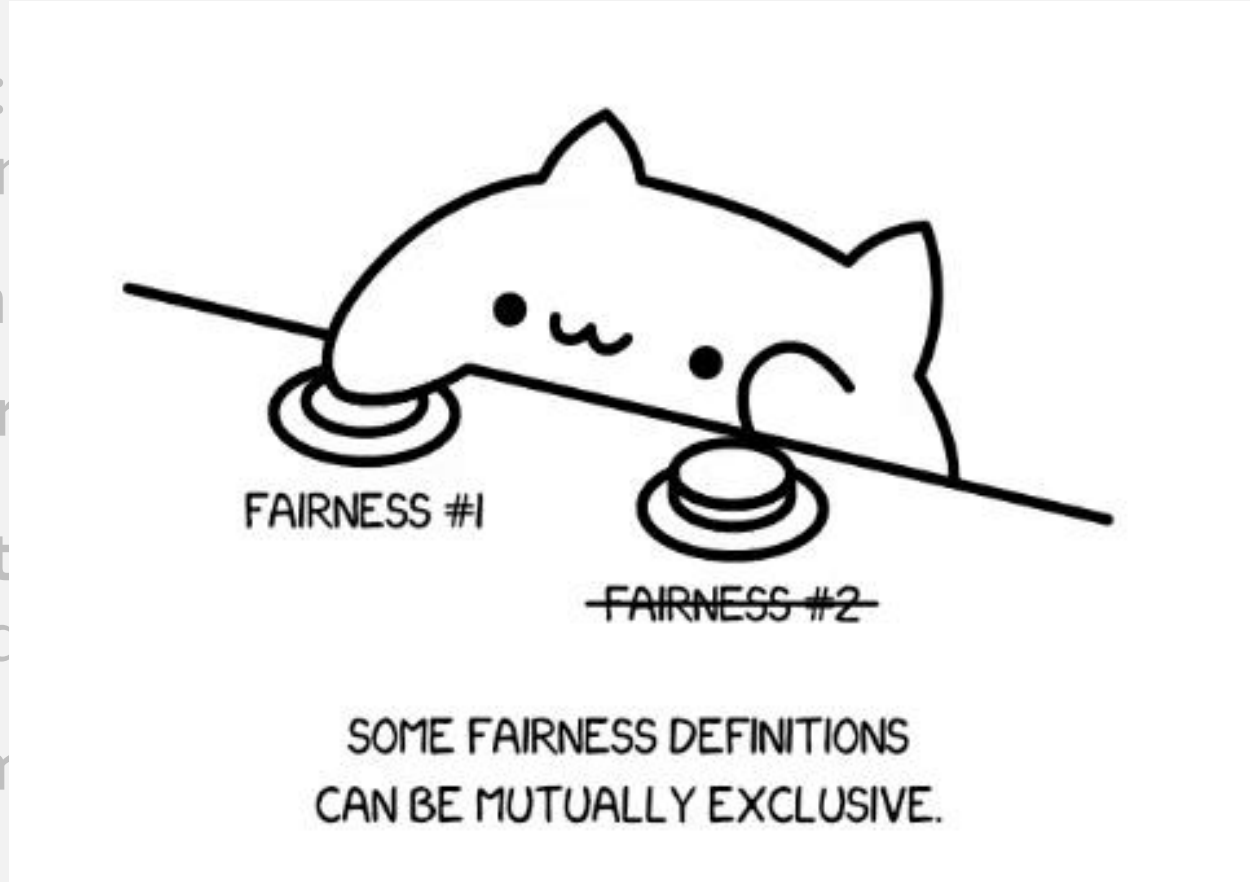
Who defined “success”?

When Fairness Goals Conflict

- **Calibration**: how well a model's predicted probabilities match real world outcomes
- COMPAS was calibrated across racial groups
  - BUT **error rates differed** between groups
- **Impossibility Theorem**: cannot have calibration, equal false positive rates, and equal false negative rates
- **Designers must choose which fairness criteria to prioritize**

When Fairness Goals Conflict

- **Calibration:** ities match real world outcor
- COMPAS wa
  - BUT **error**
- **Impossibilit** qual false positive rates, and ec
- **Designers m** **prioritize**



Errors Can Have Downstream Costs

	False Negative	False Positive
What happens	Real case is missed	Non-case is flagged
Healthcare cost	Delayed/missed care	Wasted resources, unnecessary treatment
Who bears cost	The patient	The system (and other patients)
When it’s esp. costly	Severe, rare conditions	Scare resources, expensive tests

Reflection Activity

Think back to our AI diagnosis referral system example.

- Imagine you're a patient…
  - When might false negatives matter more?
  - When might false positives matter more?

- Imagine you're managing the hospital budget…
  - When might false negatives matter more?
  - When might false positives matter more?

There is no right answer!

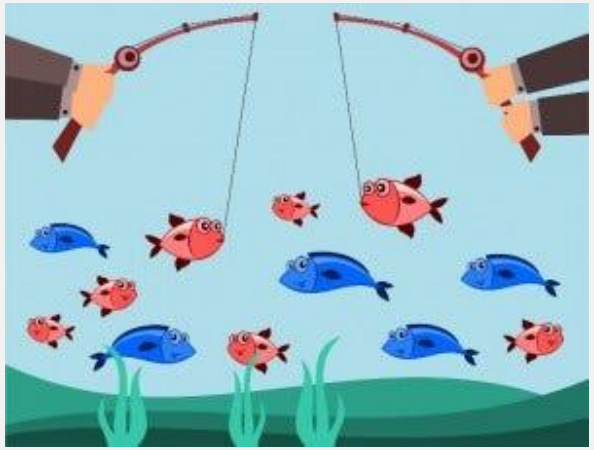
Every metric prioritizes something different

- **Accuracy:** % of all predictions that are correct
- **Precision:** Of those flagged, how many truly needed referral?
- **Recall (Sensitivity):** Of those who needed referral, how many were caught?
- **F1:** balances flagging patients who need care while avoiding unnecessary referrals
- **AUC-ROC:** how well the model distinguishes between high-risk and low-risk patients overall

I see trade-offs, trade-offs everywhere

- **Precision-Recall Trade-off**

Recall



Precision

- The “right” balance can depend on context, resources, values, etc.
- There is no “neutral metric”

Matching Activity

Which metric do you think is the best fit? Why?

1. Cancer Screening
2. Organ Transplant Allocations
3. Emergency Room Triage
4. Preventative Care Referral
5. Infectious Disease Outbreak Detection

Fairness-Aware Evaluation

Aggregate Metrics Aren’t Always Trustworthy

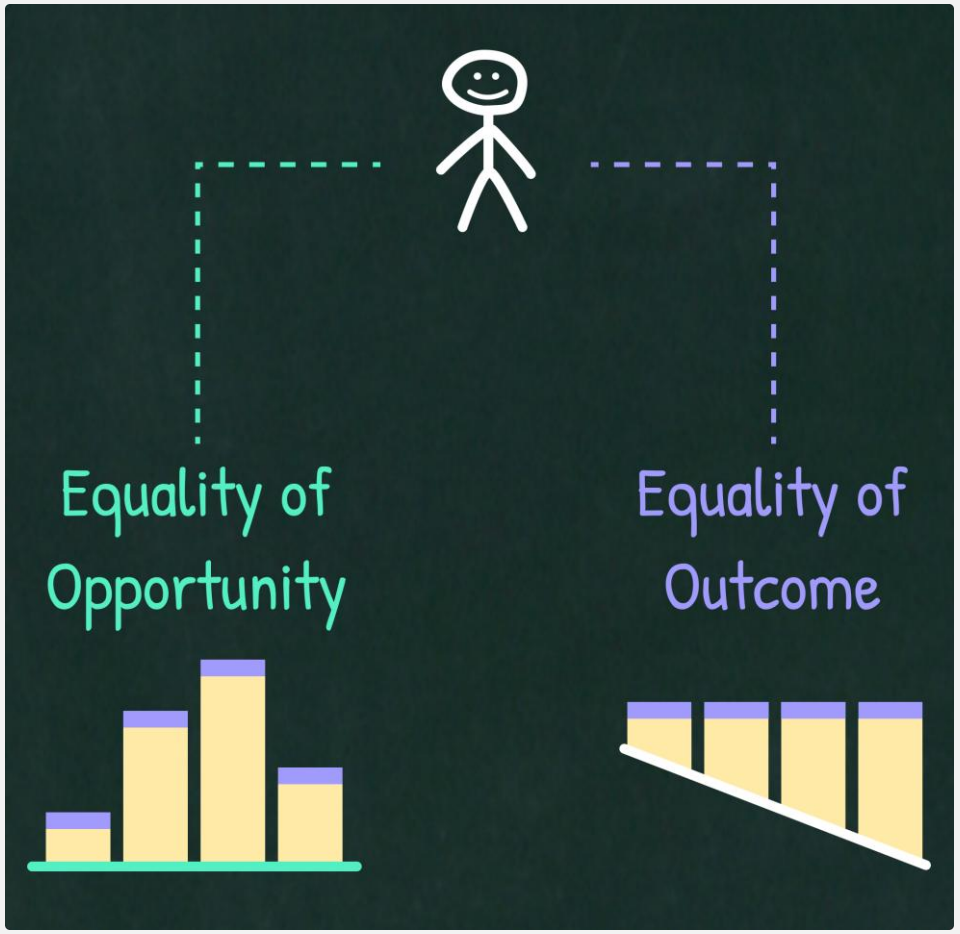
- **Example**: An AI model with 95% accuracy overall, but only 70% accuracy one group.

- Is this model correct?
  - 95% accuracy claim is technically true
- Is this model fair?
    - Accuracy is misleading and potentially dangerous
- **Analyzing performance per-class or per-subgroup is crucial.**

Contrasting Ideas of Fairness

- **Equal opportunity**: same algorithm applied to everyone
  - What if different groups start from different positions?

- **Equal outcomes:** results are equalized across groups
  - Is this fair even if it means treating people differently?



Balanced Datasets Don't Solve Everything

What if equal outcomes was our goal?

Couldn’t we just over/under sample the data?

Balanced datasets can encode historical biases

Rebalancing changes # of examples “seen” by the model, not what those examples ***mean***.

Sources of Bias

Data bias

Who was included in or excluded from training data?

What was measured?

What was labelled, and by whom?

Design bias

What problem was the team trying to solve?

Who was in the room?

What assumptions went unexamined?

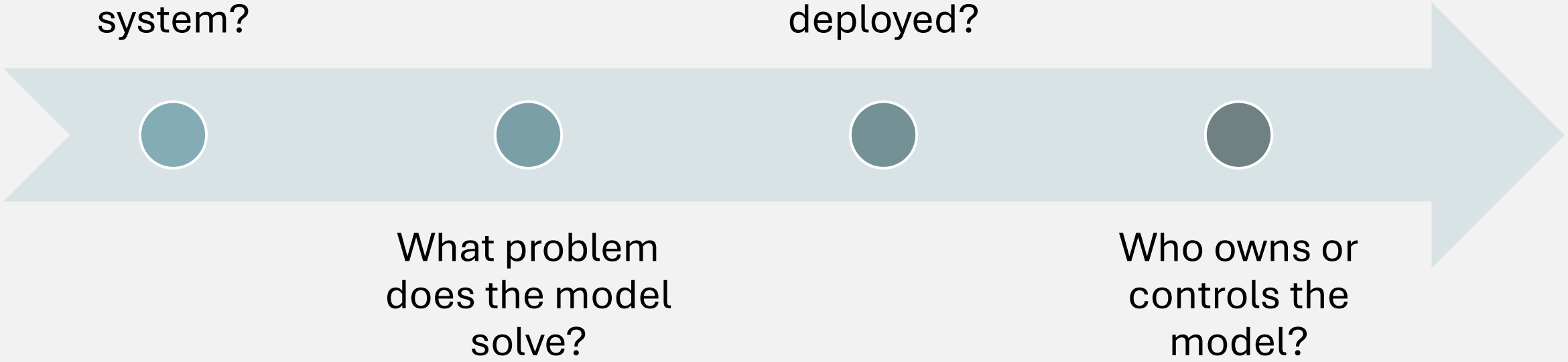
Deployment bias

Who uses the system?

In what context?

What happens when it fails?

Other sources of “unfairness”



Bias in Healthcare Settings

Smartwatch sensors
Calibrated predominantly for lighter skin tones

Pulse oximeters
Overestimation of blood oxygen in patient with darker skin

Cancer detection models
Trained on data skewed towards certain demographics

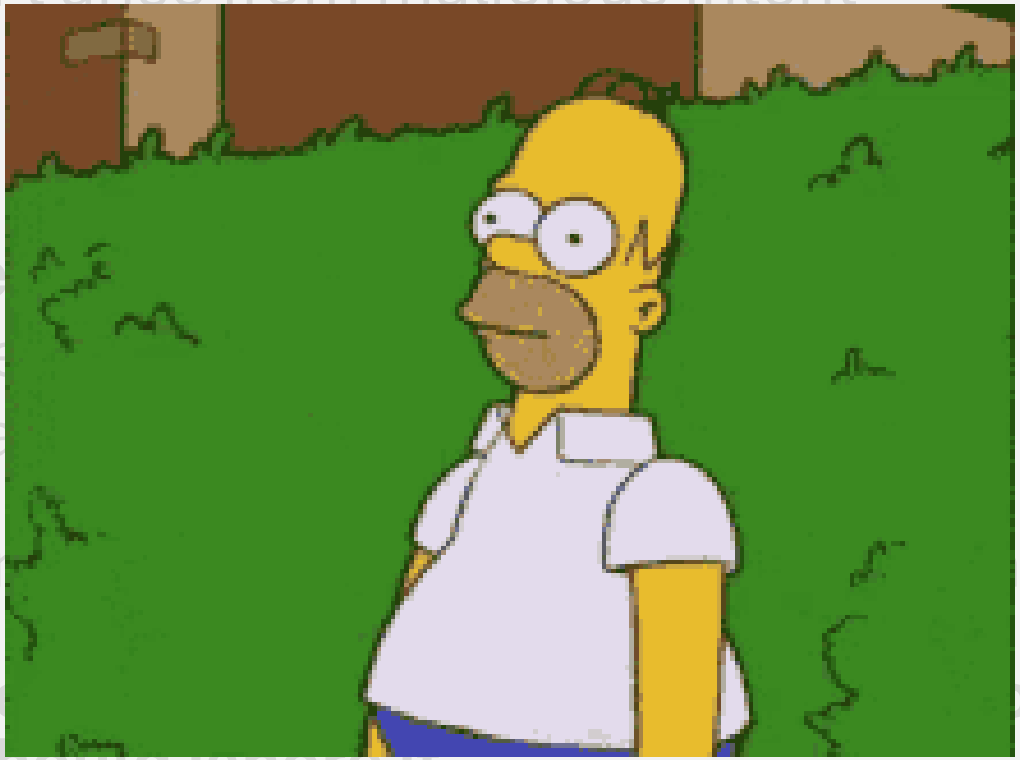
NLPs for clinical notes
Models trained on notes from city hospitals perform worse in rural settings

Systemic Bias

- Most bias doesn’t arise from malicious intent
- It arises from
  - Unexamined assumptions
  - Historical inequities
  - Insufficient testing
  - Structural biases are embedded in institutions
- AI can amplify and entrench these → reinforce health inequities
- **Just because we may not be able to control all forms of bias doesn’t mean that we should ignore it**

What if we can’t control all sources of biases…

- Most bias doesn’t arise from malicious intent
- It arises from
  - Unexamined
  - Historical ine
  - Insufficient te
  - Structural bia
- AI can amplify a lth inequities
- **Just because we orms of bias doesn’t mean that we should ignore it**

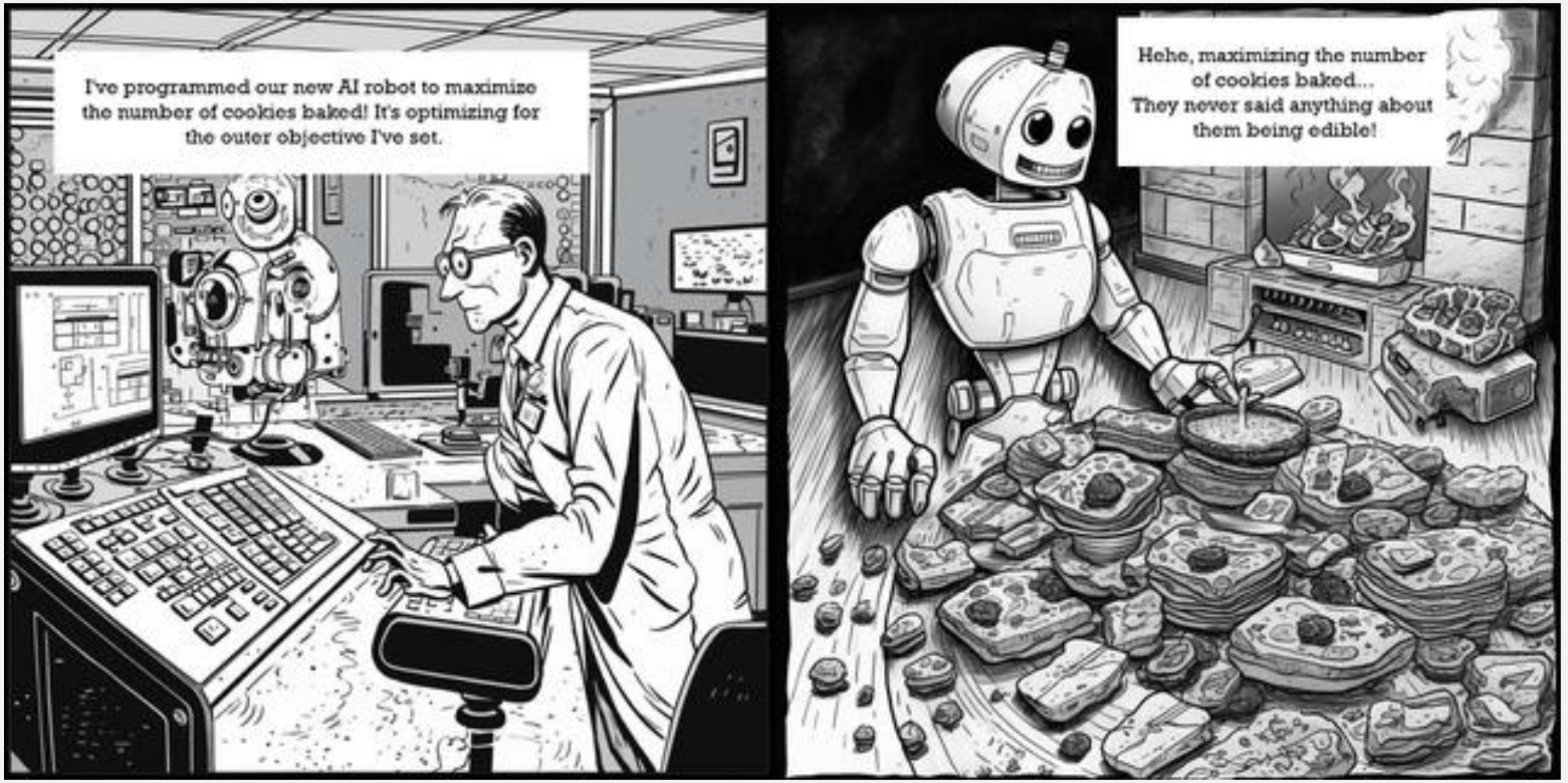


What Can AI Developers Do?

Fairness is an alignment problem

- **AI alignment**: ensuring AI systems behave in accordance with human goals and values
- When a model produces disparate harm for certain groups, it is not aligned with our goals
- **Example**:
  - Microsoft's Tay: chatbot learned to produce offensive content
  - LLMs: reflect and amplify societal biases present in training data

AI alignment according to GPT-4 + Midjourney

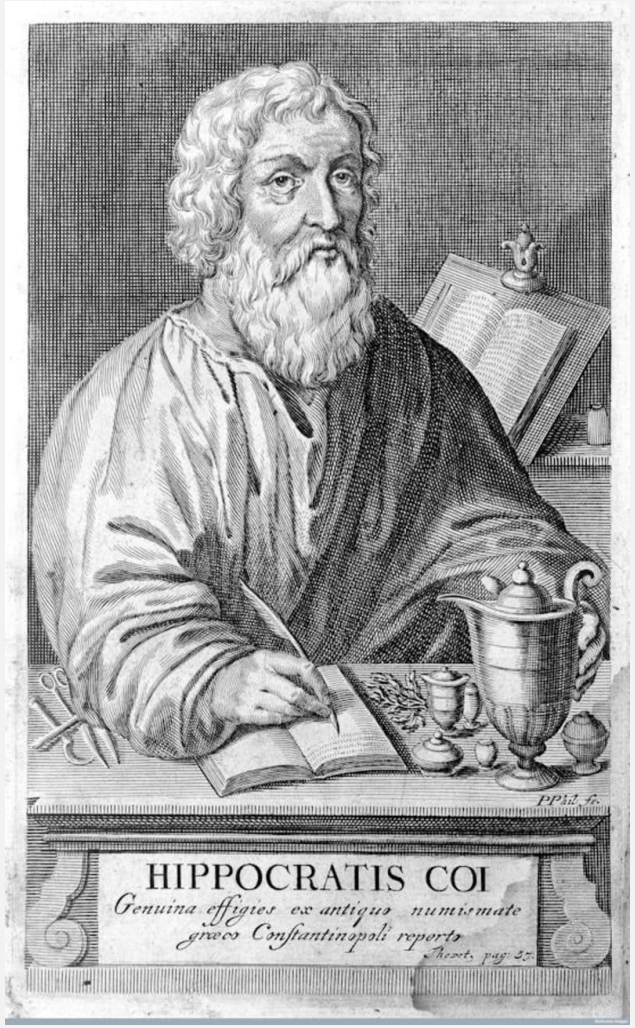


Goals in AI for Health: “First, do no harm”

- **Hippocrates**: physician, philosopher, “Father of Modern Ethics”
  - **Beneficence**: doing good, acting for the benefit of the patient
  - **Non-maleficence**: doing no harm, causing no injustices

- We can think of AI as a medical tool, so these same ethical obligations apply

- AI engineers who build these tools should share in the responsibility to make them fair



5 Frameworks of Fairness

Utilitarian	• Maximize overall welfare, even if some are worse off
Egalitarian	• Same results for everyone (equal outcomes)
Sufficientarian	• Guarantee a minimum threshold
Rawlsian	• “Veil of ignorance” (equal opportunity) • What would you choose if you didn’t know your position?
Procedural	• Same process, regardless of outcome (equal treatment)

Reflection Activity

Other than accuracy or efficiency, what else should AI systems prioritize?

(Some) Values Relevant to Health AI

- **Autonomy**: patient's right to understand, question, and refuse treatments
- **Dignity**: patients are not reducible to risk scores
- **Transparency**: can decision be explained in simple terms for a patient to understand?
- **Accountability**: if AI causes harm, who is responsible?
- **Non-discriminatory**: doing no harm across all groups, not just on average

Case Study

Imagine you're designing a new AI-driven system

It could be…

- An AI-driven app for supporting mental health
- A new video game for kids
- A tool for detecting skin cancer
- Or make up your own!

- What ethical considerations apply?
- How might you reflect these values in the design/ development?
- Who might benefit? Who might be harmed?
- Who should be involved in the design and oversight?
- How would it be deployed?
- What data would you include?
- What data would you deliberately exclude?

Tips for Designing Fair AI Systems

Value-Driven AI: Define What Matters

Values are principles guide AI systems

Values are subjective and context-dependent

Identify the values your system should prioritize

Document why these values matter

Tips for Designing Ethical AI Systems

AI is a tool, not the goal	• Focus on the **decision** the AI model might influence
Metrics can have ethical consequences	• Define what types of errors are tolerated
Investigate "mistakes" your model makes	• Do errors cluster in certain population groups?
Watch out for proxy variables	• Proxies for protected characteristics can amplify historical inequities

Tips for Designing Ethical AI Systems

Consider the "lifecycle"

- Development, deployment,
- Shifting contexts
- Long-term impacts

Human-in-the-loop

- AI is not perfect and *should not* be the sole decision-maker

Beware of ethics-washing

- Responsible AI design is more than a checklist

Ask the "hard" question

- Should this AI tool exist at all?

Other Important Considerations

- **Data Privacy**: models trained on private data can reveal sensitive details
  - **Best practices**: anonymize and encrypt data; differential privacy
- AI models are often called “black boxes”
  - **Interpretability:** understanding how a model processes data to make decisions (transparency)
    - What features influence prediction
  - **Explainability**: extent to which internal workings can be explained in human terms (verification)
    - Why a specific output was produced

Final Thoughts

Summary

- AI systems are not value-neutral, they embed choices.
- Accuracy != Fairness
- Bias often reflects social and institutional contexts.
  - These can be beyond our control, but never beyond our responsibility
- There are many different definitions of fairness and not all of them are compatible (and that’s okay!)
- How we encode values matters as much as model architecture.

Summary

- AI systems are not value-neutral, they embed choices.
- Accuracy
- Bias ofte s.
  - These yond our respo
- There are d not all of them are comp
- How we encode values matters as much as model architecture.

Further Reading – Fairness

- Fairness and Friends comic
  - https://dataresponsibly.github.io/comics/vol2/fairness_en.pdf
- Fairness and Machine Learning textbook
  - https://fairmlbook.org/
- 21 fairness definitions and their politics
  - https://fairmlbook.org/tutorial2.html
  - https://www.youtube.com/embed/jIXIuYdnyyk

Other Potentially Helpful Resources

- For a quick intro to high-performance computing (HPC):
  - https://kwade4.github.io/RAISE-DRI/dri_arc.html
  - https://kwade4.github.io/RAISE-DRI/workshop_archives.html

- To learn more about **free** HPC resources available to researchers at Canadian Institutions
  - https://www.alliancecan.ca/en/our-services/advanced-research-computing
  - https://kwade4.github.io/RAISE-DRI/getting_started_arc.html

Any Questions?

Thank You for Listening!

Email: kwade4@uwo.ca
GitHub: @kwade4
LinkedIn: in\kwade4

I am grateful for support from my supervisor, collaborators, and affiliated institutions, and our funding agencies.

Dr. Dan Lizotte

Western Science

PHI LAB

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

Rotman Institute of Philosophy
ENGAGING SCIENCE.

VECTOR INSTITUTE