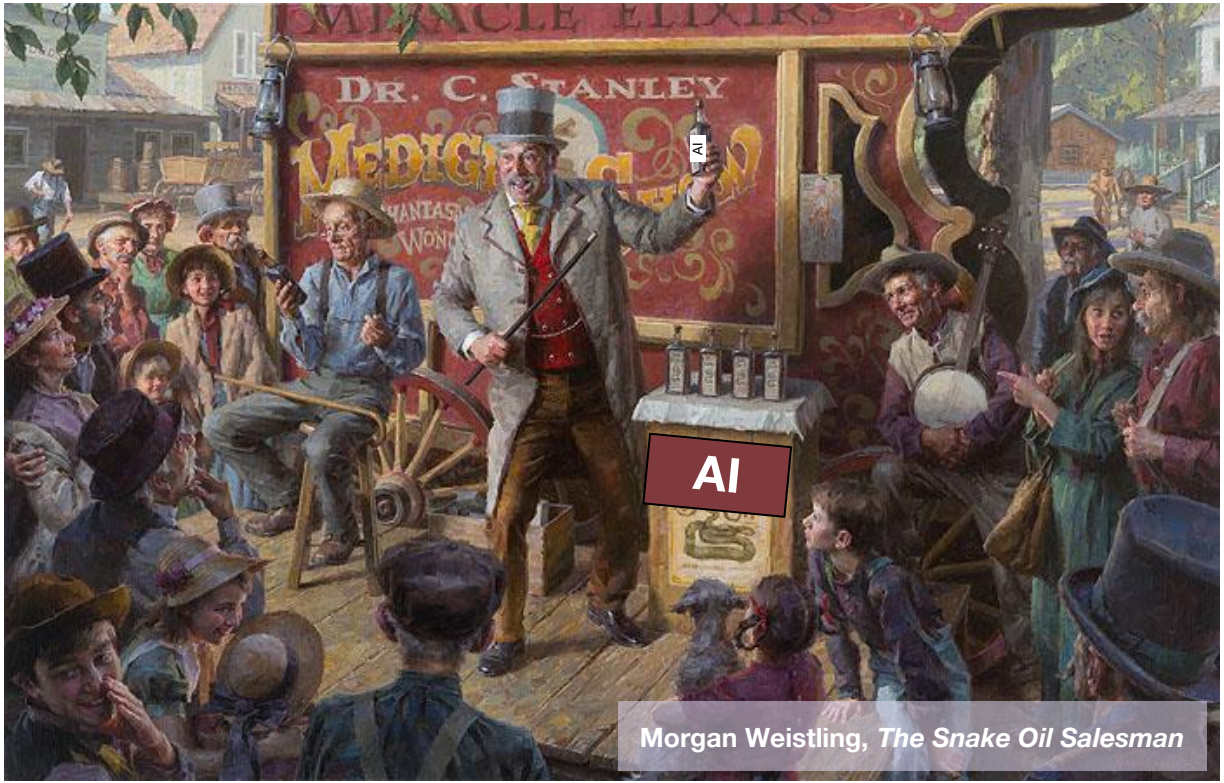
Generative AI for Health: Promises and Pitfalls

Margaret Mitchell

Hugging Face 

17th Annual GoldLab Symposium
15 May, 2026

AI, the magic everything machine!



Morgan Weistling, *The Snake Oil Salesman*

Seeing past the AI hype machine

Key questions to ask:

1. What kind of AI is this?
2. How does it work compared *to* people?
   …when used *with* people?
3. Has utility been found in a real application, or is it theoretical/promising?
4. What effect does it have on people over time?

Categorical Distinction

Predictive AI

Identifying diagnostic markers
Finite set of correct answers
Easier to evaluate

Generative AI

Generating synthetic patient data
Open-ended outputs
Harder to evaluate

What is *non*-Generative AI?

Discriminative Models

Learn the conditional probability $p(Y|X)$: Probability of outcome Y given input data X.

Use

For a specific input x,
output the probability of each possible outcome y.

Examples:

- p(tumor|scan)
- p(atrial fibrillation|patient data)

Generative Models

Learn the probability $p(X)$ or joint probability $p(X, Y)$: the distribution of data (and labels, if relevant).

Use

Sample from the modeled distribution to produce novel, synthetic data resembling the training set.

Examples:

- p(patient data) → synthetic data
- p(medical images, diagnostic labels) → synthetic image of pnemonia

Generative models = Generative AI??

Problem: “Generative AI” ≠ Generative model

Commercial “AI systems” are almost **always** built from **multiple components**.

Diverse Model Types:

- Supervised (Generative / Discriminative)
- Reinforcement Learning
- Clustering
- Density Estimation…

Many Flavors of “AI”

Predictive AI Tasks

tells you something about something that already exists

Characteristics:

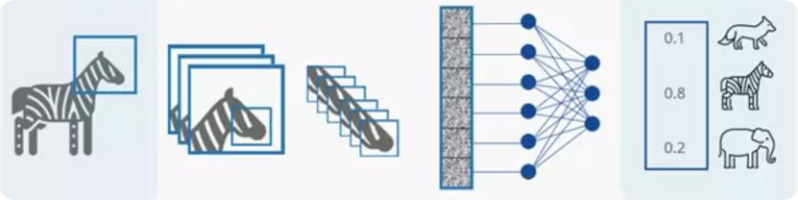
- Labeling and organizing content
- Finite set of correct answers
- Easier to evaluate

Types:

Medical imaging: Detection-focused CNNs

Drug discovery: Graph Neural Networks

Risk scoring: Random Forests / XGBoost



Generative AI Tasks

makes something new

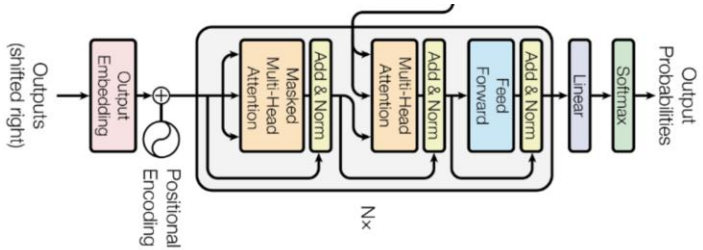
Characteristics:

- Creating poetry or generating video
- Open-ended answers
- Harder to evaluate

Types:

Generating Text: Transformers-based LLMs

Generating Images: U-Net-based Diffusion Models



AI screening for diabetic retinopathy

Task: Detect diabetic retinopathy and diabetic macular edema

- The first FDA authorized AI diagnostic system in any field of medicine
- Helps prevent vision loss in thousands of people with diabetes annually

Clinical trial results

Sensitivity	Specificity	Imageability Rate
87.2%	90.7%	96.1%

Primary "AI": Convolutional Neural Network (CNN) trained for detection

Source: Abràmoff et al., "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices", Nature npj Digital Medicine, 2018

Flagging atrial fibrillation risk

Task: Detect the electrocardiographic signature of atrial fibrillation present during normal sinus rhythm

- Associated with stroke, heart failure, and death
- Frequently asymptomatic, underdetected
- Existing screening methods require prolonged monitoring, are limited by cost & low yield.

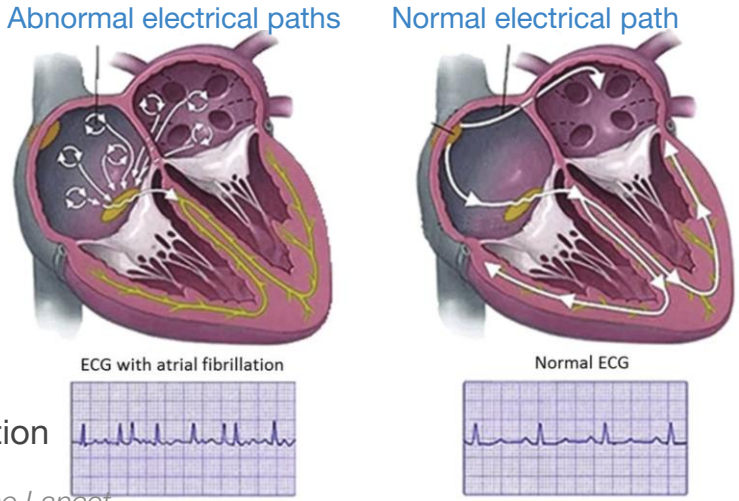
Correct detection up to

10 years

before clinical diagnosis

A single "AI"-enabled ECG performance

AUC	Sensitivity	Specificity	Accuracy
.87	79%	79.5%	79.4%



Primary "AI": Convolutional Neural Network (CNN) trained for detection

Source: Attia et al., "An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm". The Lancet, 2019

Breast cancer detection

Detection Rate

AI group detection rate

6.7 per 1,000

Control group detection rate

5.7 per 1,000

17.6% higher

statistically superior

(95% CI: +5.7%, +30.8%)

PPV of Recall

Recall (AI vs. Control)

noninferior

37.4 vs 38.3 per 1,000

AI group PPV

17.9%

Control group PPV

14.9%

PPV of Biopsy

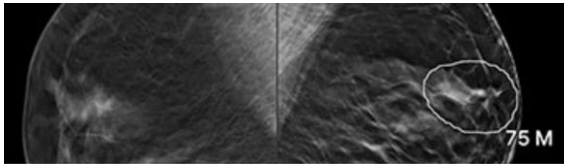
AI group PPV

64.5%

Control group PPV

59.2%

Compared to standard double reading, AI-supported double reading associated with **a higher breast cancer detection rate without negatively affecting the recall rate**.



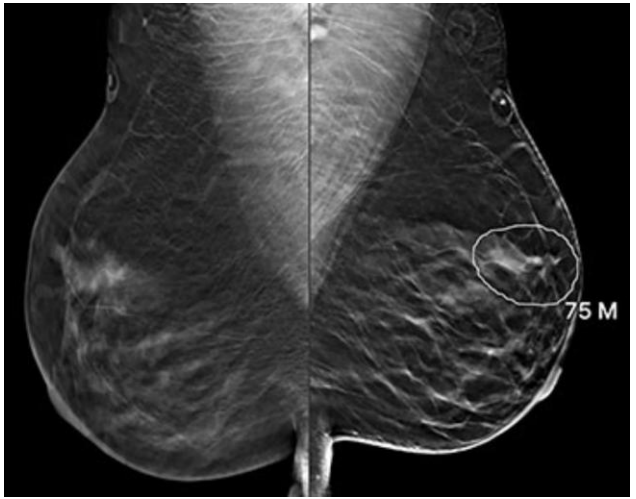
Primary "AI": Convolutional Neural Network (CNN) trained for detection

Source: Eisemann et al., "Nationwide real-world implementation of AI for cancer detection in population-based mammography screening". Nature Medicine, 2025

Catching what humans missed

Task: Detecting and localizing interval cancers at screening Digital Breast Tomosynthesis

True Positives	True Negatives	False Positives Correction
Correctly localized	Correctly categorized	Correctly flagged
84.4%	**85.9%**	**73.3%**
of screening-detected cancers	of true-negative cases as negative	of false-positive cases as negative



Correctly localized **nearly one-third of interval cancers** at retrospective evaluation of screening DBT exams

Primary "AI": Convolutional Neural Network (CNN) trained for detection

Source: Bahl et al., "AI to Reduce the Interval Cancer Rate of Screening Digital Breast Tomosynthesis". Radiology, 2025.

Vaccine preservation & distribution in Nigeria

Problem

Storage: Unreliable electricity, complex refrigeration needs → vaccines spoil

Distribution: Opaque logistic information, unclear need → vaccines late to arrive

Solutions

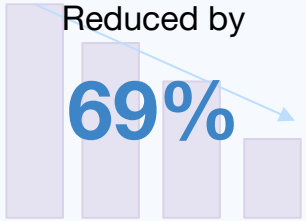
Storage: IoT sensors provide real-time data;
Machine learning detects anomalies, stability, helps optimise temperature

Distribution: Machine learning forecasts demand, optimizes distribution

Vaccine Spoilage

Reduced by

69%



Delivery Lead Time

Improved to

2.5 days

down from 3.9 days

Stock-outs

Reduced to

5%

down from 14%

Primary "AI": Multiple different types in multiple systems

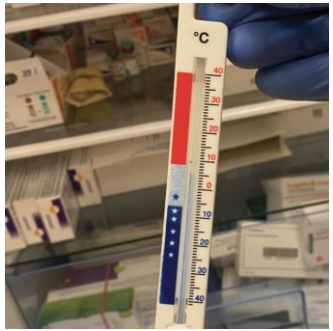
Source: Bature et al., "AI-driven cold chain monitoring systems and preservation and distribution of vaccines in Primary Health Care centres in the Federal Capital Territory, Abuja." ABJAM, 2026

Predictive AI to Assist Clinical Work

Retinopathy

Specificity

90.7%



Vaccine Spoilage

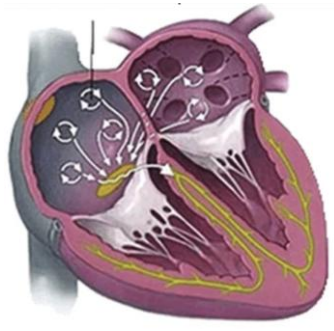
Reduced by

69%

Breast cancer detection

17.6% higher

with AI assistance



Flagging atrial fibrillation risk

Correct detection up to

10 years

before clinical diagnosis

When AI assistance erodes clinician skill

Study: Adenoma Detection Rate (ADR) before vs. after AI adoption for endoscopy

AI-Assisted Boost	Standard ADR Before AI	Standard ADR After AI
ADR increase	Baseline	Dropped to
+8.1%	**28.4%**	**22.4%**
during AI-assisted colonoscopy (meta-analysis of 20 RCTs)	non-AI colonoscopy (n=795)	non-AI colonoscopy (n=648)

After continuous AI exposure, standard non-AI ADR fell by a **6% absolute reduction**, consistent with a clinician **"de-skilling" effect**

Primary "AI": Convolutional Neural Network (CNN) trained for detection

Source: Budzyń et al., "Endoscopist de-skilling after exposure to artificial intelligence in colonoscopy...". The Lancet Gastroenterology & Hepatology, 2025

AI as a scribe in clinical settings

- Clinicians can speak directly to their patients, looking at them rather than a screen.
- The machine captures everything, assembles a structured note; record is essentially complete within seconds.
- **Burnout reduction:** 51.9% to 38.8% after just 30 days of use.
- **Time savings:** ~26% saved on documentation time.

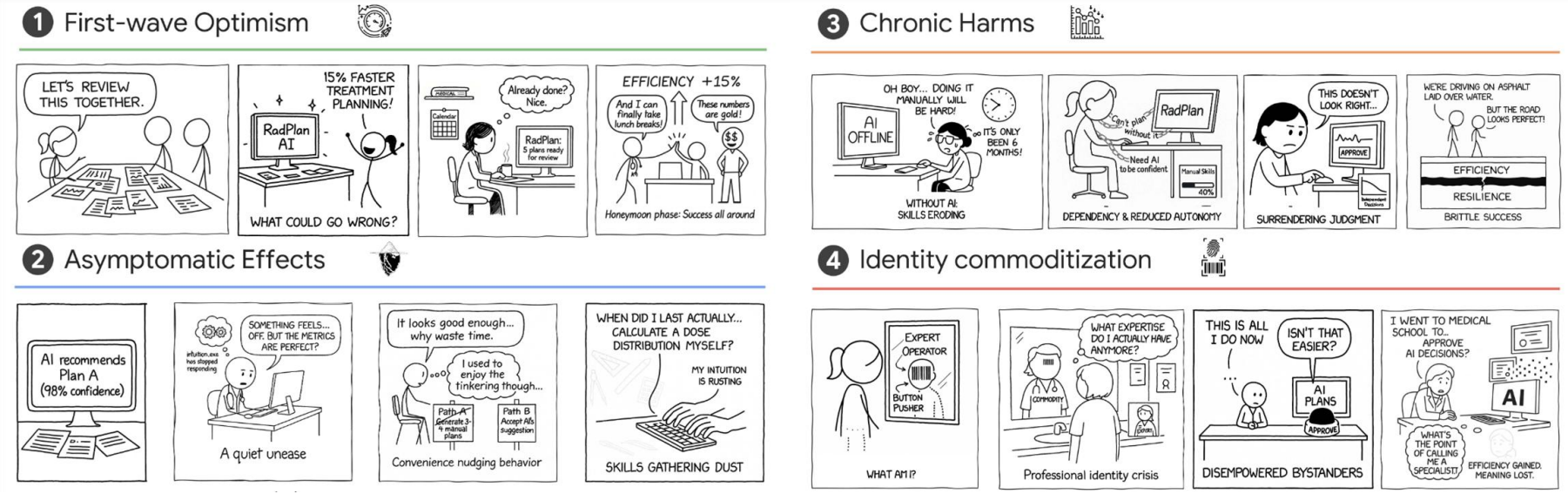
"What I was actually doing was offloading the work of clinical curation — one of the most cognitively demanding and clinically important things a GP does — to the post-consultation review process."

"When I sat down to review a patient I had seen six weeks previously….I did not recognise it."

– Dr. Benn Gooch

Source: Gooch. "I Was an Enthusiastic Early Adopter of AI Scribes. Here's Why I Stopped". Substack blog, 2026

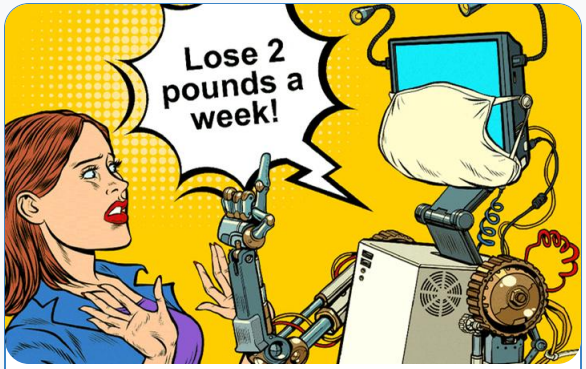
The AI-as-Amplifier Paradox



The strengths that AI offers erodes the very same human expertise it's said to support

Source: Ehsan et al., "From Future of Work to Future of Workers: Addressing Asymptomatic AI Harms to Foster Dignified Human-AIInteraction", CHI 2026

AI Advice in Practice: Real-World Harm



Eating Disorder Helpline

The helpline fired its AI chatbot for providing harmful weight loss advice to vulnerable callers after replaces human staff.

Image Source: NYPost

Dangerous Medical Advice

A case of bromide poisoning linked to AI medical advice that recommended replacing table salt with sodium bromide.

Image Source: Pharma Simplified

Documentation Risks

Clinical AI summaries faulty, containing hallucinated and factually incorrect content, or omitting critical information.

Image Source: Jamie Grill / Getty Images

Generative AI: The Pitfalls

The Allure

Remove tedious, rote work

Generate voluminous content

Often correct

The Reality

Biases:

- Overrepresents some populations, underrepresents others
- Stereotypes

"Hallucinations": Unpredictably adding incorrect content.

"Nuance smoothing": Leaves out critical nuances needed for more appropriate conclusions.

A system correct 99% of time can be more harmful than one consistently wrong.

Leads to:

- Overtrust & Overreliance
- Automation & Anchoring bias

Creates **cognitive debt**, eventually leading to **cognitive surrender** and **deskilling** as we engage less in critical thinking.

AI Assessment Framework

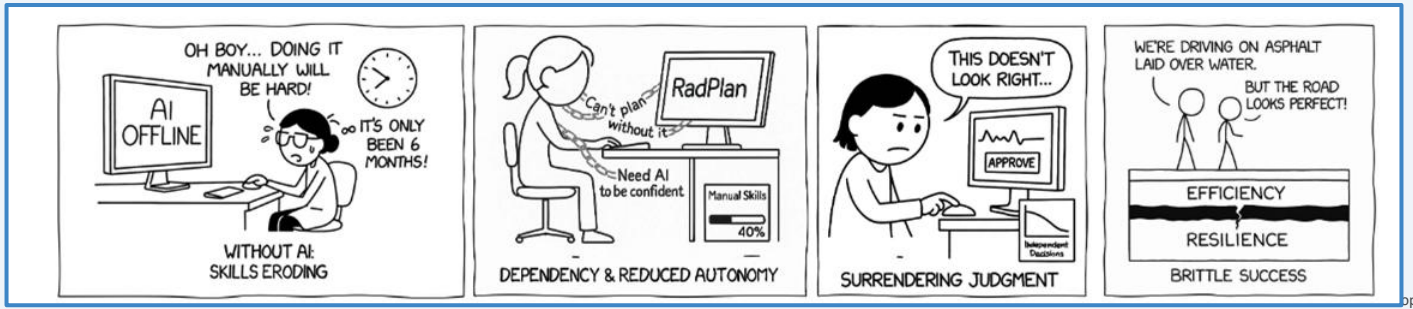
1. When someone says "AI", *what are they talking about?*

chatbot
generative models
discriminative models
generative task
chatbot atop a larger system of many models
diffusion models
large language models
clustering
system of many models
machine learning
neural network
deep learning
predictive task
transformer architecture

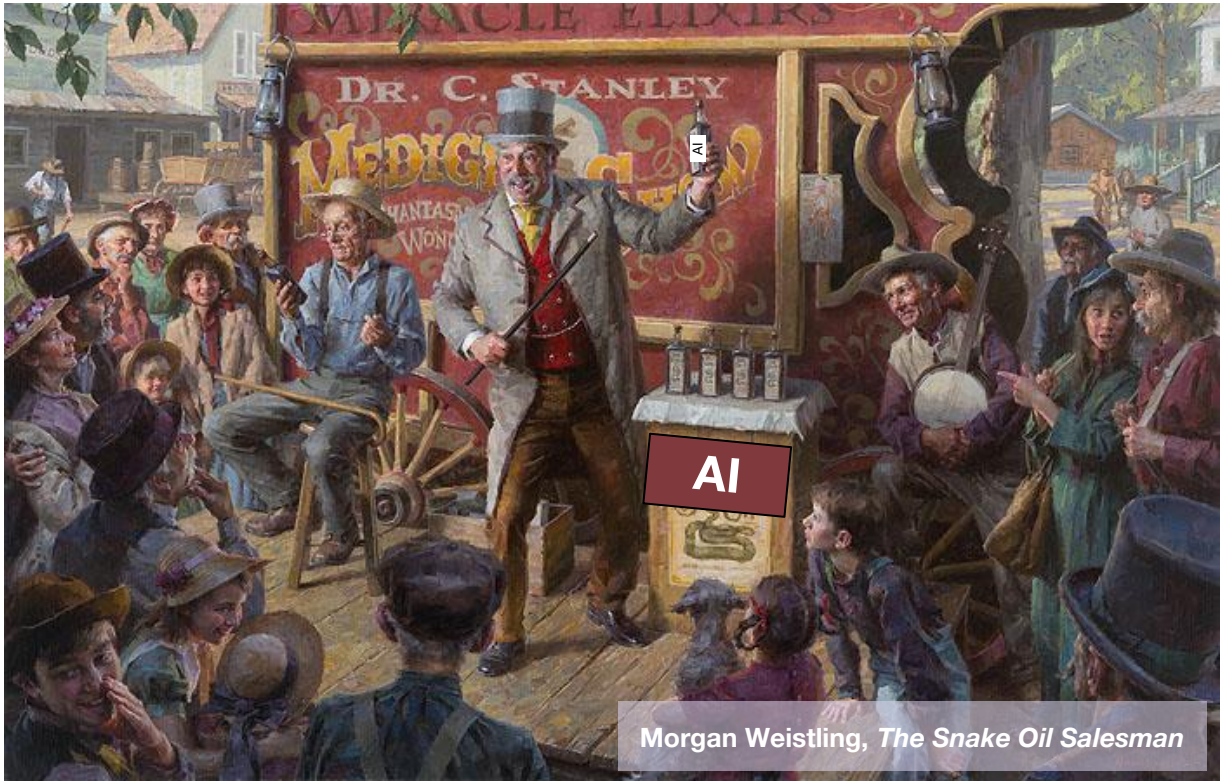
2. If it's "superhuman", *how was that measured?*

false positives?
theoretically?
human-system collaboration?
clinical trial?
accuracy?
compared to humans?
"promising initial results in controlled environment"?
test population diversity?
what an OpenAI salesperson said?
false negatives?

3. What are the longitudinal effects?



AI, the magic everything machine!



Morgan Weistling, *The Snake Oil Salesman*

Generative Artificial Intelligence Use in Healthcare: Opportunities for Clinical Excellence and Administrative Efficiency

Research | Open access | Published: 16 January 2025
Volume 49, article number 10, (2025) Cite this article

Soumitra S. Bhuyan, Vidyoth Sateesh, Naya Mukul, Alay Galvankar, Asos Mahmood, Muhammad Nauman, Akash Rai, Kahuwa Bordoloi, Urmi Basu & Jim Samuel

Abstract

Generative Artificial Intelligence (Gen AI) has transformative potential in healthcare to enhance patient care, personalize treatment options, train healthcare professionals, and advance medical research. This paper examines various clinical and non-clinical applications of Gen AI. In clinical settings, Gen AI supports the creation of customized treatment plans, generation of synthetic data, analysis of medical images, nursing workflow management, risk prediction, pandemic preparedness, and population health management. By automating administrative tasks such as medical documentations, Gen AI has the potential to reduce clinician burnout, freeing more time for direct patient care. Furthermore, application of Gen AI may enhance surgical outcomes by providing real-time feedback and automation of certain tasks in operating rooms. The generation of synthetic data opens new avenues for model training for diseases and simulation, enhancing research capabilities and improving predictive accuracy. In non-clinical contexts, Gen AI

Generative Artificial Intelligence Use in Healthcare: Opportunities for Clinical Excellence and Administrative Efficiency

Research | Open access | Published: 16 January 2025
Volume 49, article number 10, (2025) Cite this article

Soumitra S. Bhuyan, Vidyoth Sateesh, Naya Mukul, Nauman, Akash Rai, Kahuwa Bordoloi, Urmi Basu & Jim

Medical literature describing "GenAI" is instead describing AI as a whole!

Abstract

Generative Artificial Intelligence (Gen AI) has transformati
enhance patient care, personalize treatment options, train healthcare professionals, and advance medical research. This paper examines various clinical and non-clinical applications of Gen AI. In clinical settings, Gen AI supports the creation of customized treatment plans, generation of synthetic data, analysis of medical images, nursing workflow management, risk prediction, pandemic preparedness, and population health management. By automating administrative tasks such as medical documentations, Gen AI has the potential to reduce clinician burnout, freeing more time for direct patient care. Furthermore, application of Gen AI may enhance surgical outcomes by providing real-time feedback and automation of certain tasks in operating rooms. The generation of synthetic data opens new avenues for model training for diseases and simulation, enhancing research capabilities and improving predictive accuracy. In non-clinical contexts, Gen AI

predictive and generative

predictive