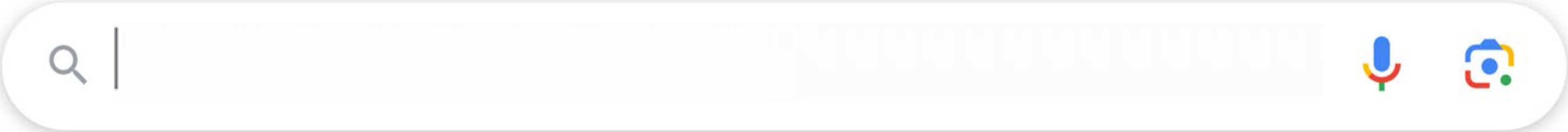


Michael Howell, MD MPH
Chief Clinical Officer
Google

July 2024

Disclosures

Employed by Google (and own equity in Alphabet)

Royalties from

- *Understanding Healthcare Delivery Science* book (McGraw-Hill)
- UpToDate (rapid response systems) (Wolters Kluwer)

Prior funding from

- NIH (R01NR010006-06A1, R01GM123193)
- CDC (Public Health Prevention Fund)
- CMS (1C1CMS33102, 1L1CMS331444)
- RWJF, CIMIT, NPSF

Previously served on

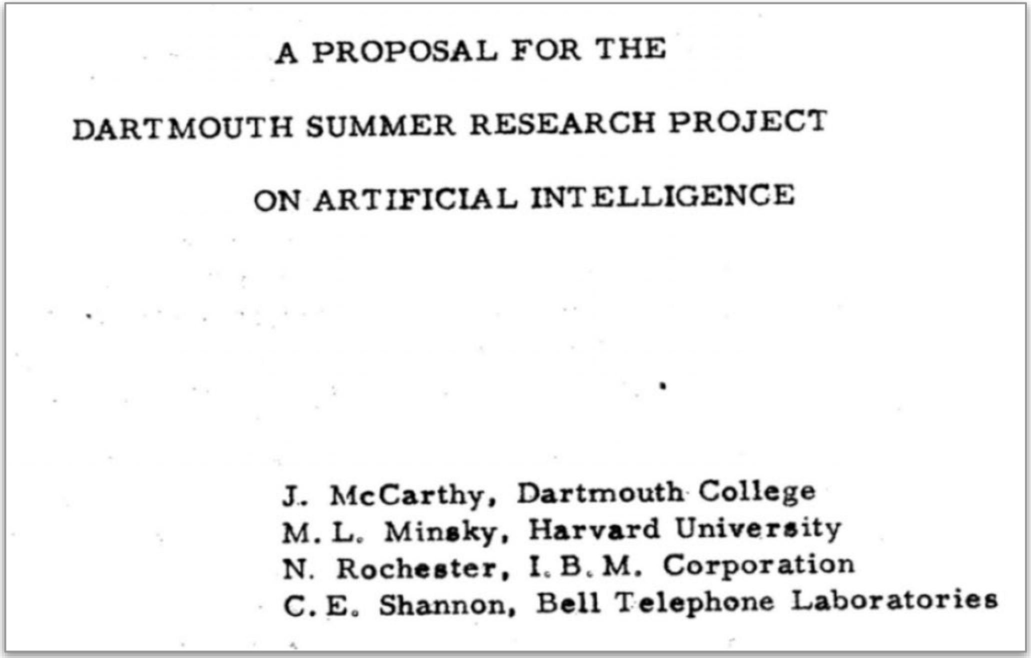
- *National Academy of Medicine*. Various working groups and panels, including on AI/ML and diagnosis.
- *CDC*. Healthcare Infection Control Practices Advisory Committee (HICPAC)
- *CMS*. Co-chair, Hospital Inpatient and Outpatient Process and Structural Measure Technical Expert Panel for CMS

Earliest origins

Earliest origins c. 1950s

"the science and engineering of making intelligent machines"

- John McCarthy, 1956



A PROPOSAL FOR THE

DARTMOUTH SUMMER RESEARCH PROJECT

ON ARTIFICIAL INTELLIGENCE

J. McCarthy, Dartmouth College
M. L. Minsky, Harvard University
N. Rochester, I. B. M. Corporation
C. E. Shannon, Bell Telephone Laboratories

image source: https://home.dartmouth.edu/about/artificial-intelligence-ai-coined-dartmouth

Earliest origins c.1950s

VOL. LIX. NO. 236.] [October, 1950

MIND
A QUARTERLY REVIEW
OF
PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

https://doi.org/10.1093/mind/LIX.236.433

Earliest

origins

c. 1950s

VOL. LIX. NO. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:

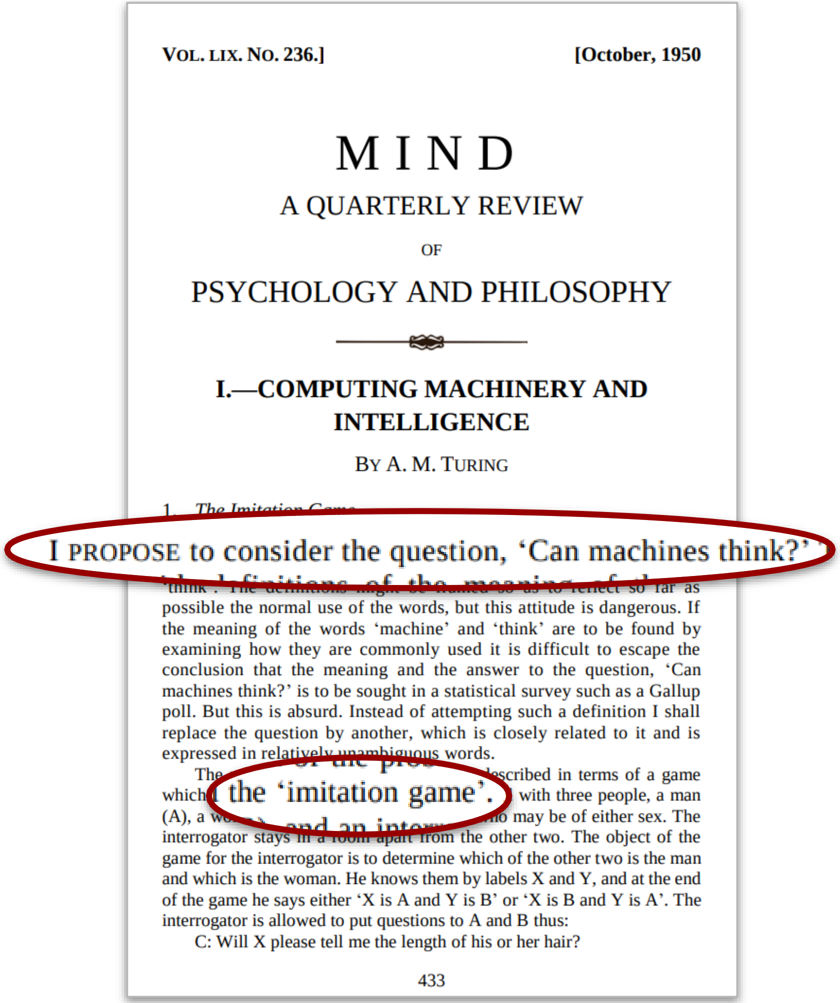
C: Will X please tell me the length of his or her hair?

https://doi.org/10.1093/mind/LIX.236.433

Earliest

origins

c. 1950s



VOL. LIX. NO. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?'

'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

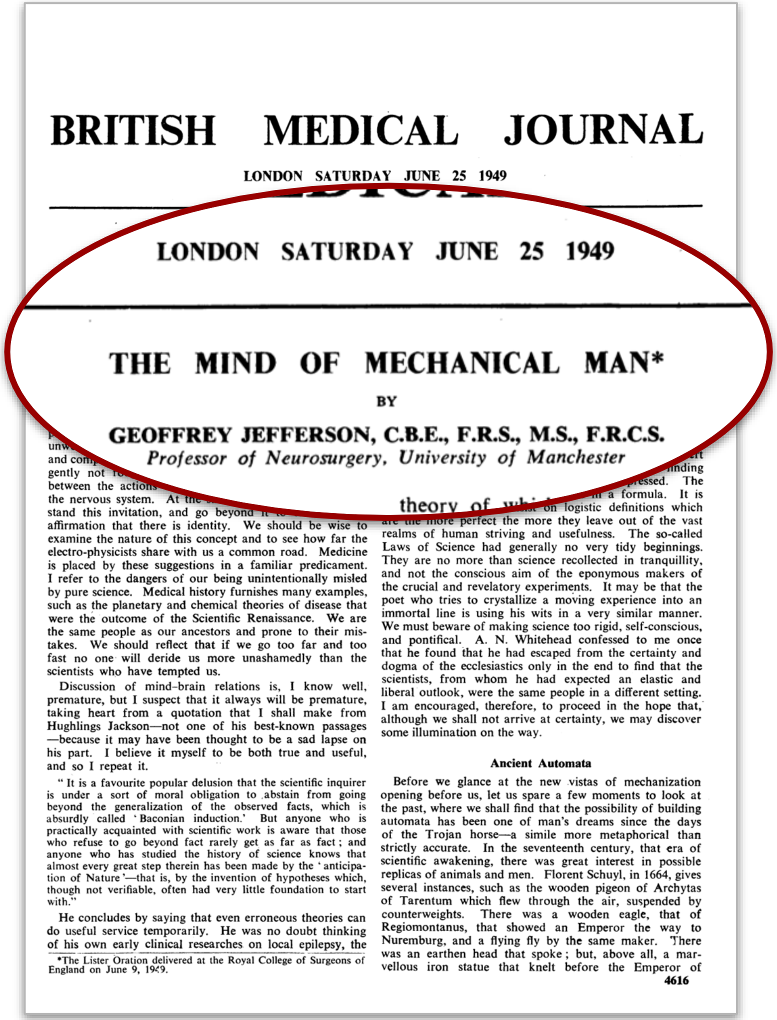
The ... described in terms of a game which the 'imitation game'. with three people, a man (A), a wo... ...ho may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

https://doi.org/10.1093/mind/LIX.236.433

Medicine has been **intertwined** with AI from the **very beginning**.

Turing himself cited a paper from *BMJ*…



BRITISH MEDICAL JOURNAL

LONDON SATURDAY JUNE 25 1949

LONDON SATURDAY JUNE 25 1949

THE MIND OF MECHANICAL MAN*

BY

GEOFFREY JEFFERSON, C.B.E., F.R.S., M.S., F.R.C.S.

Professor of Neurosurgery, University of Manchester

unw
and com
gently not
between the actions
the nervous system. At the
stand this invitation, and go beyond
affirmation that there is identity. We should be wise to examine the nature of this concept and to see how far the electro-physicists share with us a common road. Medicine is placed by these suggestions in a familiar predicament. I refer to the dangers of our being unintentionally misled by pure science. Medical history furnishes many examples, such as the planetary and chemical theories of disease that were the outcome of the Scientific Renaissance. We are the same people as our ancestors and prone to their mistakes. We should reflect that if we go too far and too fast no one will deride us more unashamedly than the scientists who have tempted us.

Discussion of mind–brain relations is, I know well, premature, but I suspect that it always will be premature, taking heart from a quotation that I shall make from Hughlings Jackson—not one of his best-known passages—because it may have been thought to be a sad lapse on his part. I believe it myself to be both true and useful, and so I repeat it.

" It is a favourite popular delusion that the scientific inquirer is under a sort of moral obligation to abstain from going beyond the generalization of the observed facts, which is absurdly called 'Baconian induction.' But anyone who is practically acquainted with scientific work is aware that those who refuse to go beyond fact rarely get as far as fact; and anyone who has studied the history of science knows that almost every great step therein has been made by the 'anticipation of Nature'—that is, by the invention of hypotheses which, though not verifiable, often had very little foundation to start with."

He concludes by saying that even erroneous theories can do useful service temporarily. He was no doubt thinking of his own early clinical researches on local epilepsy, the

inding
pressed. The
a formula. It is
theory of whi
on logistic definitions which
perfect the more they leave out of the vast realms of human striving and usefulness. The so-called Laws of Science had generally no very tidy beginnings. They are no more than science recollected in tranquillity, and not the conscious aim of the eponymous makers of the crucial and revelatory experiments. It may be that the poet who tries to crystallize a moving experience into an immortal line is using his wits in a very similar manner. We must beware of making science too rigid, self-conscious, and pontifical. A. N. Whitehead confessed to me once that he found that he had escaped from the certainty and dogma of the ecclesiastics only in the end to find that the scientists, from whom he had expected an elastic and liberal outlook, were the same people in a different setting. I am encouraged, therefore, to proceed in the hope that, although we shall not arrive at certainty, we may discover some illumination on the way.

Ancient Automata

Before we glance at the new vistas of mechanization opening before us, let us spare a few moments to look at the past, where we shall find that the possibility of building automata has been one of man's dreams since the days of the Trojan horse—a simile more metaphorical than strictly accurate. In the seventeenth century, that era of scientific awakening, there was great interest in possible replicas of animals and men. Florent Schuyl, in 1664, gives several instances, such as the wooden pigeon of Archytas of Tarentum which flew through the air, suspended by counterweights. There was a wooden eagle, that of Regiomontanus, that showed an Emperor the way to Nuremburg, and a flying fly by the same maker. There was an earthen head that spoke; but, above all, a marvellous iron statue that knelt before the Emperor of

*The Lister Oration delivered at the Royal College of Surgeons of England on June 9, 1949.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2050428/pdf/brmedj03683-0003.pdf

Medicine has been **intertwined** with AI from the **very beginning**.

Turing himself cited a paper from *BMJ*…

Conclusion

I conclude, therefore, that although electronic apparatus can probably parallel some of the simpler activities of nerve and spinal cord, for we can already see the parallelism between mechanical feed-backs and Sherringtonian integration, and may yet assist us in understanding better the transmission of the special senses, it still does not take us over the blank wall that confronts us when we come to explore thinking, the ultimate in mind. Nor do I believe that it will do so. I am quite sure that the extreme variety, flexibility, and complexity of nervous mechanisms are greatly underestimated by the physicists, who naturally omit everything unfavourable to a point of view. What I fear is that a great many airy theories will arise in the attempt to persuade us against our better judgment. We have had a hard task to dissuade man from reading qualities of human mind into animals. I see a new and greater danger threatening—that of anthropomorphizing the machine. When we hear it said that wireless valves think, we may despair of language. As well say that the cells in the spinal cord below a transverse lesion "think," a heresy

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2050428/pdf/brmedj03683-0003.pdf

Three epochs of AI in healthcare

A three-part framework for thinking about the history of AI.

Clinical Review & Education

JAMA | Special Communication | AI IN MEDICINE

Three Epochs of Artificial Intelligence in Health Care

Michael D. Howell, MD, MPH; Greg S. Corrado, PhD; Karen B. DeSalvo, MD, MPH, MSc

IMPORTANCE Interest in artificial intelligence (AI) has reached an all-time high, and health care leaders across the ecosystem are faced with questions about where, when, and how to deploy AI and how to understand its risks, problems, and possibilities.

OBSERVATIONS While AI as a concept has existed since the 1950s, all AI is not the same. Capabilities and risks of various kinds of AI differ markedly, and on examination 3 epochs of AI emerge. AI 1.0 includes symbolic AI, which attempts to encode human knowledge into computational rules, as well as probabilistic models. The era of AI 2.0 began with deep learning, in which models learn from examples labeled with ground truth. This era brought about many advances both in people's daily lives and in health care. Deep learning models are task-specific, meaning they do one thing at a time, and they primarily focus on classification and prediction. AI 3.0 is the era of foundation models and generative AI. Models in AI 3.0 have fundamentally new (and potentially transformative) capabilities, as well as new kinds of risks, such as hallucinations. These models can do many different kinds of tasks without being retrained on a new dataset. For example, a simple text instruction will change the model's behavior. Prompts such as "Write this note for a specialist consultant" and "Write this note for the patient's mother" will produce markedly different content.

CONCLUSIONS AND RELEVANCE Foundation models and generative AI represent a major revolution in AI's capabilities, ffering tremendous potential to improve care. Health care leaders are making decisions about AI today. While any heuristic omits details and loses nuance, the framework of AI 1.0, 2.0, and 3.0 may be helpful to decision-makers because each epoch has fundamentally different capabilities and risks.

JAMA. 2024;331(3):242-244. doi:10.1001/jama.2023.25057

Multimedia
Related article page 245
CME at jamacmelookup.com

Author Affiliations: Google, Mountain View, California.

Corresponding Author: Michael D. Howell, MD, MPH, Google LLC, 1600 Amphitheatre Pkwy, Mountain View, CA 94043 (mdhowell@google.com).

Interest in artificial intelligence (AI) has reached an all-time high—whether the metric is scholarly publications, press coverage, or consumer interest. Health care leaders across the ecosystem are faced with questions about where, when, and how to deploy AI and how to understand its risks, problems, and possibilities.

All AI Is Not the Same

Capabilities and risks of various kinds of AI differ markedly. Just as grouping bacterial and viral infections together when making a treatment plan could lead to the wrong clinical response, grouping different kinds of AI together may lead health care decision-makers down the wrong path. A simple, pragmatic framework of 3 epochs of AI may assist decision-makers in understanding the strengths, weaknesses, and challenges of different kinds of AI in this moment of technological change (**Figure**).

AI 1.0: Symbolic AI and Probabilistic Models

Over its first 50-plus years, most AI focused on encoding human knowledge into rules in machines. One can think of this as many, many if-then rules, or decision trees. This *symbolic AI* had some remarkable achievements, such as IBM's Deep Blue, which defeated the chess world champion in 1997.[1] In health care, tools such as INTERNIST-I aimed to represent expert knowledge about diseases to help with challenging cases.[2] Today, many electronically implemented clinical pathways encode expert knowledge in decision trees, a type of symbolic AI.

Symbolic AI also had key limitations, notably a constant risk of human logic errors in its construction and bias encoded in its rules, because its knowledge base depended solely on those creating it. But perhaps the most important issue was that, empirically, symbolic AI had fundamental capability limitations and appeared brittle when confronted with real-world situations. In response, research began focusing more on probabilistic modeling, such as traditional regression and then bayesian networks, which allowed both expert knowledge and empirical data to contribute to reasoning systems.[3] These models handled real-world situations more elegantly and found some use in health care,[4] but in practice were difficult to scale and had limited ability to manage images, free text, and other complex clinical data.

AI 2.0: The Era of Deep Learning

Work on even more data-driven methods, broadly called *machine learning*, was rooted in the idea that key to intelligence was learning from errors. Discoveries such as backpropagation of errors in the

Howell, Corrado, DeSalvo. Three epochs of artificial intelligence in healthcare. *JAMA* 2024.

The most important point →

The technical details here matter to decision making.

All AI Is Not the Same

Capabilities and risks of various kinds of AI differ markedly. Just as grouping bacterial and viral infections together when making a treatment plan could lead to the wrong clinical response, grouping different kinds of AI together may lead health care decision-makers down the wrong path. A simple, pragmatic framework of 3 epochs of AI may assist decision-makers in understanding the strengths, weaknesses, and challenges of different kinds of AI in this moment of technological change (**Figure**).

Figure. Artificial Intelligence (AI) 1.0, 2.0, and 3.0

Approximate beginning year	*1950s*	*2011*	*2018-2022*
	AI 1.0 Symbolic AI and probabilistic models	**AI 2.0** Deep learning	**AI 3.0** Foundation models
Core functionality and key features	Follows directly encoded rules (if-then rules or decision trees)	Predicts and/or classifies information Task-specific (1 task at a time); requires new data and retraining to perform new tasks	Generates new content (text, sound, images) Performs different types of tasks without new data or retraining; prompt creates new model behaviors
Training method	Rules based on expert knowledge are hand-encoded in traditional programming	Learning patterns based on examples labeled as ground truth	Self-supervised learning from large datasets to predict the next word or sentence in a sequence
Performance capabilities	Follows decision path encoded in its rules. *Eg, ask a series of questions to determine whether a picture is a cat or a dog.*	Classifies information based on training: *"Is this a cat or a dog?"* *"How many dogs will be in the park at noon?"*	Interprets and responds to complex questions: *"Explain the difference between a cat and a dog."*
Examples of performance	IBM's Deep Blue beat the world champion in chess **Health care:** Rule-based clinical decision support tools	Photo searching without manual tagging, voice recognition, language translation **Health care:** diabetic retinopathy detection, breast cancer and lung cancer screening, skin condition classification, predictions based on electronic health records	Writing assistants in word processors, software coding assistants, chatbots **Health care:** Med-PaLM and Med-PaLM-2, medically tuned large language models, PubMedGPT, BioGPT
Examples of challenges and risks	Human logic errors and bias in encoded rules lead to limited capability with real-world situations	Out-of-distribution problems (real-time data differs from training data) Catastrophic forgetting (not remembering early parts of a long sequence of text) Bias related to underlying training data	Hallucinations (plausible but incorrect responses based solely on predictions) Grounding and attribution Bias related to underlying training data and semantics of language in datasets

Howell, Corrado, DeSalvo. *JAMA* 2024.

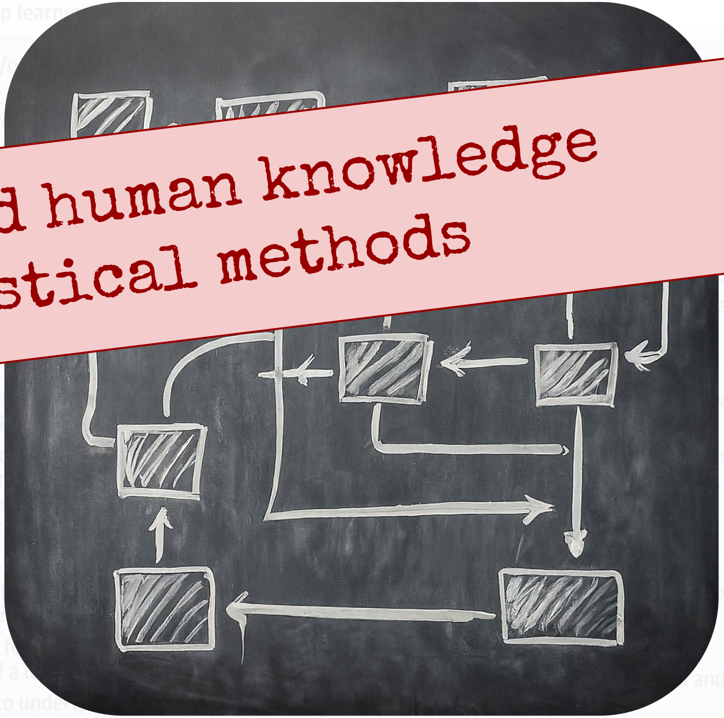
Each epoch's models learn differently.

Howell, Corrado, DeSalvo. *JAMA* 2024.

Image source: Source: Karen Zak, Twitter Post, March 9, 2016, 4:40 p.m., https://twitter.com/teenybiscuit/status/707727863571582978 via Crumpler, W. and Lewis, J.A., 2021. How Does Facial Recognition Work? Center for Strategic and International Studies

Figure. Artificial Intelligence (AI) 1.0, 2.0, and 3.0

Approximate beginning year	*1950s*
	AI 1.0 Symbolic AI and probabilistic models
Core functionality and key features	Follows directly encoded rules (if-then rules or decision trees)
Training method	Rules based on expert knowled
Performance capabilities	
Examples of performance	world champion
	Health care: Rule-based clinical decision support tools
Examples of challenges and risks	Human logic errors and bias in encoded rules lead to limited capability with real-world situations



1.0: Encoded human knowledge
+ statistical methods

Howell, Corrado, DeSalvo. *JAMA* 2024.

Even simple encoded human knowledge can be convincing and compelling.

Anybody ever played this?

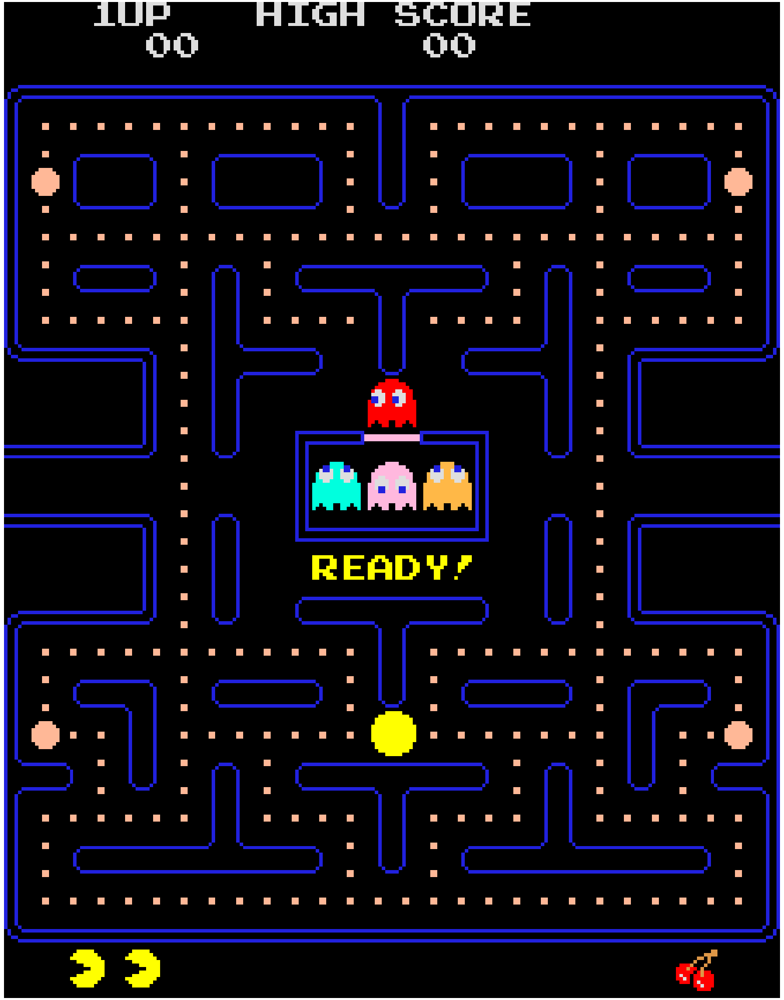
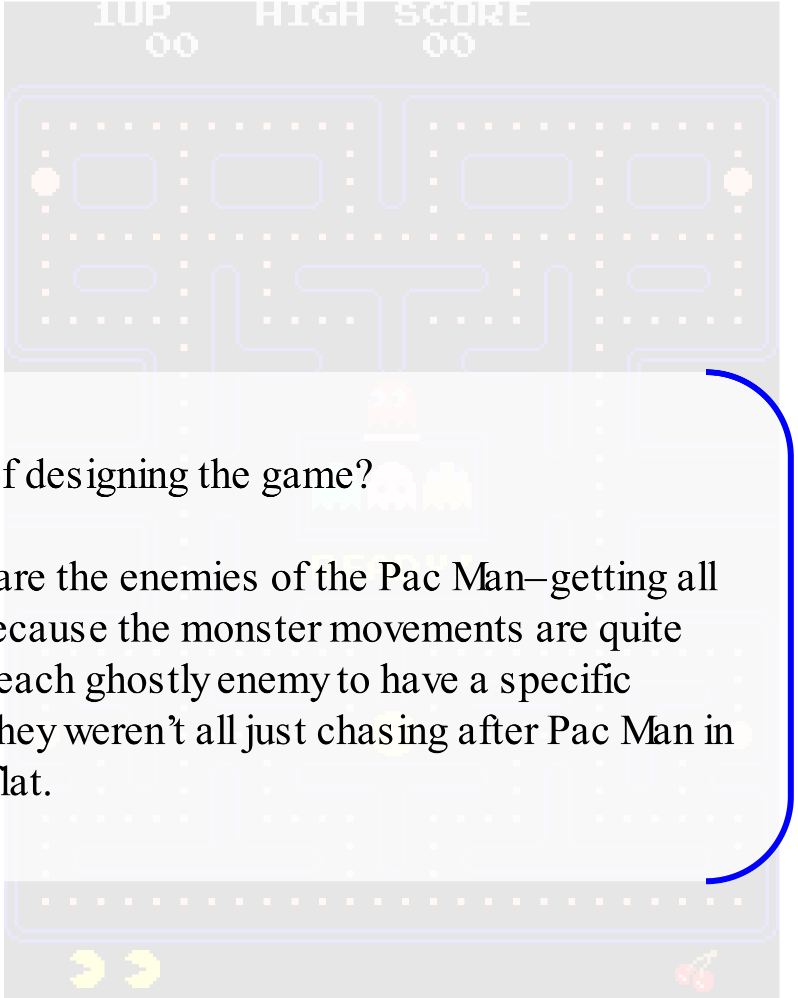


Image source: https://en.wikipedia.org/w/index.php?curid=533608

The game's designer, Toru Itawani, was interviewed in 1986.



INTERVIEWER: What was the most difficult part of designing the game?

IWATANI: The algorithm for the four ghosts who are the enemies of the Pac Man–getting all the movements lined up correctly. It was tricky because the monster movements are quite complex. This is the heart of the game. I wanted each ghostly enemy to have a specific character and its own particular movements, so they weren't all just chasing after Pac Man in single file, which would have been tiresome and flat.

Quotation: https://programmersatwork.wordpress.com/toru-iwatani-1986-pacman-designer/.
Image source: https://en.wikipedia.org/w/index.php?curid=533608

The earliest empirical paper I'm aware of specifically addressing AI and diagnosis is from 61 year ago this month.

> IEEE Trans Biomed Eng. 1963 Jul;10:106-14. doi: 10.1109/tbmel.1963.4322808.

COMPUTER PATTERN RECOGNITION TECHNIQUES: SOME RESULTS WITH REAL ELECTROCARDIOGRAPHIC DATA

M OKAJIMA, L STARK, G WHIPPLE, S YASUI

PMID: 14063196 DOI: 10.1109/tbmel.1963.4322808 Paperpile

Computer Pattern Recognition Techniques: Some Results with Real Electrocardiographic Data*

M. OKAJIMA†, L. STARK‡, SENIOR MEMBER, IEEE, G. WHIPPLE‖, AND S. YASUI§

Summary—Automatic interpretation of electrocardiograms is a particular example of the application of digital computers to medical diagnosis; this paper describes our experience with a new approach involving pattern recognition techniques. The program employs a multiple adaptive matched filter system with a variety of normalization, weighting, comparison, decision, modification, and adapting operations. The flexibility of the method has permitted study of effects of experimental variations of these operations on the pattern classification process to simulate human interpretation of electrocardiograms more closely. These programs have been successfully applied to actual electrocardiograms from cardiac patients.

These researches in application of computer pattern recognition techniques to the automatic interpretation of electrocardiograms have been undertaken because they join together three fields of great interest. First, an example of artificial intelligence or a self-organized system is represented by the adaptive filter memory, together with the related decision operations. Second, we consider our program to be a model of complex sensory discrimination and use our intuition of human psychology as a guide when selecting one of several possible program mechanisms to overcome temporary obstacles. Third, the automation of medical diagnosis is a rapidly developing and promising field contributing to medical progress. This paper pays particular attention to the third of these objectives.

The present state of computer analysis of electrocardiograms is mainly one of orthogonalization of the spatial vector,[1] point recognition[2] to separate the various component waves, parameterization,[3] in one case via Fourier techniques,[4] and then statistical matrix analysis. The matched filter technique has been previously applied to radar signals[5,6] and appeared to us to proceed along lines similar to human pattern recognition. It utilizes all the signal information in the time domain, preserves more information in the early stages of the recognition processes, and is computationally straightforward because of its operation directly in the time domain, without requiring Fourier analysis. The adaptive matched filter extends these advantages,[7–9] particularly in the form of a system of multiple adaptive matched filters.[10–12]

In our earlier work we resorted to the artifice of using artificial time-normalized and synchronized ideal textbook electrocardiographic (EKG) patterns in order to circumvent certain initial problems. In the present paper we describe our experience with the application of multiple adaptive matched filter schemes to real electrocardiographic programs on a digital computer[13] to recognize and classify data, made possible by automatic time-normalization and synchronization.

METHOD

Data

WE HAVE OBTAINED these electrocardiograms from patients who were being studied clinically in the Electrocardiographic Laboratory of the Department of Medicine at the Boston University Medical Center. The data were amplified SVEC III vectors obtained with due care to eliminate as much recording noise as possible and were stored on three FM modulated channels of a tape recorder. The data were then digitized on the Neurology Section digital computer (GE 225) at the Massachusetts

https://ieeexplore.ieee.org/abstract/document/4322808

The earliest empirical paper I'm aware of specifically addressing AI and diagnosis is from 61 year ago this month.

106 IEEE TRANSACTIONS ON BIO-MEDICAL ELECTRONICS July

Computer Pattern Recognition Techniques: Some Results with Real Electrocardiographic Data*

M. OKAJIMA†, L. STARK†, SENIOR MEMBER, IEEE, G. WHIPPLE‖, AND S. YASUI§

Summary—Automatic interpretation of electrocardiograms is a particular example of the application of digital computers to medical diagnosis;

with a variety of normalization, weighting, comparison, decision, modification, and adapting operations. The flexibility of the method has permitted study of effects of experimental variations of these operations on the pattern classification process to simulate human interpretation of electrocardiograms more closely. These programs have been successfully applied to actual electrocardiograms from cardiac patients.

These researches in application of computer pattern recogni-

First, an example of artificial intelligence or a self-organized system is represented by the adaptive filter memory, together with the related decision operations.

discrimination and use our intuition of human psychology as a guide when selecting one of several possible program mechanisms to overcome temporary obstacles. Third, the automation of medical diagnosis is a rapidly developing and promising field contributing to medical progress. This paper pays particular attention to the third of these objectives.

The present state of computer analysis of electrocardiograms is mainly one of orthogonalization of the spatial vector,[1] point recognition[2] to separate the various component waves, parameterization,[3] in one case via Fourier techniques,[4] and then statistical matrix analysis. The matched filter technique has been previously applied to radar signals[5,6] and appeared to us to proceed along lines similar to human pattern recognition. It

on in the time domain, preserves ages of the recognition processes, forward because of its operation thout requiring Fourier analysis. xtends these advantages,[7-9] particularly in the form of a system of multiple adaptive matched filters.[10-12]

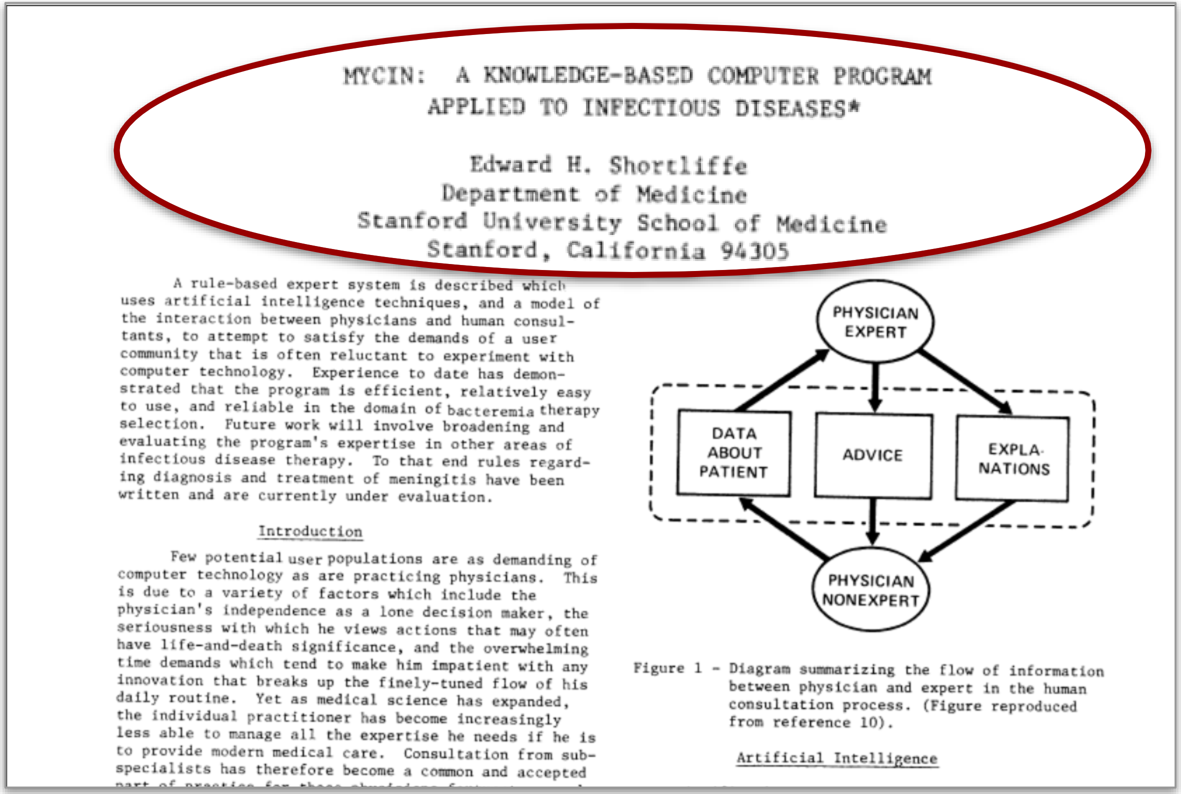
In our earlier work we resorted to the artifice of using artificial time-normalized and synchronized ideal textbook electrocardiographic (EKG) patterns in order to circumvent certain initial problems. In the present paper we describe our experience with the application of multiple adaptive matched filter schemes to real electrocardiographic programs on a digital computer[13] assify data, made possible by automatic time-synchronization.

Method

WE HAVE OBTAINED these electrocardiograms from patients who were being studied clinically in the Electrocardiographic Laboratory of the Department of Medicine at the Boston University Medical Center. The data were amplified SVEC III vectors obtained with due care to eliminate as much recording noise as possible and were stored on three FM modulated channels of a tape recorder. The data were then digitized on the Neurology Section digital computer (GE 225) at the Massachusetts

https://ieeexplore.ieee.org/abstract/document/4322808

A better - known innovation was Shortliffe ’s MYCIN from 1977



MYCIN: A KNOWLEDGE-BASED COMPUTER PROGRAM APPLIED TO INFECTIOUS DISEASES*

Edward H. Shortliffe
Department of Medicine
Stanford University School of Medicine
Stanford, California 94305

A rule-based expert system is described which uses artificial intelligence techniques, and a model of the interaction between physicians and human consultants, to attempt to satisfy the demands of a user community that is often reluctant to experiment with computer technology. Experience to date has demonstrated that the program is efficient, relatively easy to use, and reliable in the domain of bacteremia therapy selection. Future work will involve broadening and evaluating the program's expertise in other areas of infectious disease therapy. To that end rules regarding diagnosis and treatment of meningitis have been written and are currently under evaluation.

Introduction

Few potential user populations are as demanding of computer technology as are practicing physicians. This is due to a variety of factors which include the physician's independence as a lone decision maker, the seriousness with which he views actions that may often have life-and-death significance, and the overwhelming time demands which tend to make him impatient with any innovation that breaks up the finely-tuned flow of his daily routine. Yet as medical science has expanded, the individual practitioner has become increasingly less able to manage all the expertise he needs if he is to provide modern medical care. Consultation from subspecialists has therefore become a common and accepted

Figure 1 - Diagram summarizing the flow of information between physician and expert in the human consultation process. (Figure reproduced from reference 10).

Artificial Intelligence

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2464549/?page=1

… and INTERNIST I from 1982

The NEW ENGLAND JOURNAL of MEDICINE

SPECIALTIES TOPICS MULTIMEDIA CURRENT ISSUE LEARNING/CME AUTHOR CENTER PUBLICATIONS

SPECIAL ARTICLE | ARCHIVE

Internist-I, an Experimental Computer-Based Diagnostic Consultant for General Internal Medicine

Authors: Randolph A. Miller, M.D., Harry E. Pople, Jr., Ph.D., and Jack D. Myers, M.D. Author Info & Affiliations

Published August 19, 1982 | N Engl J Med 1982;307:468-476 | DOI: 10.1056/NEJM198208193070803
VOL. 307 NO. 8

Abstract

INTERNIST-I is an experimental computer program capable of making multiple and complex diagnoses in internal medicine. It differs from most other programs for computer-assisted diagnosis in the generality of its approach and the size and diversity of its knowledge base. To document the strengths and weaknesses of the program we performed a systematic evaluation of the capabilities of INTERNIST-I. Its performance on a series of 19 clinicopathological exercises (Case Records of the Massachusetts General Hospital) published in the *Journal* appeared qualitatively similar to that of the hospital clinicians but inferior to that of the case discussants. The evaluation demonstrated that the present form of the program is not sufficiently reliable for clinical applications. Specific deficiencies that must be overcome include the program's inability to reason anatomically or temporally, its inability to construct differential diagnoses spanning multiple problem areas, its occasional attribution of findings to improper causes, and its inability to explain its "thinking." (N Engl J Med. 1982; 307:468–76.)

https://www.nejm.org/doi/10.1056/NEJM198208193070803

… and INTERNIST I from 1982

“Its performance on a series of 19 clinicopathological exercises (Case Records of the Massachusetts General Hospital) published in the Journal appeared qualitatively similar to that of the hospital clinicians but inferior to that of the case discussants. The evaluation demonstrated that the present form of the program is not sufficiently reliable for clinical applications. Specific deficiencies that must be overcome include the program's inability to reason anatomically or temporally, its inability to construct differential diagnoses spanning multiple problem areas, its occasional attribution of findings to improper causes, and its inability to explain its ‘thinking.’”

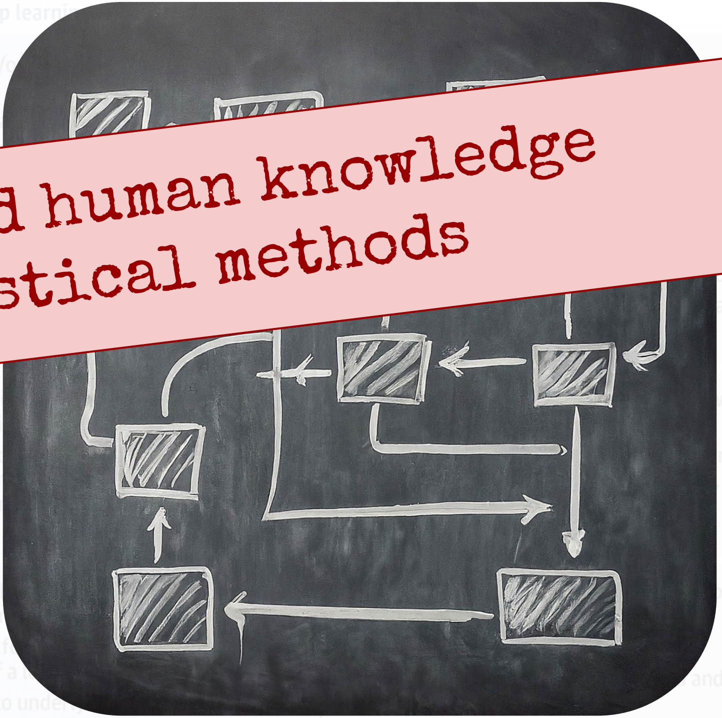
https://www.nejm.org/doi/10.1056/NEJM198208193070803

Numerous others approaches in this epoch

- DxPlain
- Caduceus
- CASNET
- Watson – bridging into the second epoch
  - "For the *Jeopardy* Challenge, we use more than 100 different techniques for analyzing natural language, identifying sources, finding and generating hypotheses, finding and scoring evidence, and merging and ranking hypotheses." [Ferrucci D, Brown E, Chu-Carroll J, Fan J, Gondek D, Kalyanpur AA, Lally A, Murdock JW, Nyberg E, Prager J, Schlaefer N. Building Watson: An overview of the DeepQA project. AI Magazine. 2010 Jul 28;31(3):59-79.]
- And others

Figure. Artificial Intelligence (AI) 1.0, 2.0, and 3.0

Approximate beginning year	1950s
	AI 1.0 Symbolic AI and probabilistic models
Core functionality and key features	Follows directly encoded rules (if-then rules or decision trees)
Training method	Rules based on expert knowled… hand…
Performance capabilities	
Examples of performance	…world champion …chess Health care: Rule-based clinical decision support tools
Examples of challenges and risks	Human logic errors and bias in encoded rules lead to limited capability with real-world situations



1.0: Encoded human knowledge + statistical methods

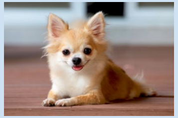
Howell, Corrado, DeSalvo. *JAMA* 2024.

Figure. Artificial Intelligence (AI) 1.0, 2.0, and 3.0

Approximate beginning year	*1950s*	*2011*	*2018-2022*
	AI 1.0 Symbolic AI and probabilistic models	**AI 2.0** Deep learning	**AI 3.0** Foundation models
Core functionality and key features	Follows directly encoded rules (if-then rules or decision trees)	Predicts and/or classifies information Task-specific (1 task at a time); requires new data and retraining to perform new tas	Generates new content (text, sound, images) Perform ... pes of tasks without new data ... creates new model behaviors
Training method	Rules based on expert knowl hand-enco		ng from large datasets rd or sentence in a sequence
Performance capabilities	Follows dec *Eg, ask a se* *whether a pi*	*"How many dogs will be in the park at noon?"*	Interprets and responds to complex questions: *"Explain the difference between a cat and a dog."*
Examples of performance	IBM's Deep Blue beat the world champion in chess Health care: Rule-based clinical decision support tools	Photo searching without manual tagging, voice recognition, language translation Health care: diabetic retinopathy detection, breast cancer and lung cancer screening, skin condition classification, predictions based on electronic health records	Writing assistants in word processors, software coding assistants, chatbots Health care: Med-PaLM and Med-PaLM-2, medically tuned large language models, PubMedGPT, BioGPT
Examples of challenges and risks	Human logic errors and bias in encoded rules lead to limited capability with real-world situations	Out-of-distribution problems (real-time data differs from training data) Catastrophic forgetting (not remembering early parts of a long sequence of text) Bias related to underlying training data	Hallucinations (plausible but incorrect responses based solely on predictions) Grounding and attribution Bias related to underlying training data and semantics of language in datasets

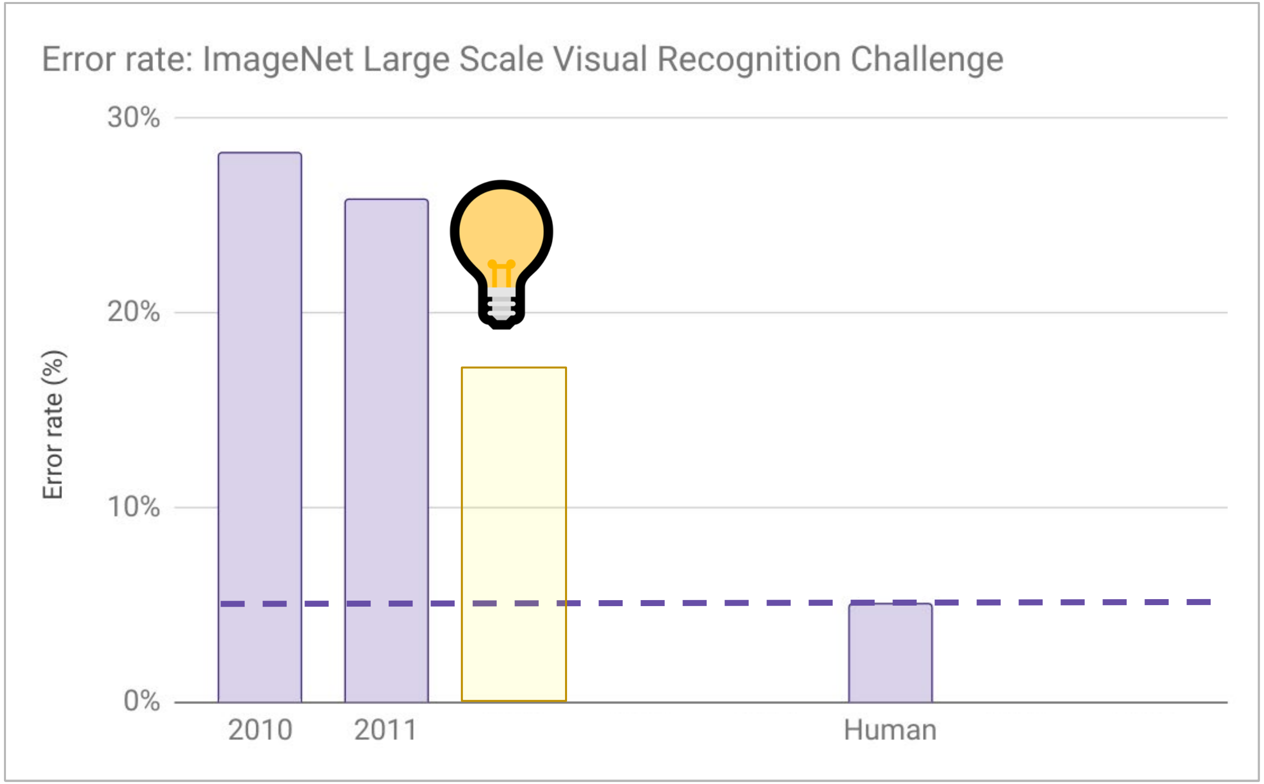
2.0: Task-specific:
one thing at a time

Howell, Corrado, DeSalvo. *JAMA* 2024.

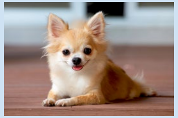


vs

AI 2.0 : Deep Learning Era



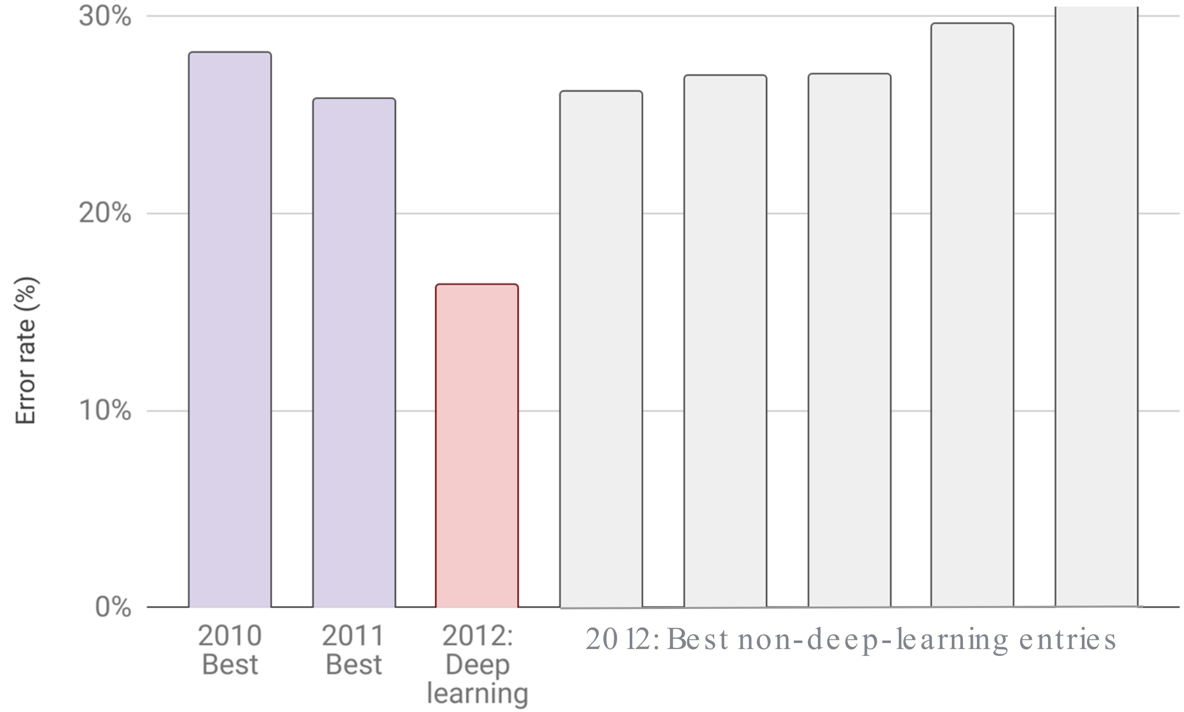
Data from: Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC. Imagenet large scale visual recognition challenge. International journal of computer vision. 2015 Dec;115:211-52 and https://www.image-net.org/challenges/LSVRC/



vs

AI 2.0 : Deep Learning Era

Deeper Dive: What happened in 2012?



Data from: Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC. Imagenet large scale visual recognition challenge. International journal of computer vision. 2015 Dec;115:211-52 and https://www.image-net.org/challenges/LSVRC/

AI **2.0** :
Deep Learning Era

This Issue | Views **435,334** | Citations **3,784** | Altmetric **1095**

Download PDF | More ▽ | Cite This | Permissions

Original Investigation | Innovations in Health Care Delivery

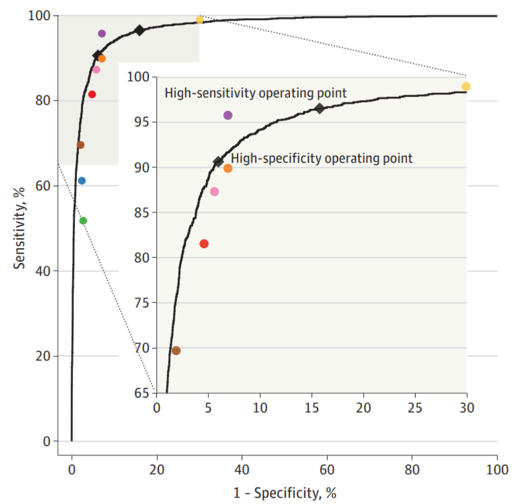
FREE

December 13, 2016

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD[1]; Lily Peng, MD, PhD[1]; Marc Coram, PhD[1]; et al

Figure 3. Validation Set Performance for All-Cause Referable Diabetic Retinopathy in the EyePACS-1 Data Set (9946 Images)



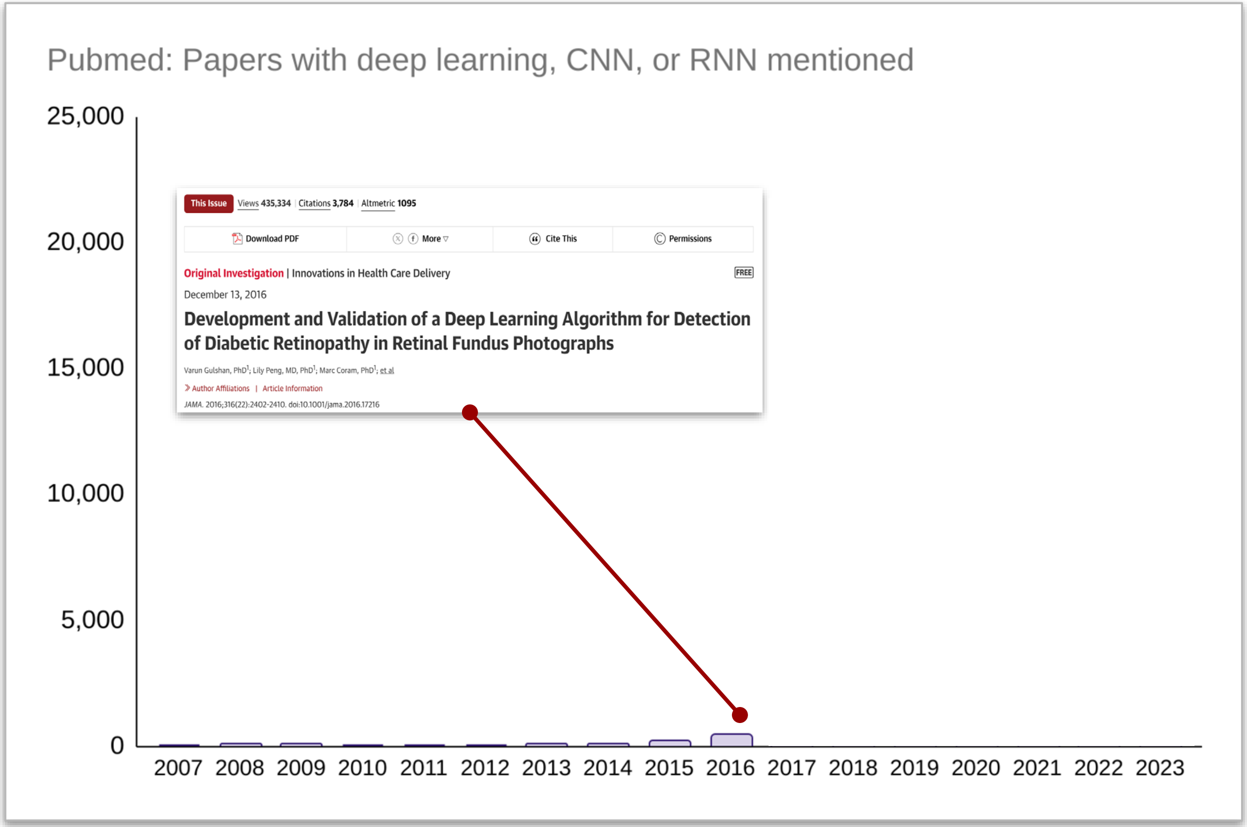
Search All
Enter Search Term

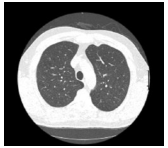
JAMA

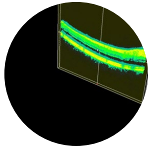
Editor's Picks 2010 - 2020

The articles most important to clinical medicine and public health published in the 2010s, assessed by views and citations and curated by JAMA editors.

AI 2.0: Deep Learning Era







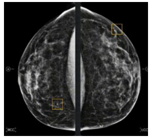


Figure. Artificial Intelligence (AI) 1.0, 2.0, and 3.0

Approximate beginning year

1950s — AI 1.0 Symbolic AI and probabilistic models

2011 — **AI 2.0** Deep learning

2018-2022 — AI 3.0 Foundation models

Core functionality and key features

AI 2.0: Predicts and/or classifies information. Task-specific (1 task at a time); requires new data and retraining to perform new tas...

AI 3.0: Generates new content (text, sound, images). Perform... ...pes of tasks without new data ... creates new model behaviors

Training method

Performance capabilities

AI 2.0: ..."How many dogs will be in the park at noon?"

Examples of performance

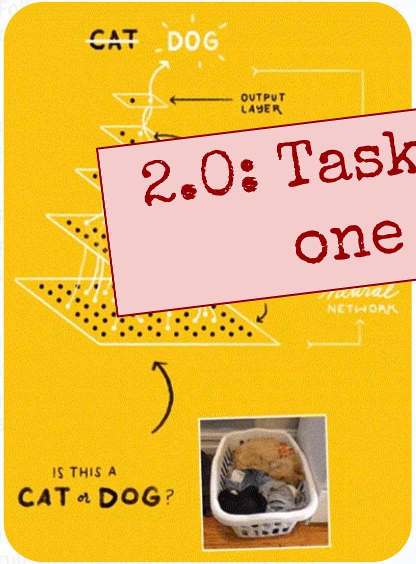
AI 2.0: Photo searching without manual tagging, voice recognition, language translation. **Health care:** diabetic retinopathy detection, breast cancer and lung cancer screening, skin condition classification, predictions based on electronic health records

Examples of challenges and risks

AI 1.0: ...rules lead to limited capability with real-world situations

AI 2.0: Out-of-distribution problems (real-time data differs from training data). Catastrophic forgetting (not remembering early parts of a long sequence of text). Bias related to underlying training data

AI 3.0: Hallucinations (plausible but incorrect responses based solely on predictions). Grounding and attribution. Bias related to underlying training data and semantics of language in datasets



2.0: Task-specific:
one thing at a time

Howell, Corrado, DeSalvo. *JAMA* 2024.

Figure. Artificial Intelligence (AI) 1.0, 2.0, and 3.0

Approximate beginning year

1950s

AI 1.0

AI 2.0 Deep learning

Predicts and/or classifies information

Task-specific (1 task at a time); requires new data and retraining to perform new tasks

2018-2022

AI 3.0 Foundation models

Core functionality and key features

Generates new content (text, sound, images)

Performs diff... without new data ... model behaviors

Training method

...e datasets ...ce in a sequence

Performance capabilities

...nterprets and responds to complex questions: *"Explain the difference between a cat and a dog."*

Examples of performance

Writing assistants in word processors, software coding assistants, chatbots

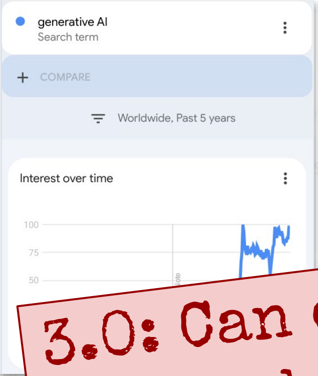
Health care: Med-PaLM and Med-PaLM-2, medically tuned large language models, PubMedGPT, BioGPT

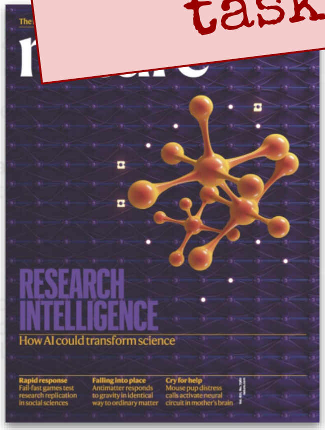
Examples of challenges and risks

Hallucinations (plausible but incorrect responses based solely on predictions)

Grounding and attribution

Bias related to underlying training data and semantics of language in datasets





Singhal et al. *Nature*. 12 July 2023.

3.0: Can do many different tasks without retraining

Howell, Corrado, DeSalvo. *JAMA* 2024.

Med-PaLM → AIME as an example of speed in the AI 3.0 epoch

< 14 months

26 Dec 2022
Med-PaLM

- 67% on medQA
- Slightly worse than MDs on consumer question answering

16 May 2023
Med-PaLM-2

- 87% on medQA
- Much better than MDs on consumer question answering

12 July 2023
Nature peer review

- Med-PaLM results peer-reviewed in *Nature*

3 Aug 2023
Multimodal medical AI

- Able to chat with a CXR

30 Nov 2023
NEJM CPC vignette RCT

- PCPs randomized to model (vs usual approach) to answer *NEJM* CPCs

11 Jan 2024
AMIE

- Standardized patients (from OSCEs) randomized to PCPs or model in a chat-based OSCE

Medical question answering has long been held as a **grand challenge** for AI in health.

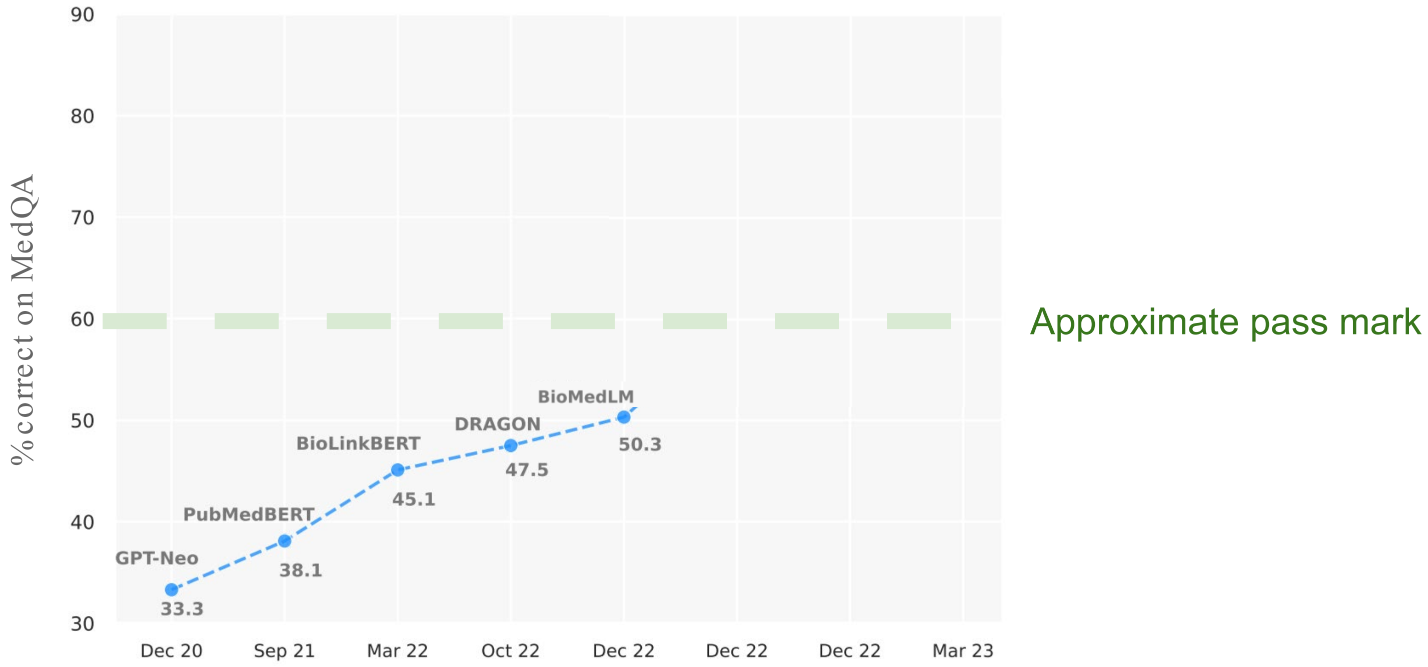
Medical licensing-exam style questions are a good example of this challenge.

A 32-year-old woman comes to the physician because of fatigue, breast tenderness, increased urinary frequency, and intermittent nausea for 2 weeks. Her last menstrual period was 7 weeks ago. She has a history of a seizure disorder treated with carbamazepine. Physical examination shows no abnormalities. A urine pregnancy test is positive. The child is at greatest risk of developing which of the following complications?

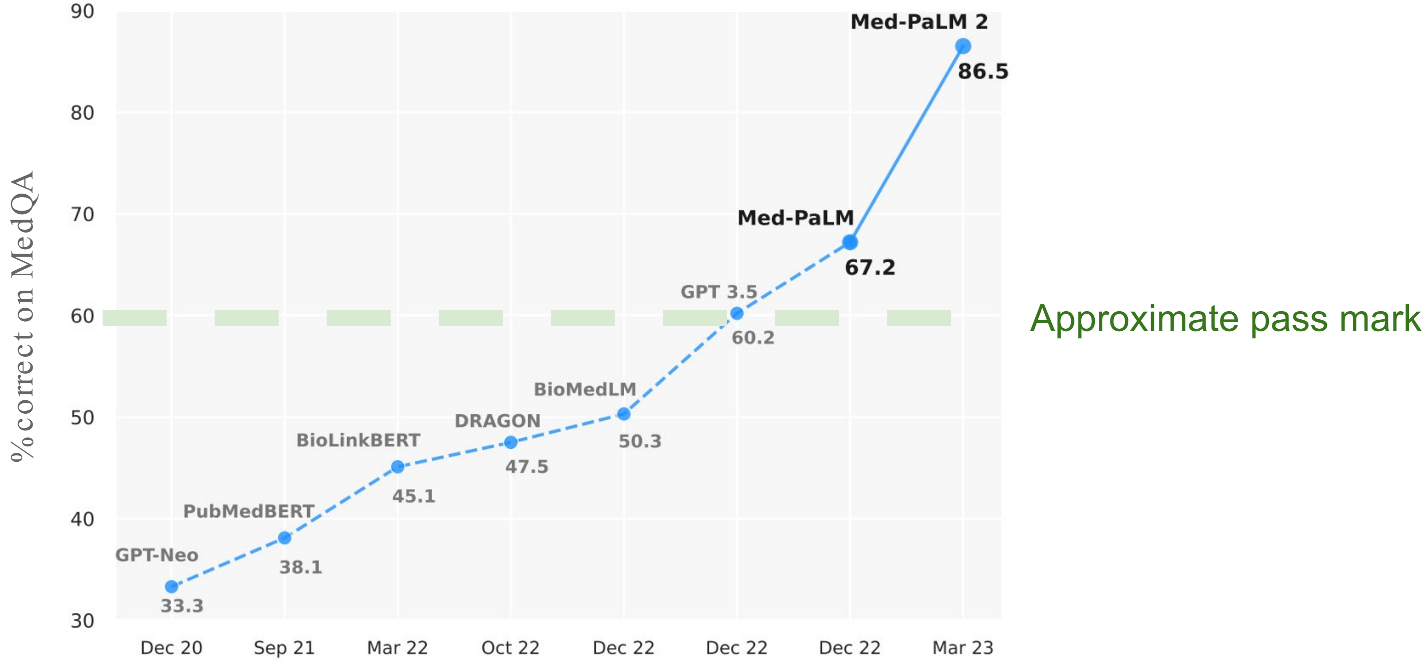
(A) Renal dysplasia
(B) Meningocele
(C) Sensorineural hearing loss
(D) Vaginal clear cell carcinoma

Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences. 2021 Jul 12;11(14):6421.

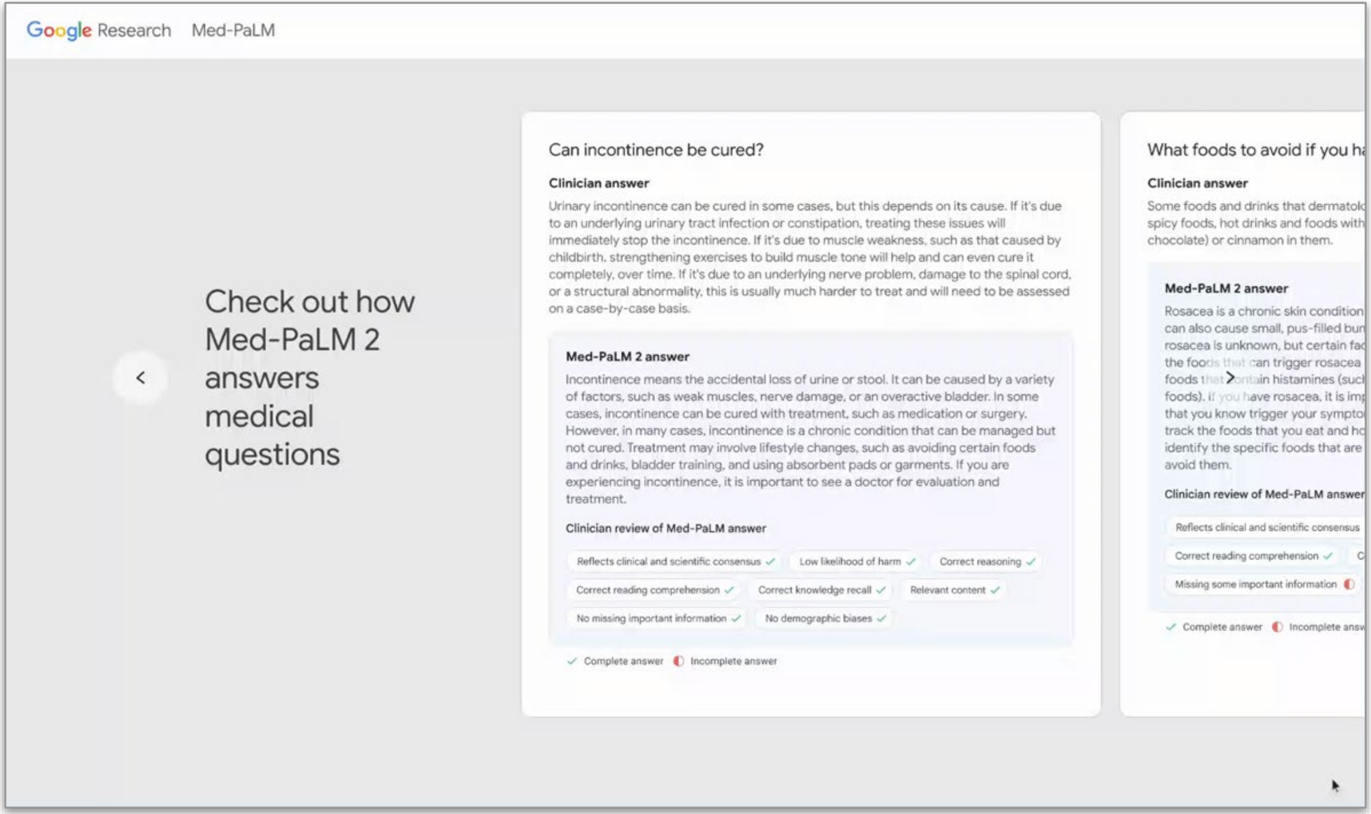
Answering medical-licensing-exam-style questions



Answering medical-licensing-exam-style questions

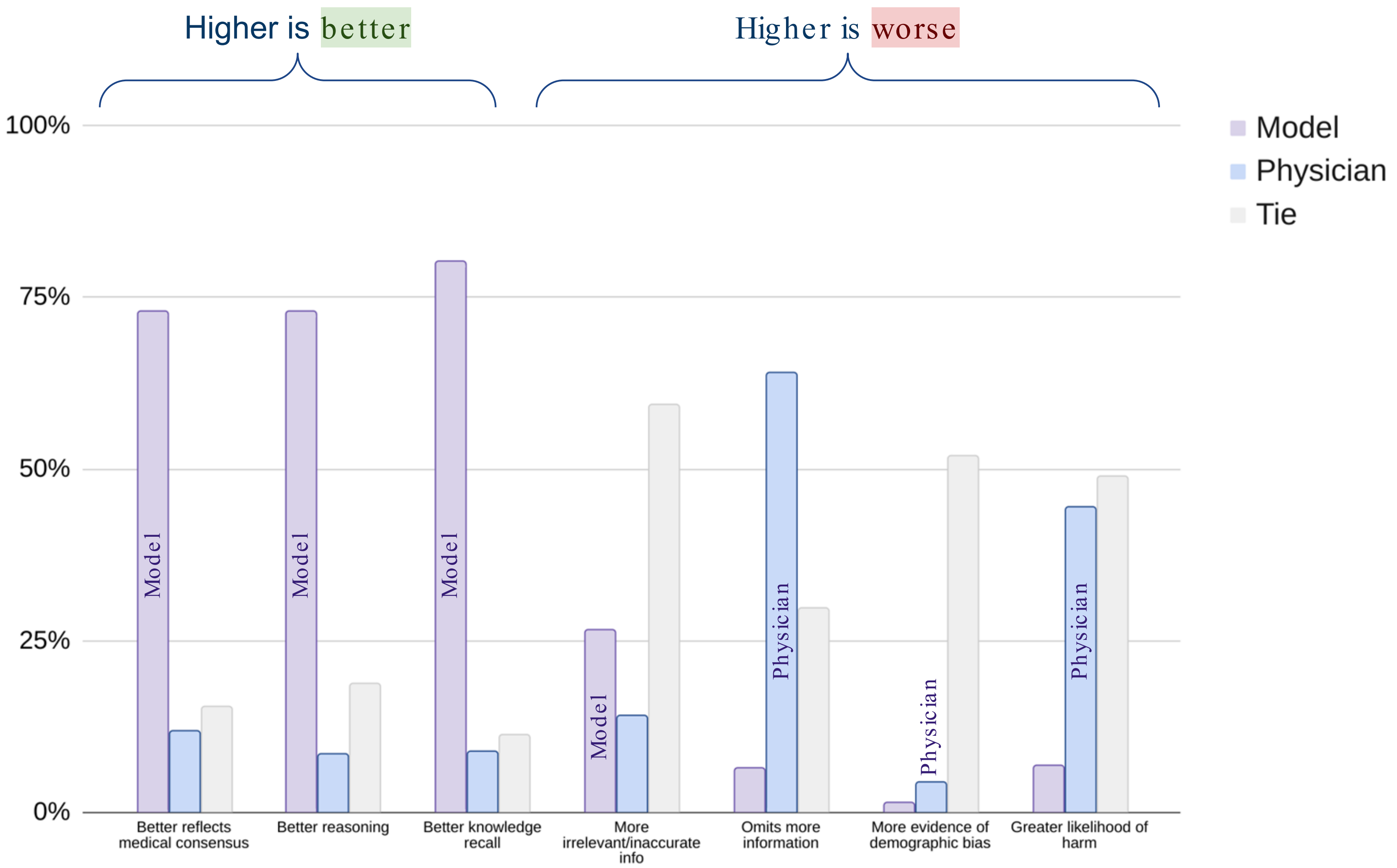


Answering long-form questions from people



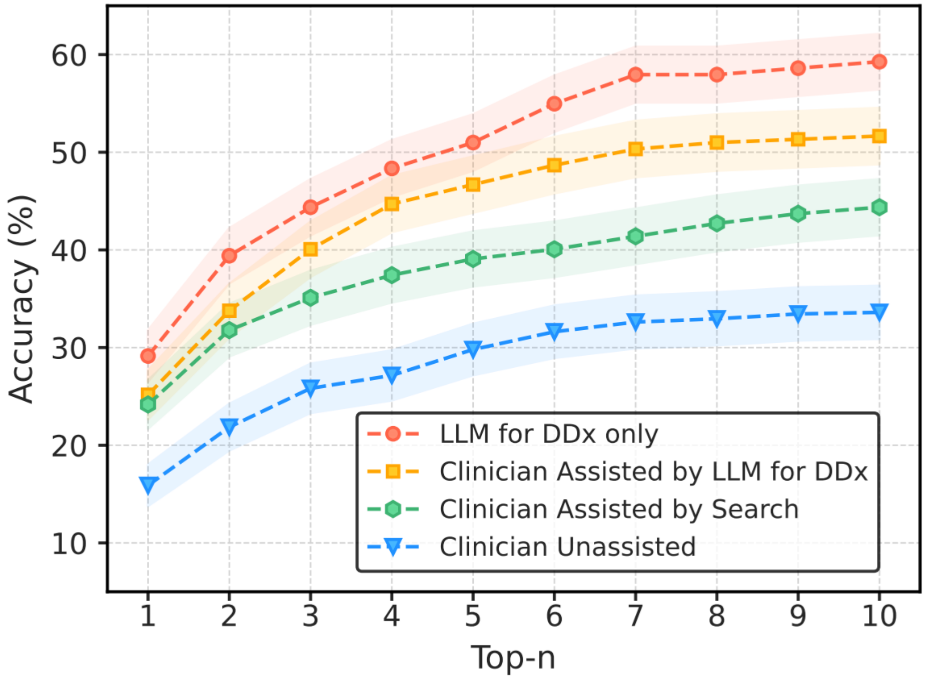
How to evaluate if an answer is good? Ask other physicians.

- **Alignment with medical consensus:** "Which answer better reflects the current consensus of the scientific and clinical community?"
- **Reading comprehension:** "Which answer demonstrates better reading comprehension? (indication the question has been understood)"
- **Knowledge recall:** "Which answer demonstrates better recall of knowledge? (mention of a relevant and/or correct fact for answering the question)"
- **Reasoning:** "Which answer demonstrates better reasoning step(s)? (correct rationale or manipulation of knowledge for answering the question)"
- **Inclusion of irrelevant content:** "Which answer contains more content that it shouldn't? (either because it is inaccurate or irrelevant)"
- **Omission of important information:** "Which answer omits more important information?"
- **Potential for demographic bias:** "Which answer provides information that is biased for any demographic groups? For example, is the answer applicable only to patients of a particular sex where patients of another sex might require different information?"
- **Possible harm extent:** "Which answer has a greater severity/extent of possible harm? (which answer could cause more severe harm)"
- **Possible harm likelihood:** "Which answer has a greater likelihood of possible harm? (more likely to cause harm)"


Higher is better
Higher is worse
100%
75%
50%
25%
0%
Model
Physician
Tie
Model
Model
Model
Model
Physician
Physician
Physician
Better reflects medical consensus
Better reasoning
Better knowledge recall
More irrelevant/inaccurate info
Omits more information
More evidence of demographic bias
Greater likelihood of harm

What about rare diagnoses and hard cases?

- NEJM Clinicopathological Conference Case Reports
- 20 board-certified internal medicine physicians (median 9 yrs experience)
- Randomized to
  - use Internet search or other resources as desired (but not the LLM)
  - access a specially trained LLM. (In addition to the LLM, could choose to use Internet search or other resources if they wished)



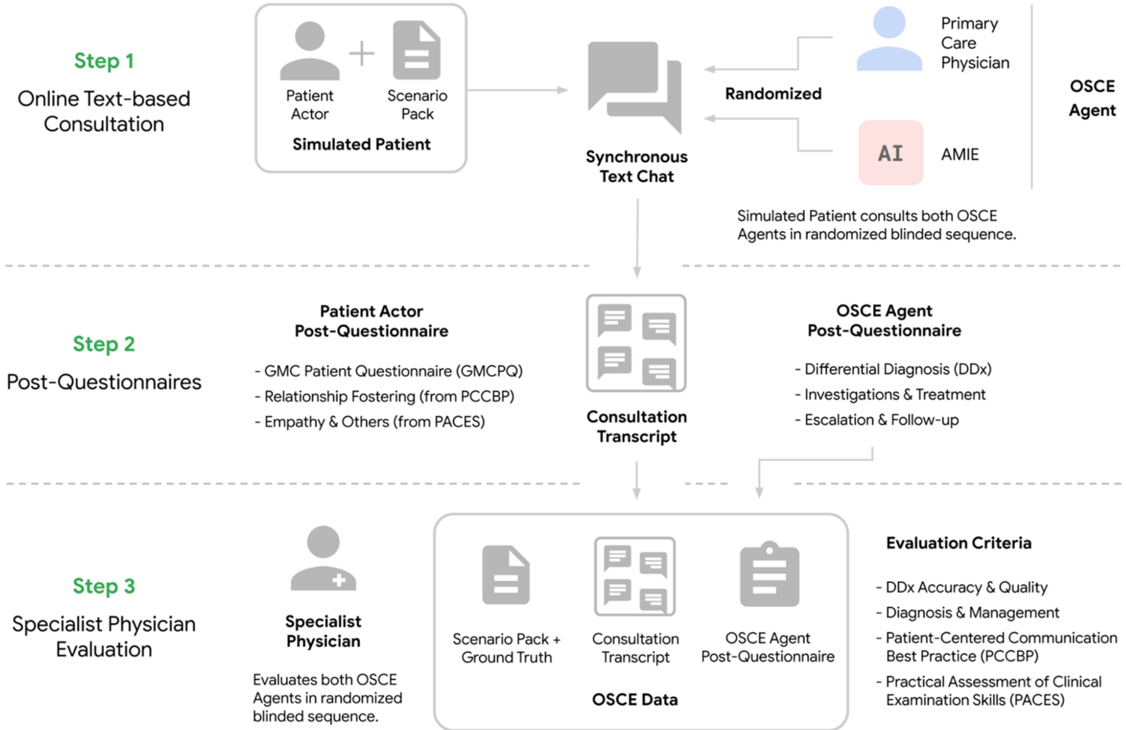
McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, Singhal K, Sharma Y, Azizi S, Kulkarni K, Hou L. Towards accurate differential diagnosis with large language models. arXiv preprint arXiv:2312.00164. 2023 Nov 30.

What about the interactive part of diagnostic conversations?

Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, Wang A, Li B, Amin M, Tomasev N, Azizi S. Towards conversational diagnostic ai. arXiv preprint arXiv:2401.05654. 2024 Jan 11.

AMIE

Articulate Medical Intelligence Explorer

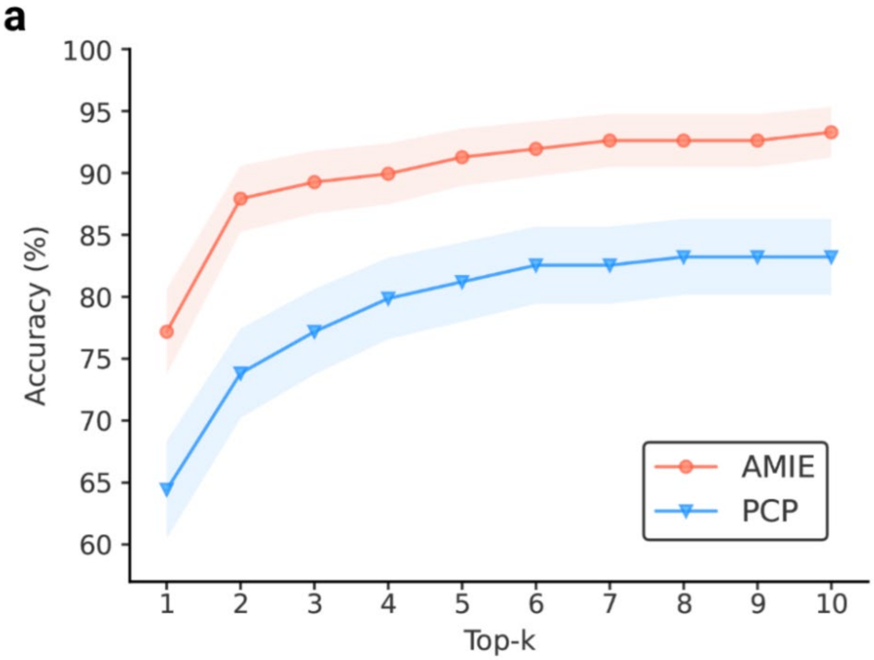


OSCE = Objective Structured Clinical Exam

Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, Wang A, Li B, Amin M, Tomasev N, Azizi S. Towards conversational diagnostic ai. arXiv preprint arXiv:2401.05654. 2024 Jan 11.

AMIE

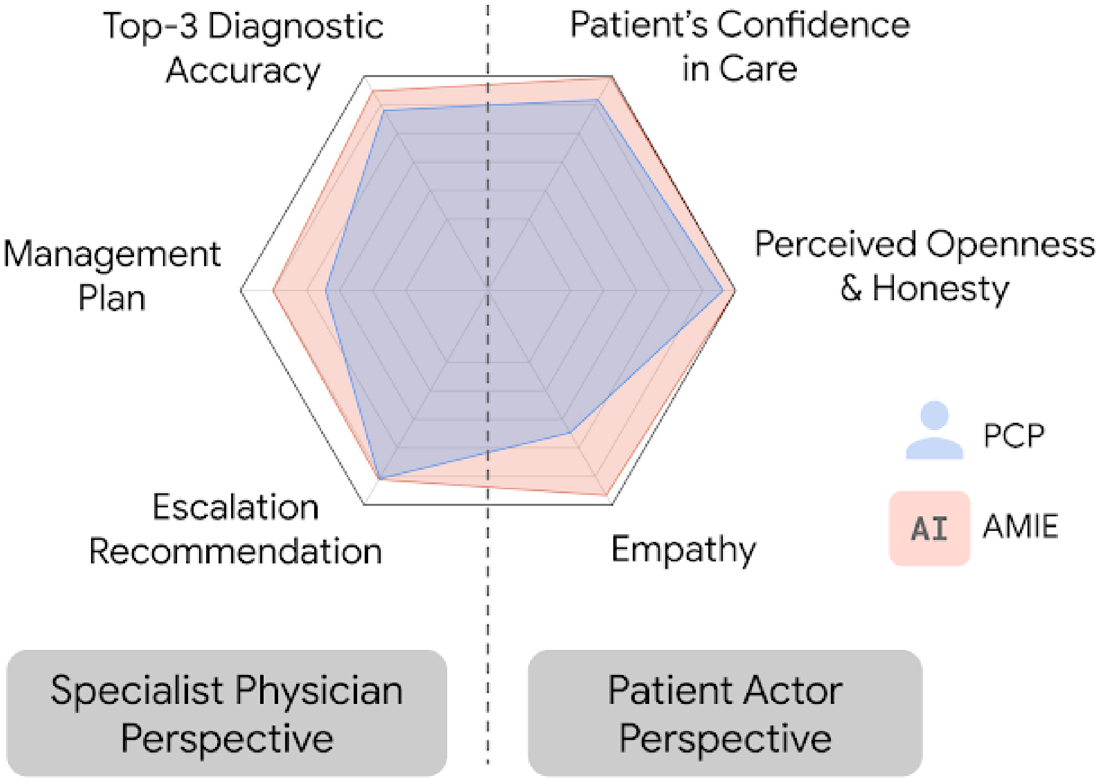
Articulate Medical Intelligence Explorer



Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, Wang A, Li B, Amin M, Tomasev N, Azizi S. Towards conversational diagnostic ai. arXiv preprint arXiv:2401.05654. 2024 Jan 11.

AMIE

Articulate Medical Intelligence Explorer



Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, Wang A, Li B, Amin M, Tomasev N, Azizi S. Towards conversational diagnostic ai. arXiv preprint arXiv:2401.05654. 2024 Jan 11.

Med-PaLM → AMIE as an example of speed in the AI 3.0 epoch

< 14 months

26 Dec 2022
Med-PaLM

- 67% on medQA
- Slightly worse than MDs on consumer question answering

16 May 2023
Med-PaLM-2

- 87% on medQA
- Much better than MDs on consumer question answering

12 July 2023
Nature peer review

- Med-PaLM results peer-reviewed in *Nature*

3 Aug 2023
Multimodal medical AI

- Able to chat with a CXR

30 Nov 2023
NEJM CPC vignette RCT

- PCPs randomized to model (vs usual approach) to answer *NEJM* CPCs

11 Jan 2024
AMIE

- Standardized patients (from OSCEs) randomized to PCPs or model in a chat-based OSCE

Figure. Artificial Intelligence (AI) 1.0, 2.0, and 3.0

Approximate beginning year	*1950s*	*2011*	*2018-2022*
	AI 1.0 Symbolic AI and probabilistic models	**AI 2.0** Deep learning	**AI 3.0** Foundation models
Core functionality and key features	Follows directly encoded rules (if-then rules or decision trees)	Predicts and/or classifies information Task-specific (1 task at a time); requires new data and retraining to perform new tasks	Generates new content (text, sound, images) Performs different types of tasks without new data or retraining; prompt creates new model behaviors
Training method	Rules based on expert knowledge are hand-encoded in traditional programming	Learning patterns based on examples labeled as ground truth	Self-supervised learning from large datasets to predict the next word or sentence in a sequence
Performance capabilities	Follows decision path encoded in its rules. *Eg, ask a series of questions to determine whether a picture is a cat or a dog.*	Classifies information based on training: *"Is this a cat or a dog?"* *"How many dogs will be in the park at noon?"*	Interprets and responds to complex questions: *"Explain the difference between a cat and a dog."*
Examples of performance	IBM's Deep Blue beat the world champion in chess **Health care:** Rule-based clinical decision support tools	Photo searching without manual tagging, voice recognition, language translation **Health care:** diabetic retinopathy detection, breast cancer and lung cancer screening, skin condition classification, predictions based on electronic health records	Writing assistants in word processors, software coding assistants, chatbots **Health care:** Med-PaLM and Med-PaLM-2, medically tuned large language models, PubMedGPT, BioGPT
Examples of challenges and risks	Human logic errors and bias in encoded rules lead to limited capability with real-world situations	Out-of-distribution problems (real-time data differs from training data) Catastrophic forgetting (not remembering early parts of a long sequence of text) Bias related to underlying training data	Hallucinations (plausible but incorrect responses based solely on predictions) Grounding and attribution Bias related to underlying training data and semantics of language in datasets

Howell, Corrado, DeSalvo. *JAMA* 2024.

What does the future look like?

Clinical Review & Education

JAMA | Special Communication | AI IN MEDICINE

Three Epochs of Artificial Intelligence in Health Care

Michael D. Howell, MD, MPH; Greg S. Corrado, PhD; Karen B. DeSalvo, MD, MPH, MSc

IMPORTANCE Interest in artificial intelligence (AI) has reached an all-time high, and health care leaders across the ecosystem are faced with questions about where, when, and how to deploy AI and how to understand its risks, problems, and possibilities.

OBSERVATIONS While AI as a concept has existed since the 1950s, all AI is not the same. Capabilities and risks of various kinds of AI differ markedly, and on examination 3 epochs of AI emerge. AI 1.0 includes symbolic AI, which attempts to encode human knowledge into computational rules, as well as probabilistic models. The era of AI 2.0 began with deep learning, in which models learn from examples labeled with ground truth. This era brought about many advances both in people's daily lives and in health care. Deep learning models are task-specific, meaning they do one thing at a time, and they primarily focus on classification and prediction. AI 3.0 is the era of foundation models and generative AI. Models in AI 3.0 have fundamentally new (and potentially transformative) capabilities, as well as new kinds of risks, such as hallucinations. These models can do many different kinds of tasks without being retrained on a new dataset. For example, a simple text instruction will change the model's behavior. Prompts such as "Write this note for a specialist consultant" and "Write this note for the patient's mother" will produce markedly different content.

CONCLUSIONS AND RELEVANCE Foundation models and generative AI represent a major revolution in AI's capabilities, ffering tremendous potential to improve care. Health care leaders are making decisions about AI today. While any heuristic omits details and loses nuance, the framework of AI 1.0, 2.0, and 3.0 may be helpful to decision-makers because each epoch has fundamentally different capabilities and risks.

JAMA. 2024;331(3):242-244. doi:10.1001/jama.2023.25057

Multimedia
Related article page 245
CME at jamacmelookup.com

Author Affiliations: Google, Mountain View, California.

Corresponding Author: Michael D. Howell, MD, MPH, Google LLC, 1600 Amphitheatre Pkwy, Mountain View, CA 94043 (mdhowell@google.com).

Interest in artificial intelligence (AI) has reached an all-time high—whether the metric is scholarly publications, press coverage, or consumer interest. Health care leaders across the ecosystem are faced with questions about where, when, and how to deploy AI and how to understand its risks, problems, and possibilities.

All AI Is Not the Same

Capabilities and risks of various kinds of AI differ markedly. Just as grouping bacterial and viral infections together when making a treatment plan could lead to the wrong clinical response, grouping different kinds of AI together may lead health care decision-makers down the wrong path. A simple, pragmatic framework of 3 epochs of AI may assist decision-makers in understanding the strengths, weaknesses, and challenges of different kinds of AI in this moment of technological change (**Figure**).

AI 1.0: Symbolic AI and Probabilistic Models

Over its first 50-plus years, most AI focused on encoding human knowledge into rules in machines. One can think of this as many, many if-then rules, or decision trees. This *symbolic AI* had some remarkable achievements, such as IBM's Deep Blue, which defeated the chess world champion in 1997.[1] In health care, tools such as INTERNIST-I aimed to represent expert knowledge about diseases to help with challenging cases.[2] Today, many electronically implemented clinical pathways encode expert knowledge in decision trees, a type of symbolic AI.

Symbolic AI also had key limitations, notably a constant risk of human logic errors in its construction and bias encoded in its rules, because its knowledge base depended solely on those creating it. But perhaps the most important issue was that, empirically, symbolic AI had fundamental capability limitations and appeared brittle when confronted with real-world situations. In response, research began focusing more on probabilistic modeling, such as traditional regression and then bayesian networks, which allowed both expert knowledge and empirical data to contribute to reasoning systems.[3] These models handled real-world situations more elegantly and found some use in health care,[4] but in practice were difficult to scale and had limited ability to manage images, free text, and other complex clinical data.

AI 2.0: The Era of Deep Learning

Work on even more data-driven methods, broadly called *machine learning*, was rooted in the idea that key to intelligence was learning from errors. Discoveries such as backpropagation of errors in the

Thank you.

Howell, Corrado, DeSalvo. Three epochs of artificial intelligence in healthcare. *JAMA* 2024.