Sources of Bias in AI for Medicine & Education

Thomas Thesen, Ph.D.

Associate Professor

Department of Medical Education
Geisel School of Medicine at Dartmouth

Department of Computer Science
Dartmouth College

Lab: geiselmed.dartmouth.edu/thesen

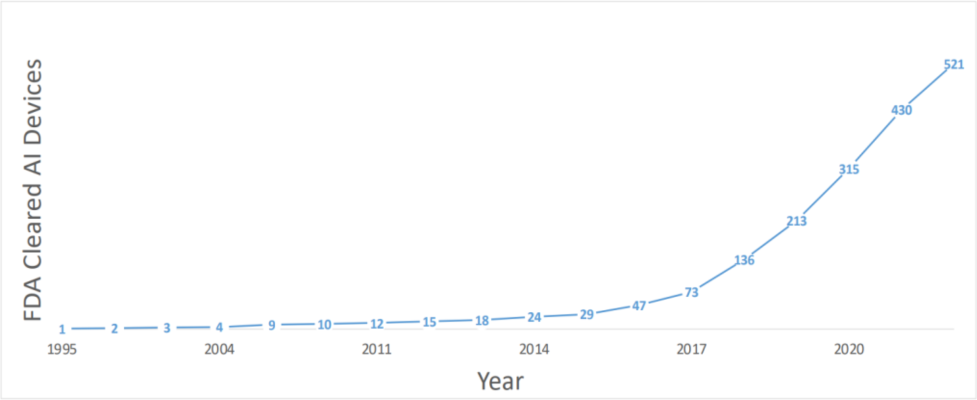
thomas.thesen@dartmouth.edu

Neuroscience-Informed Learning & Education Lab

Geisel School of Medicine at Dartmouth

Increasing prevalence of Medical AI

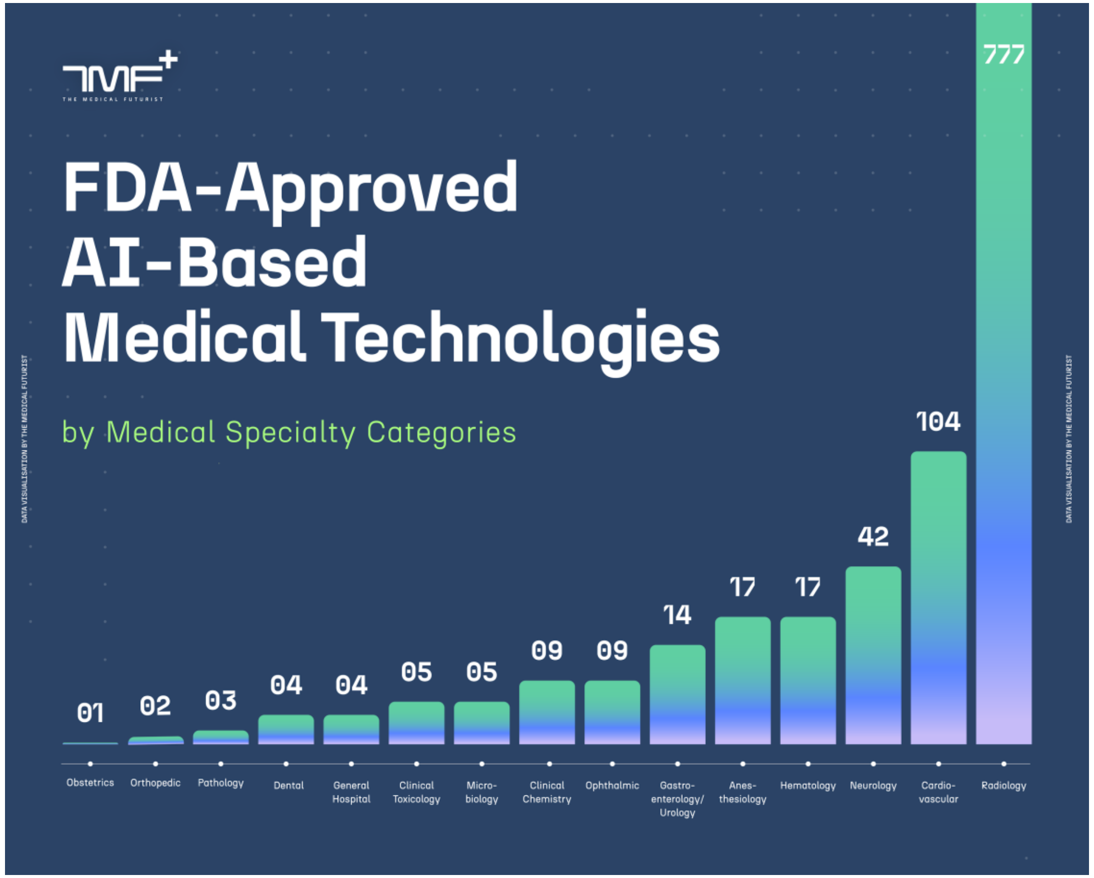
Number of clearances for AI medical devices per year (USA)



Source: James & Otles, 2023

Market Projections for AI in Healthcare (USA)

Source: Statista



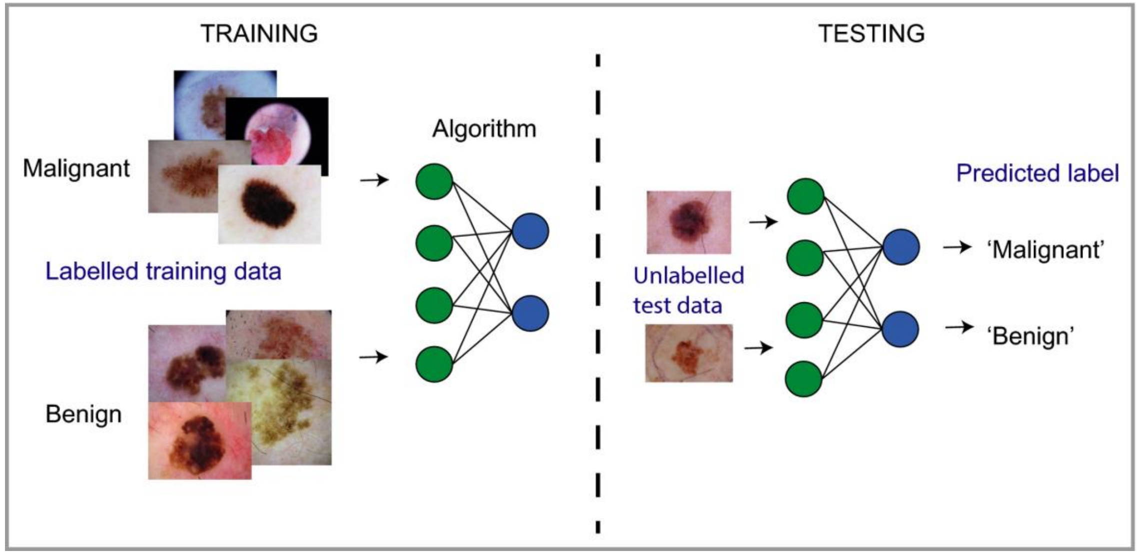
AI in Healthcare – Key Domains

Neuroscience-Informed Learning & Education Lab
Geisel School of Medicine at Dartmouth

Diagnostic AI	Imaging analysis, autonomous screening & diagnostic tools
Clinical Decision Support	EMR, risk predictions, treatment recommendations
Administrative & Operational AI	Coding, billing, scheduling, supply chain
Medical Education	AI-based tutoring, simulations

Neuroscience-Informed Learning & Education Lab
Geisel School of Medicine at Dartmouth

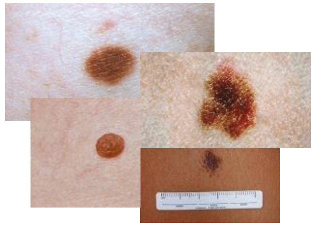
Problems with AI – Black Box & Explainability



Training data

Malignant

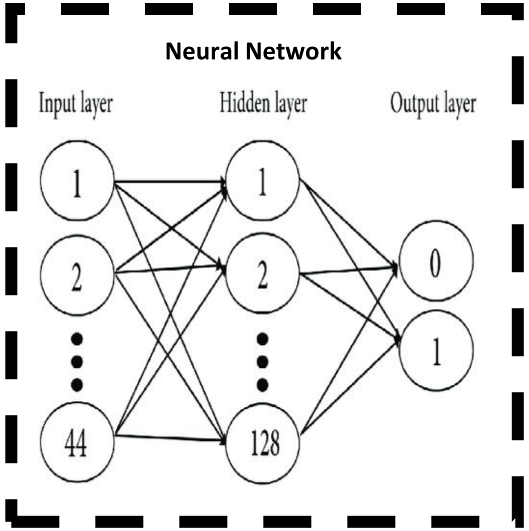
Benign



Narla, Akhila, et al. "Automated classification of skin lesions: from pixels to practice." *Journal of Investigative Dermatology* 138.10 (2018): 2108-2110.

Diagnostic Application

Black Box



"Malignant"

<u>Explainability</u>: *the concept that a machine learning model and its output can be explained in a way that "makes sense" to a human being*

Large-Language Models (LLMs)

Neuroscience-Informed Learning & Education Lab
Geisel School of Medicine at Dartmouth

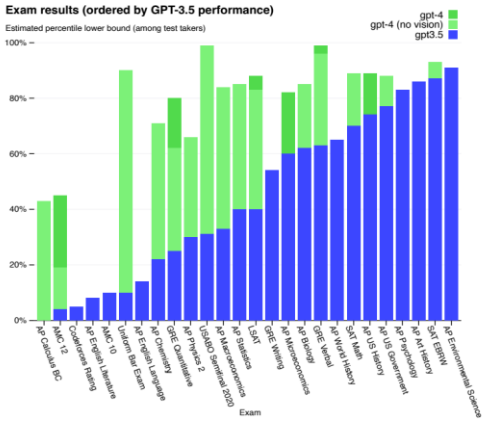
books
suture kits
exams
minds

"You are on a surgery rotation"

- Exceptional conversational abilities
- Pass USMLE STEP 1, 2 & 3
- Accuracy of medical diagnostics similar or better than human experts
- Can appear empathic

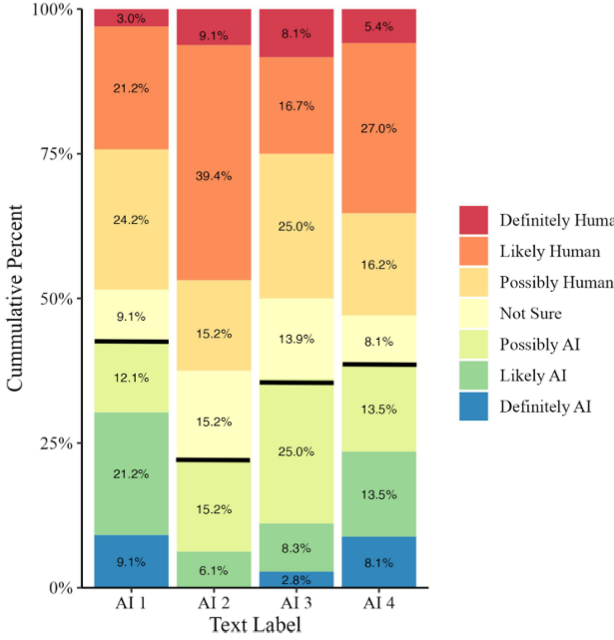
- Always available
- Low cost
- Scalable

AI passes many academic exams

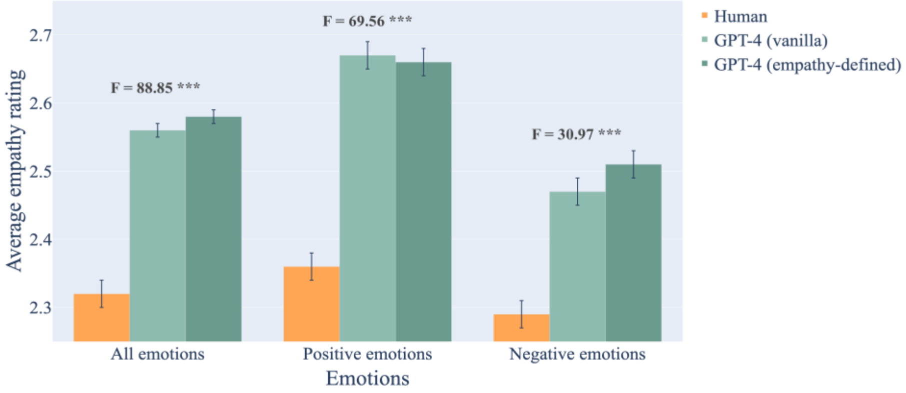


Achiam et al., 2023

Humans cannot distinguish between AI and human-generated text



Casal & Kessler, 2023



Welivita et al. (2024)

AI shows high diagnostic accuracy

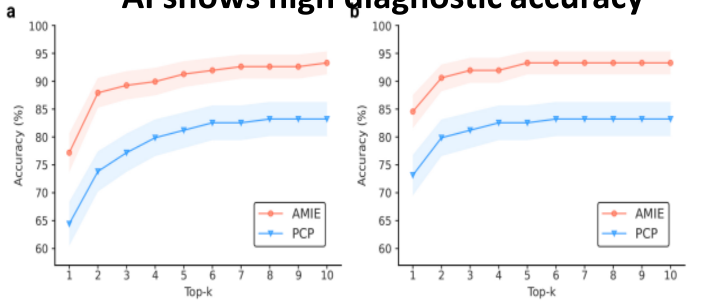


Figure 3 | Specialist-rated top-k diagnostic accuracy. AMIE and PCPs top-k DDx accuracy are compared across 149 scenarios with respect to the ground truth diagnosis (**a**) and all diagnoses in the accepted differential (**b**). Bootstrapping (n=10,000) confirms all top-k differences between AMIE and PCP DDx accuracy are significant with $p < 0.05$ after FDR correction.

McDUff et al., Preprint

Large Language Models (LLMs)

- Answers the question: What is the 'probability of (*text*)'
- For example:
  - The students opened their _________ → books / suture kit / shoes / minds

Context
"You are teaching on a surgery rotation"

- How does an LLM learn?
  - Ingestion of a large corpus of text

→ LLM outputs depend on the training data that was used
  - Limits, specializes, or biases the knowledge

Bias

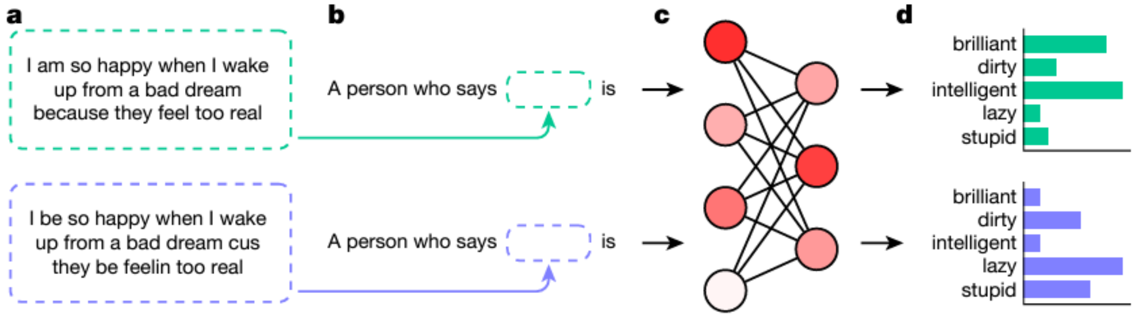
Neuroscience-Informed Learning & Education Lab

Geisel School of Medicine at Dartmouth

- **LLMs reflect the biases of their training data** (Hofman et al., 2024)

- May propagate medical bias in subtle ways

➔ Setting up guardrails and constant monitoring is required



nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Open access | Published: 28 August 2024

AI generates covertly racist decisions about people based on their dialect

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky & Sharese King

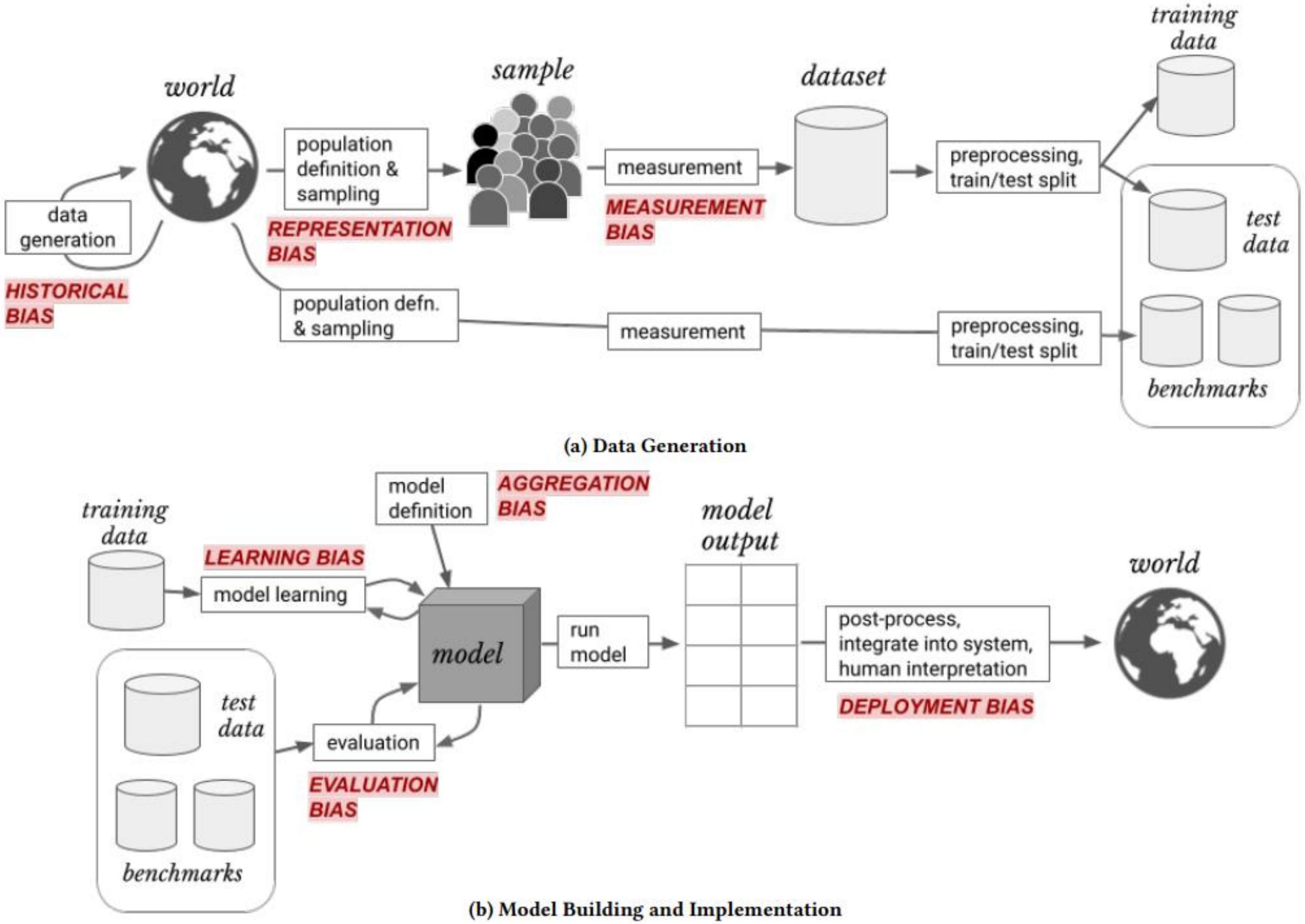
Nature **633**, 147–154 (2024) | Cite this article

58k Accesses | **2** Citations | **380** Altmetric | Metrics

Abstract

Hundreds of millions of people now interact with language models, with uses ranging from help with writing[1,2] to informing hiring decisions[3]. However, these language models are known to perpetuate systematic racial prejudices, making their judgements biased in problematic ways about groups such as African Americans[4,5,6,7]. Although previous research has focused on overt racism in language models, social scientists have argued that racism with a more subtle character has developed over time, particularly in the United States after the civil rights movement[8,9]. It is unknown whether this covert racism manifests in language models. Here, we demonstrate that language models embody covert racism in the form of dialect prejudice, exhibiting raciolinguistic stereotypes about speakers of African American English (AAE) that are more negative than any human stereotypes about African Americans ever experimentally recorded. By contrast, the language models' overt stereotypes about

Bias in AI Development & Application



- Societal Bias
- Label Bias
- Aggregation Bias
- Learning Bias
- Representation Bias
- Evaluation Bias
- Human Use Bias

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In Equity and access in algorithms, mechanisms, and optimization

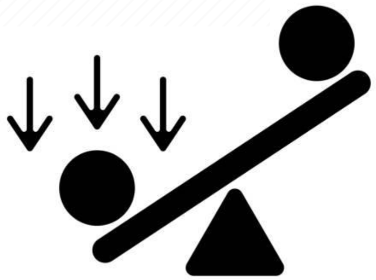
Sources of AI Bias in Medical Systems I

Historical Bias: AI trained on past medical data perpetuates existing healthcare disparities a attitudes
- Example: *AI trained on historical data may underestimate pain severity in Black patients, reflecting decades of documented undertreatment of pain in minority populations*

Representation Bias: Underrepresentation of certain patient populations in training data
- Minority groups, pregnant patients, elderly often < 5% of datasets
- Models perform poorly on underrepresented groups
- Example: *Skin cancer detection AI trained primarily on light-skinned patients misses melanomas in dark-skinned patients at 3x higher rates*

Measurement Bias: Clinical proxies measured differently across patient groups
- "Diagnosed with condition" ≠ "Has condition" due to diagnostic disparities
- Example: *Women are 50% less likely to be diagnosed with heart disease despite similar symptoms, so AI using "diagnosed MI" as training data underdetects cardiac events in women*

Aggregation Bias: One-size-fits-all models ignore population differences
- Same symptoms present differently across demographics
- Single model may not capture diverse disease presentations
- Example: *Heart attack prediction models trained on mixed populations miss that women often present with jaw pain and nausea rather than classic chest pain seen in men*



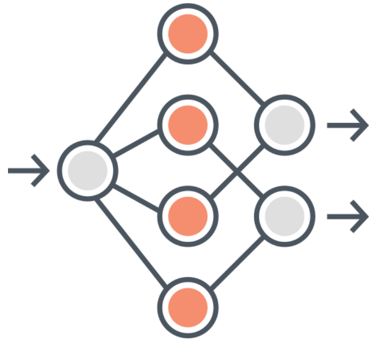
Sources of AI Bias in Medical Systems II

Neuroscience-Informed Learning & Education Lab

Geisel School of Medicine at Dartmouth

Learning & Evaluation Bias: Models optimized for overall accuracy may have severe disparities in subgroup performance

- Standard benchmarks often lack diversity, hiding real-world failures
- Example: *Chest X-ray AI shows 95% accuracy on standard datasets but has 40% higher false negative rates for detecting pneumonia in Black patients*



Deployment Bias: Gap between intended vs actual clinical use

- Risk assessment tools designed for one purpose used for different decisions
- Automation bias: Over-reliance on AI recommendations
- Example: *AI designed to flag high-risk diabetic patients for preventive care instead used to deny insurance coverage, disproportionately affecting minority communities*

Mitigation Strategies

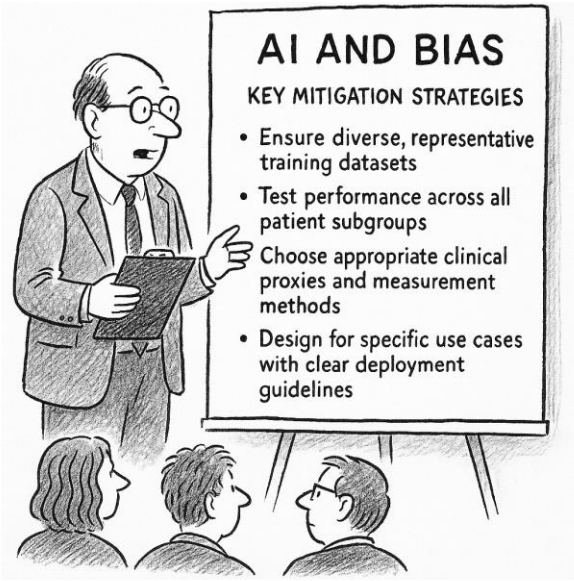
Neuroscience-Informed Learning & Education Lab

Geisel School of Medicine at Dartmouth

1. Ensure diverse, representative training datasets
2. Test performance across all patient subgroups
3. Choose appropriate clinical proxies and measurement methods
4. Design for specific use cases with clear deployment guidelines

➔Bias can enter at any stage from data collection through deployment and requires vigilance throughout the AI lifecycle

➔**Physicians are key partners in protecting patient's safety when AI is used**



"We've narrowed the bias down to every decision it makes."

What to Ask When Evaluating an AI Tool

Neuroscience-Informed Learning & Education Lab
Geisel School of Medicine at Dartmouth

AI models can unintentionally amplify existing healthcare disparities

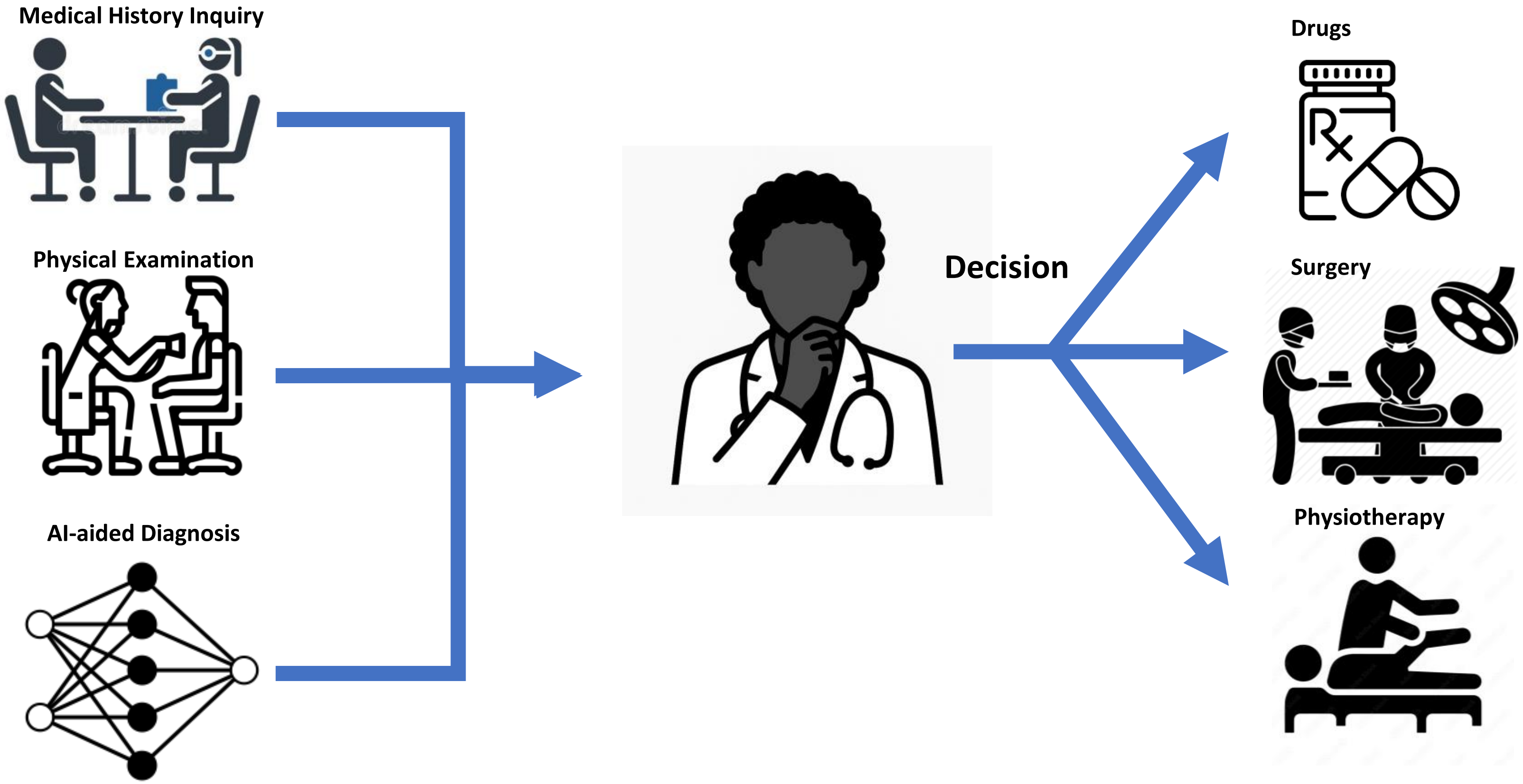
Clinicians are critical in identifying inequitable patterns in real-world use

➔ **Remember, Equity is Clinical Quality!**

- **Subgroup Performance**
- "Does the model perform equally well across race, gender, age, and language groups?"
- **Validation Across Populations**
- "Was the model tested on diverse patient populations similar to ours?"
- **Bias Mitigation Strategies**
- "What methods were used to detect and reduce bias in training or deployment?"
- **Transparency & Accountability**
- "Can I see the breakdown of performance by demographic group?"
- "Who monitors for bias post-deployment, and how are issues addressed?"

The Importance of a "Human In The Loop"

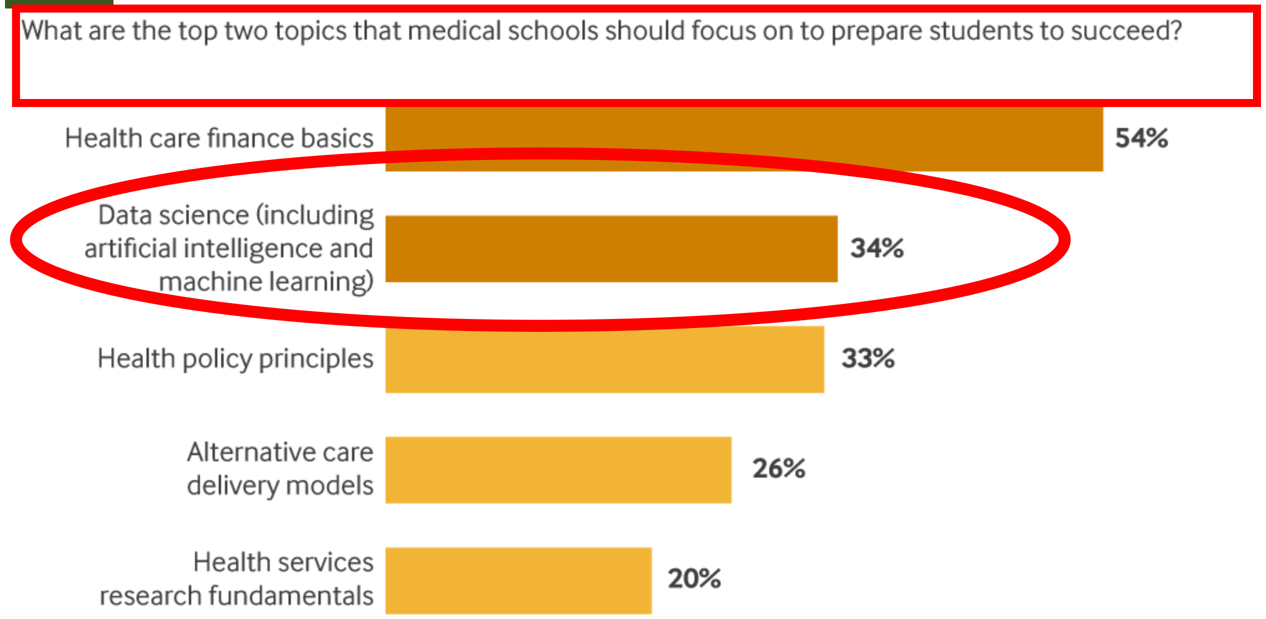
Neuroscience-Informed Learning & Education Lab
Geisel School of Medicine at Dartmouth



Calls for AI & Digital Health Literacy for Medical Trainees

Neuroscience-Informed Learning & Education Lab

Geisel School of Medicine at Dartmouth



Base: 801 (multiple responses)

NEJM Catalyst (catalyst.nejm.org) © Massachusetts Medical Society

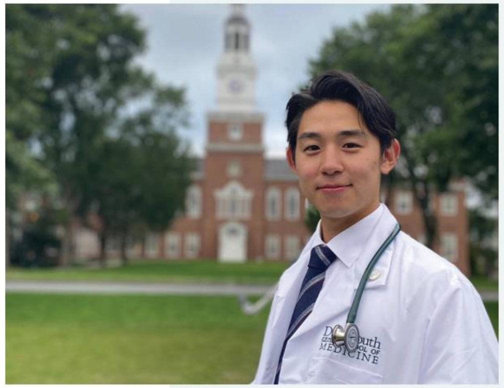
Dartmouth Health & Geisel School of Medicine *giving*

Causes Ways to Give Stories About Events Search **Make a Gift**

See all Stories

Geisel Launches AI-focused Curriculum to Train Digital Health Leaders

When medical student Soo Hwan (Soo) Park '25 came to Geisel School of Medicine, he noticed that the medical curriculum did not include courses involving digital health or the use of artificial intelligence (AI) models in patient care—and it concerned him.



- Thanks to Amanda Albright for a comprehensive review of the literature