# Sources of Bias in Al for Medicine & Education

Thomas Thesen, Ph.D.

**Associate Professor** 

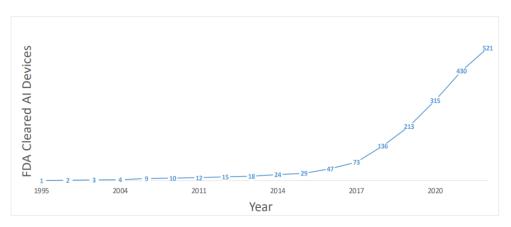
Department of Medical Education
Geisel School of Medicine at Dartmouth

Department of Computer Science
Dartmouth College

## Increasing prevalence of Medical Al

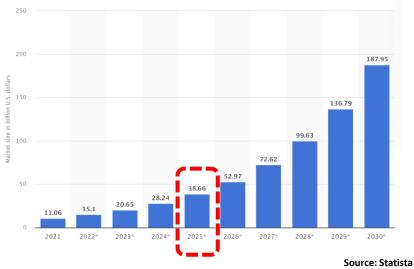


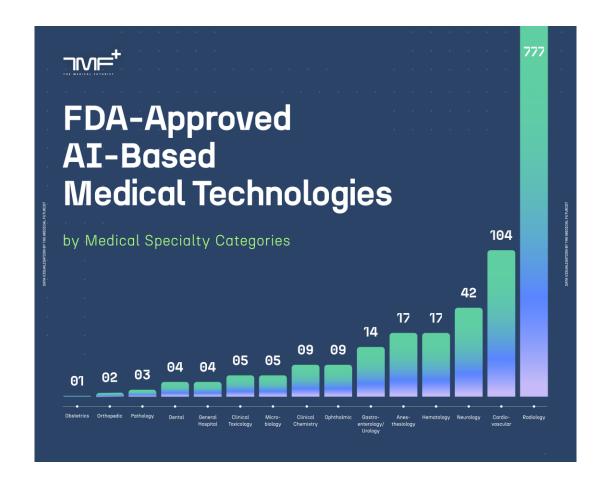
### Number of clearances for AI medical devices per year (USA)



Source: James & Otles. 2023

#### **Market Projections for AI in Healthcare (USA)**





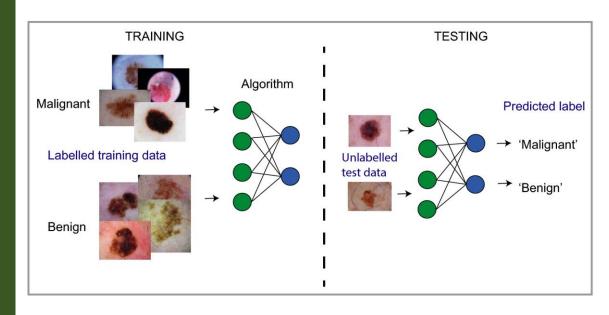
## Al in Healthcare – Key Domains



U	Diagnostic Al	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

### **Problems with AI – Black Box & Explainability**



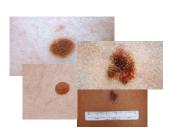


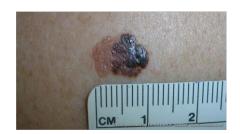
#### **Training data**

Malignant



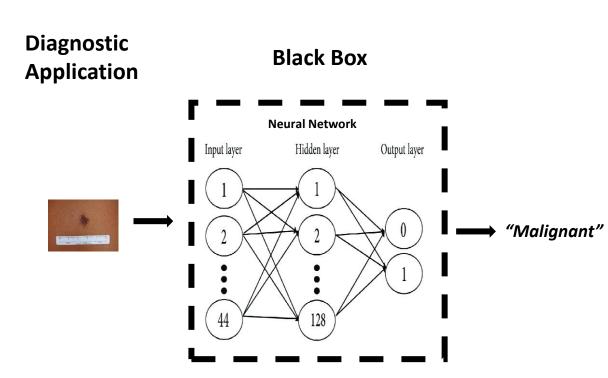
Benign







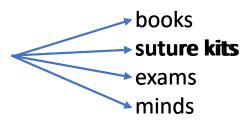
Narla, Akhila, et al. "Automated classification of skin lesions: from pixels to practice." *Journal of Investigative Dermatology* 138.10 (2018): 2108-2110.



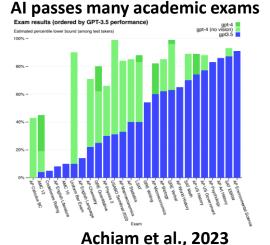
**Explainability:** the concept that a machine learning model and its output can be explained in a way that "makes sense" to a human being

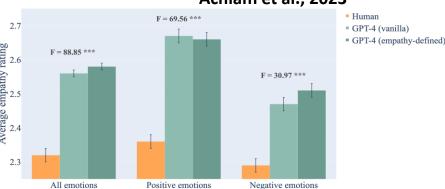
## Large-Language Models (LLMs)

- 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



"You are on a surgery rotation"

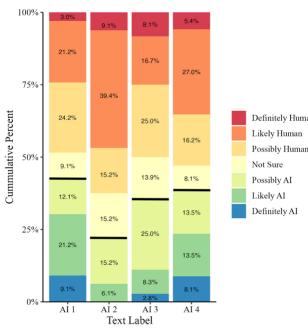




Welivita et al. (2024)



### Humans cannot distinguish between AI and human-generated text



Casal & Kessler, 2023

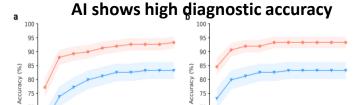


Figure 3 | Specialist-rated top-k diagnostic accuracy. AMIE and PCPs top-k DDx accuracy are compared across 14 scenarios with respect to the ground truth diagnosis (a) and all diagnoses in the accepted differential (b). Bootstrappin (n=10,000) confirms all top-k differences between AMIE and PCP DDx accuracy are significant with p < 0.05 after FDE

McDUff et al., Preprint

## Large Language Models (LLMs)



- Answers the question: What is the 'probability of (text)'
- For example:
  - The students opened their \_\_\_\_\_sut



- 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

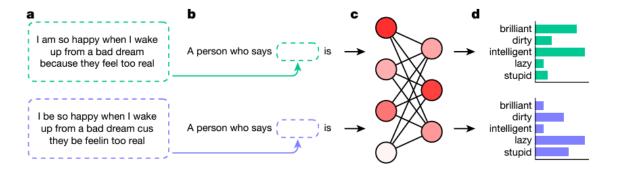
#### Context

"You are teaching on a surgery rotation"

### Bias



- 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

### Al generates covertly racist decisions about people based on their dialect

Valentin Hofmann <sup>™</sup>, Pratyusha Ria Kalluri, Dan Jurafsky & Sharese King <sup>™</sup>

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

58k Accesses | 2 Citations | 380 Altmetric | Metrics

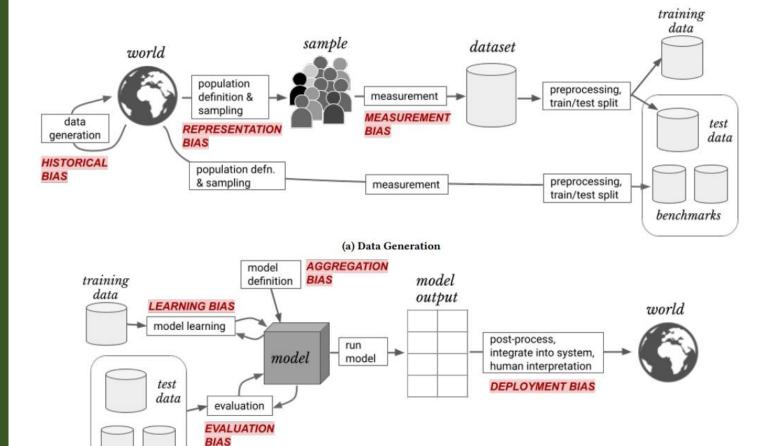
#### Abstract

Hundreds of millions of people now interact with language models, with uses ranging from help with writing \$\frac{1}{2}\$ to informing hiring decisions \$\frac{3}{2}\$. However, these language models are known to perpetuate systematic racial prejudices, making their judgements biased in problematic ways about groups such as African Americans \$\frac{4}{2}\frac{5}{2}\frac{1}{2}\$. 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 \$\frac{8}{2}\$. 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

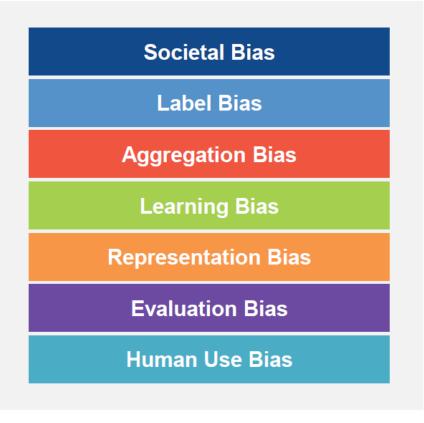
## Bias in Al Development & Application



Geisel School of Medicine at Dartmouth



(b) Model Building and Implementation



benchmarks

### Sources of Al Bias in Medical Systems I

**Historical Bias**: Al trained on past medical data perpetuates existing healthcare disparities a attitudes

• Example: Al 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



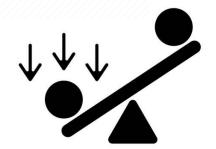
ol of Medicine Dartmouth











### Sources of Al Bias in Medical Systems II

Neuroscience-Informed
Learning & Education
Lab

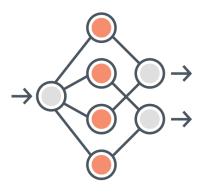
Geisel
School of Medicine
at Dartmouth

**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: Al designed to flag high-risk diabetic patients for preventive care instead used to deny insurance coverage, disproportionately affecting minority communities



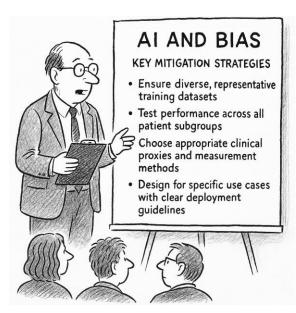


## **Mitigation Strategies**

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
- 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 Al 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 Al Tool



Al models can unintentionally amplify existing healthcare disparities

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

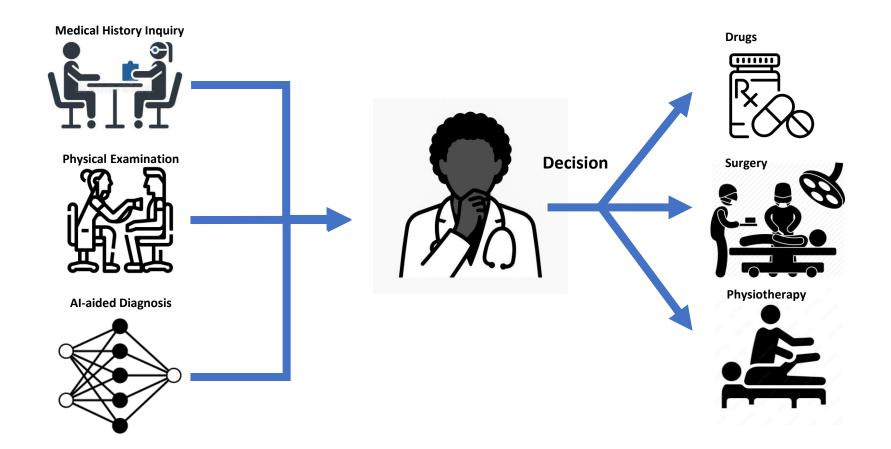
**→** Remember, Equity is Clinical Quality!



- 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"





### **Calls for AI & Digital Health Literacy for Medical Trainees**



Geisel

School of Medicine at Dartmouth

The Empirical Challenge of 21st-Century Data Science as a Core Competency in Undergraduate Medical Medical Education Education in the Age of Artificial Intelligence in Health Care Puneet Seth1\*, BSc, MD; Nancy Hueppchen2\*, MD; Steven D Miller3\*, MD; Frank Rudzicz4.56\*, PhD; Jerry Ding7 MMI, MD; Kapil Parakh8\*, MD; Janet D Record2\*, MD Facing challenges wrought by science and technology as well as societal change, the curriculum is increasingly physician and the sacrosanct doctorof probabilistic reasoning, and the <sup>1</sup>Department of Family Medicine, McMaster University, Hamilton, ON, Canada cultivation of empathy and compassic in accordance with ethical standards. Department of Gynecology and Obstetrics, Johns Hopkins University School of Medicine, Baltimore, MD, United State <sup>3</sup>Division of Pediatric Gastroenterology, Hepatology, and Nutrition, Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimor out of synch with new needs in teaching content and medical practice. The Given these needs, it is imperative 4 Faculty of Computer Science Dalhousie University Halifay NS Canada difficult because of a variety of factors, era. To address these challenges, the so doing, strive to make the hard accreditation process. Indeed, even the education: knowledge capture and 21st century. The author provides first very definition of what it means to be a professional is changing with profound Medical education has always been body of knowledge. But today the rapid with those outside the profession challenged to keep pace with advances in science and technology, and this is certainly but it is important to note that what is generally shared is information, dissemination of medical information outside of the profession is diminishing npj Digital Medicine

What do medical students actually need to know about artificial intelligence?

Liam G. McCoy()1251, Sujay Nagaraj()13, Felipe Morgado()14, Vinyas Harish()12, Sunit Das<sup>1,8</sup> and Leo Anthony Celi ()14, Sunit Das<sup>1,8</sup> and Sunit Das<sup>1,8</sup>

and future physicians about the technology is growing. Alongside comes the question of what, precisely, should medical students be taught. While competencies for the clinical usage of Al are broadly similar to those for any other novel technology, there are qualitative differences of critical importance to concerns regarding explainability, health equity, and data security. Drawing on experiences at the University of Toronto Faculty of Medicine and MIT Critical Data's "datathons", the authors advocate for a dualfocused approach: combining robust data science-focused additions to baseline health research curricula and extracurricula

npj Digital Medicine (2020)3:86; https://doi.org/10.1038/s41746-020-0294-

single restricted context may not always be transferable. It is also single restricted consess may not aways be disnessenate, it a migration to be aware of factors which may decrease the performance of algorithms for specific patient groups.\*

All has been commonly criticized for the "black box" effect—that is, the mechanism by which a model arrives at a decision may be With emerging innovations in artificial intelligence (Al) poised to substantially impact medical practice, interest in training current and future physicians about Al is growing1. Alongside this interest comes the guestion of what, precisely, medical students should learn\*. While competencies for the clinical usage of Al are broadly similar to those for any other novel technology in medicine, there are qualitative differences of critical importance to concerns indecipherable<sup>1</sup>. This lack of technical "explainability", however does not discharge the obligations of (iii). To satisfy requirements of informed consent and clinical collaboration, a physician may be called upon to communicate their understanding of the origin nature, and justification of an algorithm's results to patients regarding explainability, health equity, and data security<sup>3-3</sup>. We advocate for a dual-focused approach: combining robust, learner-centered Al additions to baseline curricula and extracurricular families, and colleagues. programs to cultivate leadership in this space.

WHAT DO PHYSICIANS NEED TO UNDERSTAND ABOUT AI IN THE CLINICAL CONTEXT?

Most directly, physicians need to understand Al in the same way Most directly, prysicians need to understand AI in the same with that they need to understand any technology impacting clinical decision-making. A physician utilizing MRI, for example, does not need to understand the particle spin physics differentiating T1 and T2 weighted scans, but they do need to be able to:

Use it—identify when the technology is appropriate for a given clinical context, and what inputs are required to receive meaningful results.
 Interpret it—understand and interpret the results with a

reasonable degree of accuracy, including awareness of sources of error, bias, or clinical inapplicability.

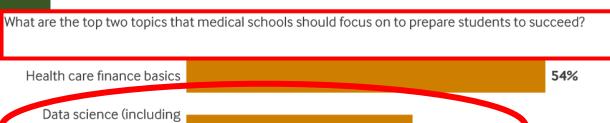
(iii) Explain it—be able to communicate the results and the processes underlying them in a way that others (e.g. alied health professionals and patients) can understand.

These skills take on particular nuances in the context of Al. For (i) and (ii), it is critical for physicians to appreciate the highly context-specific nature of Al, and the fact that performance in a

WHAT DO PHYSICIANS NEED TO UNDERSTAND ABOUT AI IN THE BROADER PROFESSIONAL CONTEXT? The professional obligations of physicians extend beyond the clinical role into leadership and health advocacy. The disruptive prospects of AI in healthcare raise significant ethical and operational challenges which physicians must collectively be

operational chainings which physicians must consciously be prepared to engage with for the sake of ensuring patient welfare. Substantial concerns exist regarding the impact of algorithmic clinical decision support on health equity, due to factors such as the use of datasets lacking representation from minority popula tions<sup>1</sup>, and the possibility for algorithms to learn from and perpetuate existing biases\*. Risks around data security and privac are also becoming rapidly apparent. There is also, however, the potential for Al itself to alleviate some of medicine's existing problems with bias and unfairness. Physicians should be aware of both possibilities and be equipped to advocate for the develop ment and deployment of ethical and equitable systems. Finally, physicians must act as responsible stewards for patient data to ensure that the foundational trust between provider and patient is

Vector Institute for Artificial Intelligence, Toronto, ON, Canada <sup>6</sup>Department of Computer Science, University of Toronto, Toronto, ON, Canada Schulich School of Medicine and Dentistry, Western University, London, ON, Canada Department of Medicine, Georgetown University, Washington, DC, United State JMIR MEDICAL EDUCATION Medical Student Training in eHealth: Scoping Review Jean-Francois Echelard<sup>1</sup>, MD: Francois Méthot<sup>1\*</sup>, BBA: Hue-Anh Neuven<sup>1\*</sup>, MD: Marie-Pascale Pomey<sup>2,3,4</sup>, MD <sup>3</sup>Department of Management, Evaluation and Health Policy, School of Public Health, Universit\(\tilde{e}\) de Montr\(\tilde{e}\), Montreal, QC, Canada <sup>4</sup>Department of Farmily Medicine and Emergency Medicine, Faculty of Medicine, Université de Montréal, Montreal, OC, Canada Cell Reports Medicine CelPress Teaching artificial intelligence as a fundamental toolset of medicine Estin Öttes, 1957: Comerius A., James, <sup>1</sup> Einberry D. Lomin, <sup>1</sup> and James O. Woolfstraff \*United Science Training Program, Unwards of Biologia Medical Science, Ann Arbor, M. USA \*Department of relocation and Commissions Engineering, University of Microgan, Ann Arbor, M. USA \*Department of Problems, University of Microgan, Ann Arbor, M. USA Antibial Intelligence (A) is transforming the practice of medicine. Systems assessing chest rackographs, perticulage class, and early warning systems embassied in intercruck hashin moorist. (DRR) and subcoming statistics of medical procedure in residual composition. Despite first, emblished moorist control to the consensation selection to distance and enablash All systems, learning them confor properties from future clinical practice. We must work procedure for the consensation of the consensation of the consensation, we grantly to distinct endergolastics in collection of the connection of the conne pensy to extend undergonate resource extend component of medical practice that is intro-one that medical estudators treat. All as a critical component of medical practice that is intro-and impossible with the other com-The promote of definition intelligence (20) or an offer production from the control of the contr Cell Regards Madissine J., 108034, December 39, 3082 0 2082 The Authority. 1
This is an open access white arrive the CC 87. NCAD between International Commission and Authority. 1



artificial intelligence and 34% machine learning) 33% Health policy principles Alternative care 26% delivery models

Health services 20% research fundamentals

Base: 801 (multiple responses)

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

**Dartmouth Health & Geisel School of Medicine** 

Ways to Give V Stories

About V Events V Search

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

