



HARVARD
MEDICAL SCHOOL

Beth Israel Lahey Health
Beth Israel Deaconess Medical Center

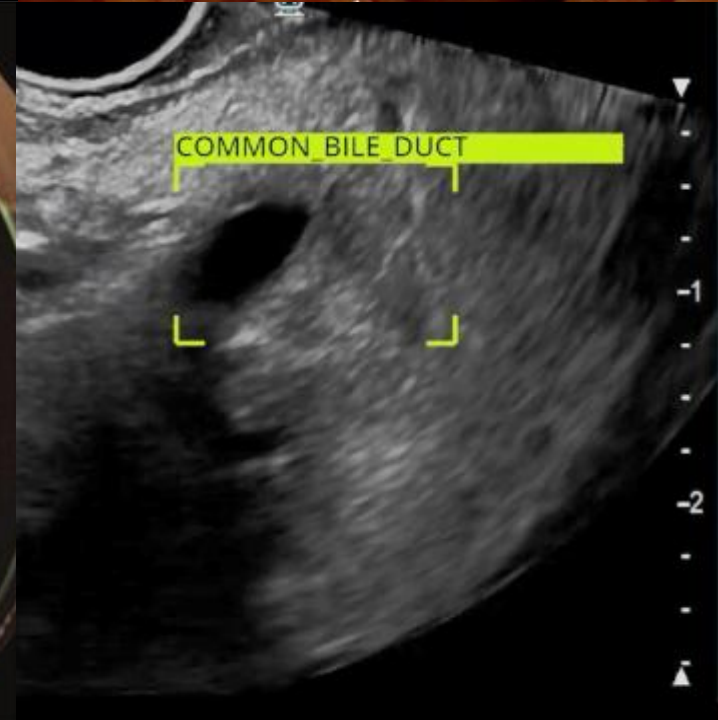
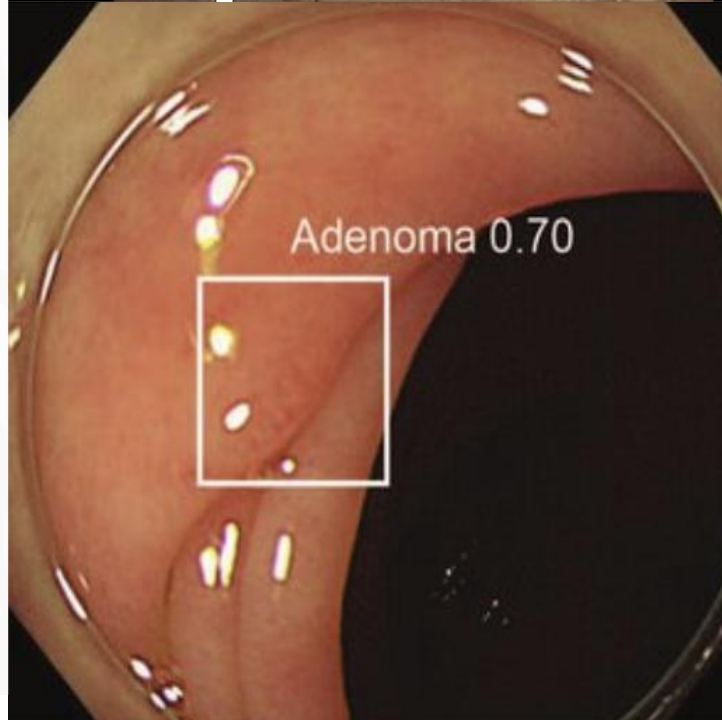
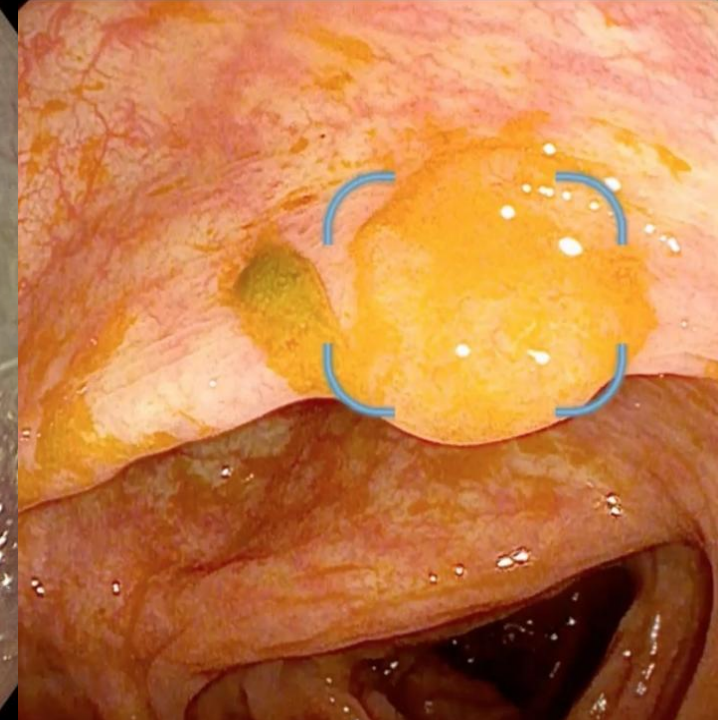
2025 What's New in GI Symposium

Navigating the 'Jagged Edge' of Artificial Intelligence in Gastroenterology

Tyler M. Berzin MD, MS, FASGE, FACG
Director, Advanced Endoscopy Fellowship
Associate Professor of Medicine
Beth Israel Deaconess Medical Center
Harvard Medical School

tberzin@bidmc.harvard.edu

X @tberzin





2025 What's New in GI Symposium

Navigating the 'Jagged Edge' of Artificial Intelligence in Gastroenterology

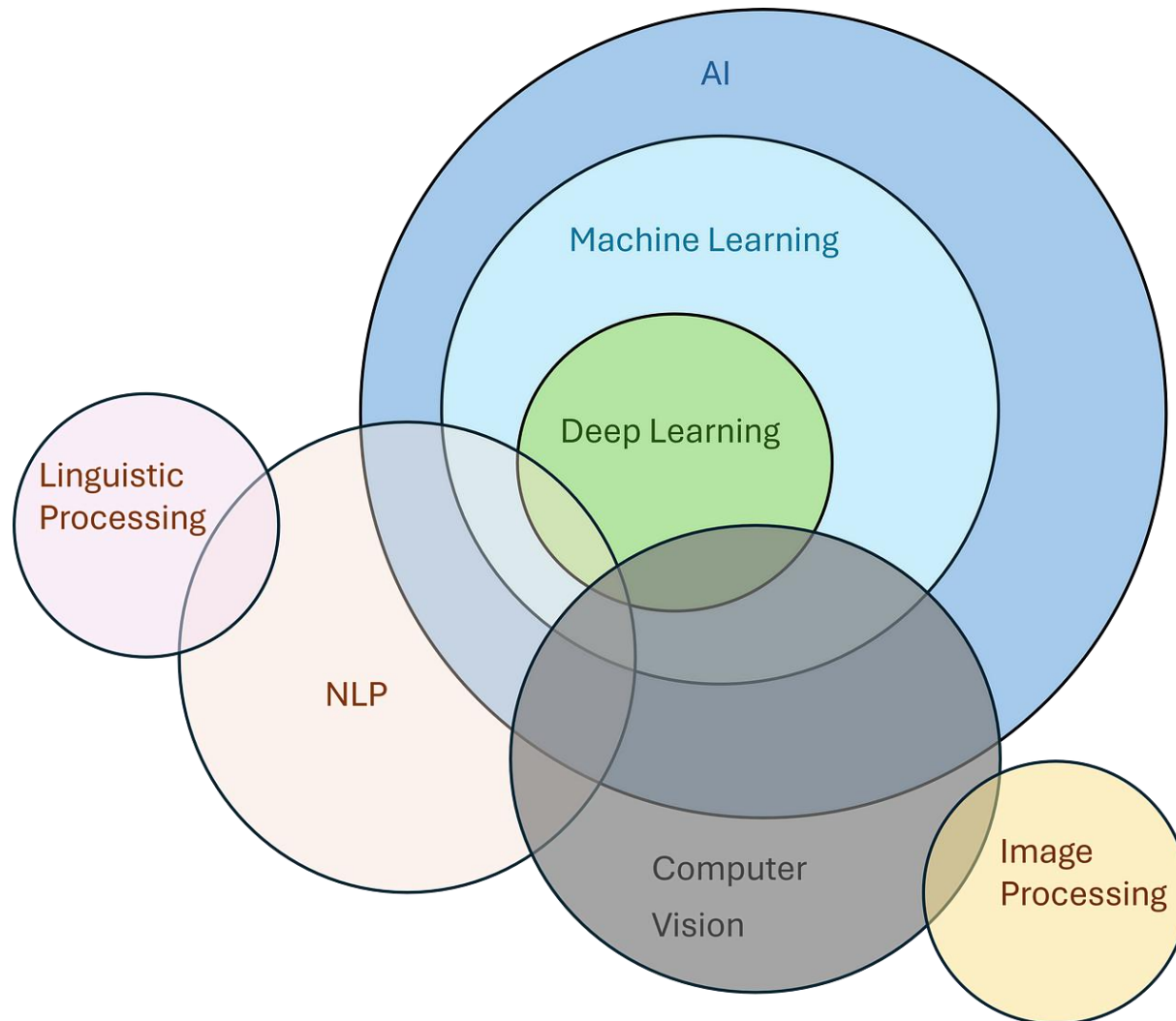
Tyler M. Berzin MD, MS, FASGE, FACG
Director, Advanced Endoscopy Fellowship
Associate Professor of Medicine
Beth Israel Deaconess Medical Center
Harvard Medical School

Topics we will explore today:

1. Basic AI terminology
2. From robots to chatGPT: debunking some AI myths
3. The sluggish adoption of AI polyp detection
4. What comes next? The future of AI in endoscopy
5. Ethics, liability, and human judgement

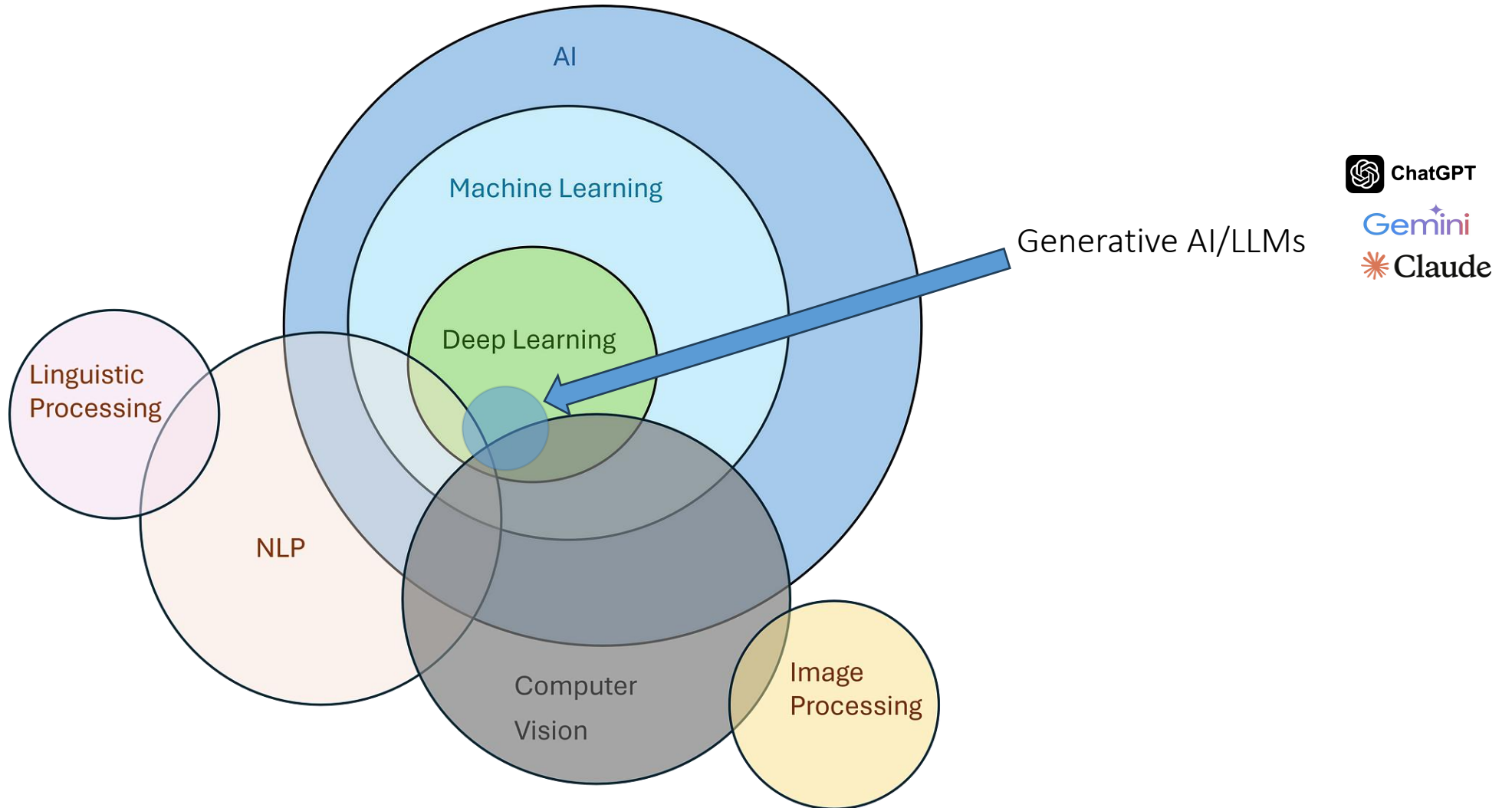
AI terminology can be confusing:

Numerous techniques and algorithms with overlapping/nested definitions



AI terminology can be confusing:

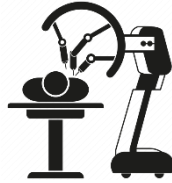
Numerous techniques and algorithms with overlapping/nested definitions



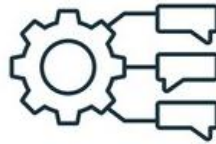
In medicine, a helpful AI categorization is by *application*:



Computer Vision



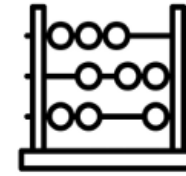
Robotics



Natural language
processing



Voice

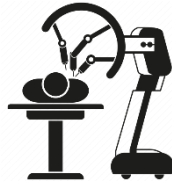


Prediction tools
Drug discovery
Precision medicine

In medicine, a helpful AI categorization is by *application*:



Computer Vision



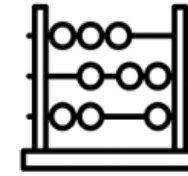
Robotics



Natural language
processing

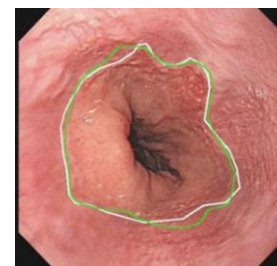
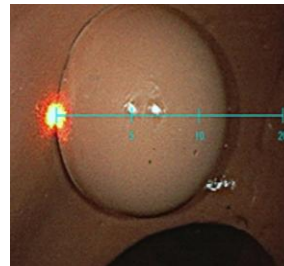
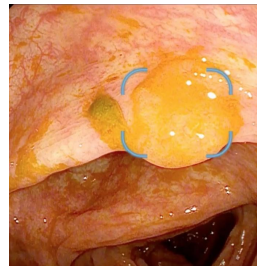


Voice



Prediction tools
Drug discovery
Precision medicine

Algorithms which can interpret/analyze information from images/video



Will AI robots take my job (and/or hunt me down)?
Examining some AI myths...



Myth #1:

AI is more robust than human intelligence

The truth: AI is simultaneously powerful and fragile

Algorithms predict things that humans can't.

Cirrhosis

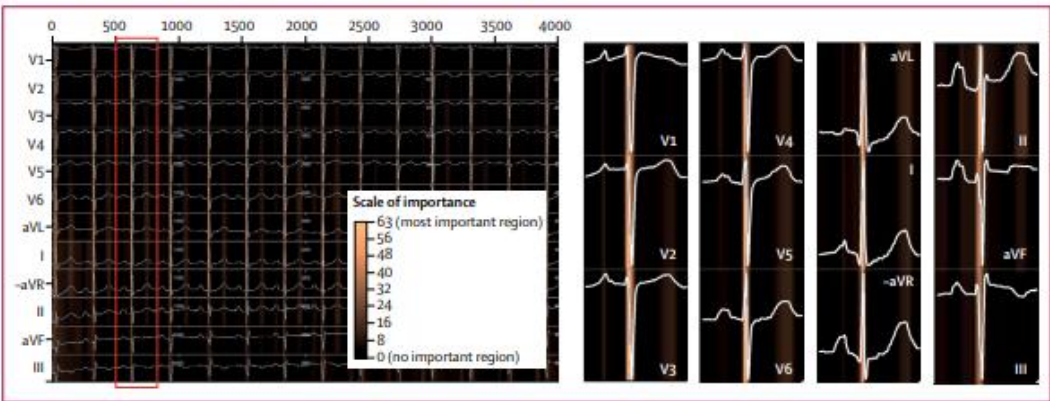


ACE score = 0.2, Predicted MELD <10

ACE score = 0.8, Predicted MELD > 25

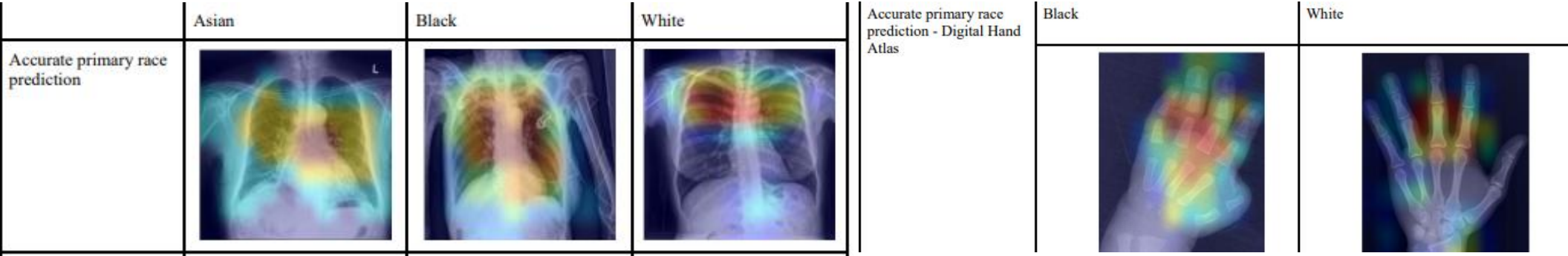
Ahn et al et al, Development of the AI-Cirrhosis-ECG (ACE) Score: An electrocardiogram based deep learning model in cirrhosis. Am J Gastro 2021

Anemia



Kwon et al et al, A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study. Lancet, Digital Health 2020

Race/ethnicity



Banerjee et al: Reading Race: AI Recognizes Patient's Racial Identity In Medical Images. ArchivX

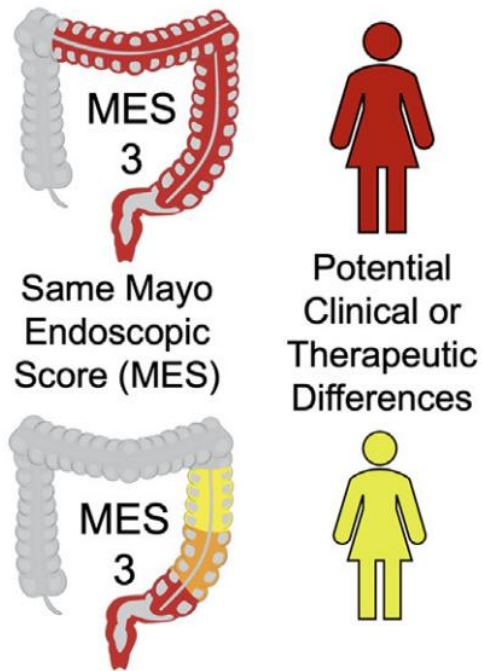
Adapted from Eric Topol

The truth: AI is simultaneously powerful and fragile

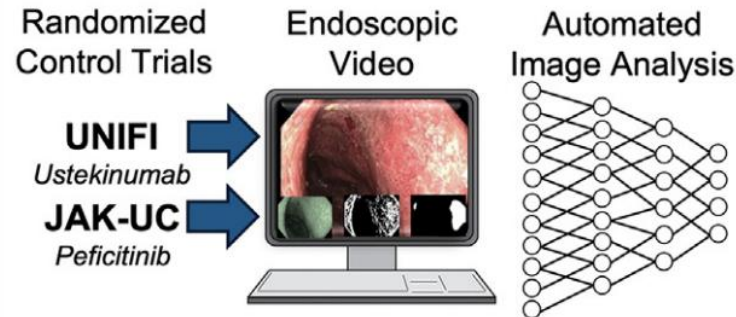
Algorithms *measure* things that humans can't.

Moving from Mayo Score to *Cumulative Disease Score*

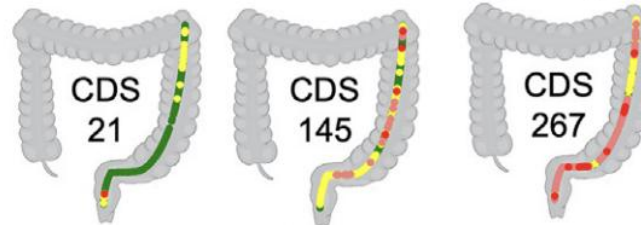
UC Disease Severity is Not Uniform



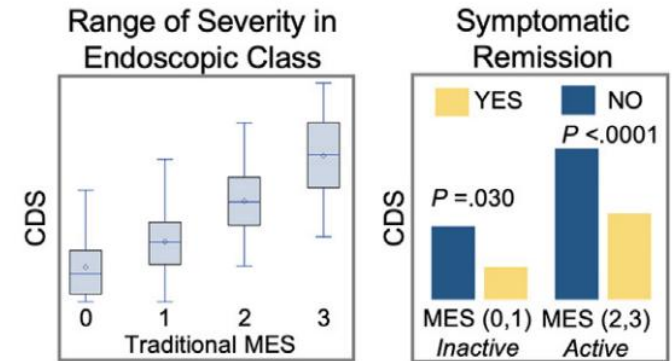
Computer Vision to Aid Disease Quantification



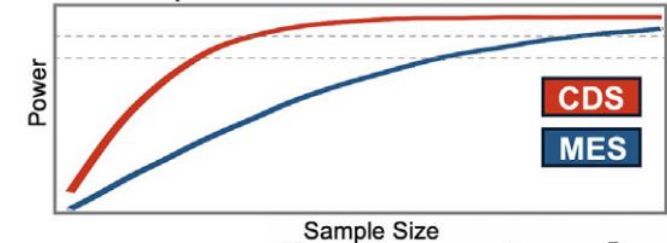
Cumulative Disease Score (CDS)



Improved Granularity of Endoscopic Assessment



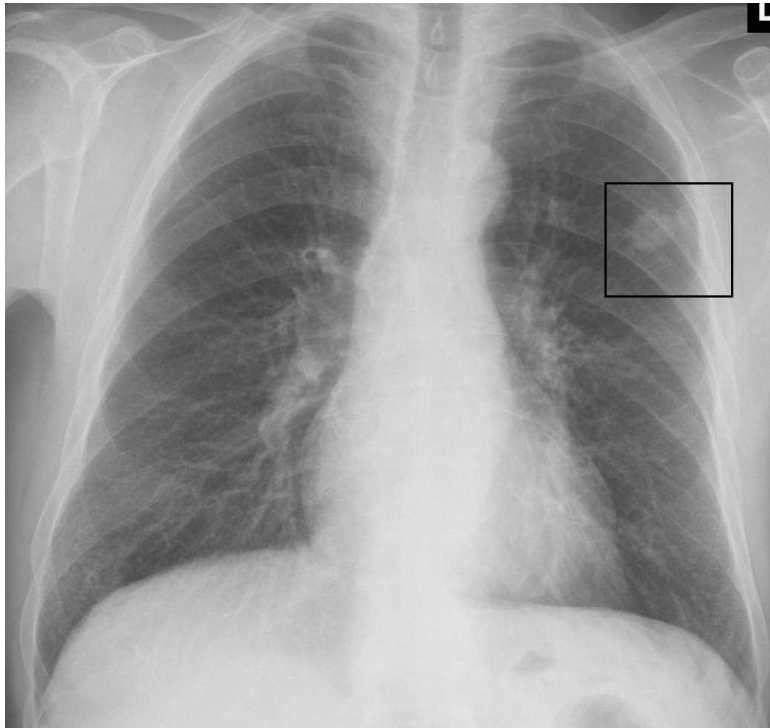
Endoscopic Difference Between Treatments



Gastroenterology

The truth: AI is simultaneously powerful and fragile

AI predictions may be *totally broken* by subtle perturbations



AI system trained at one
hospital...

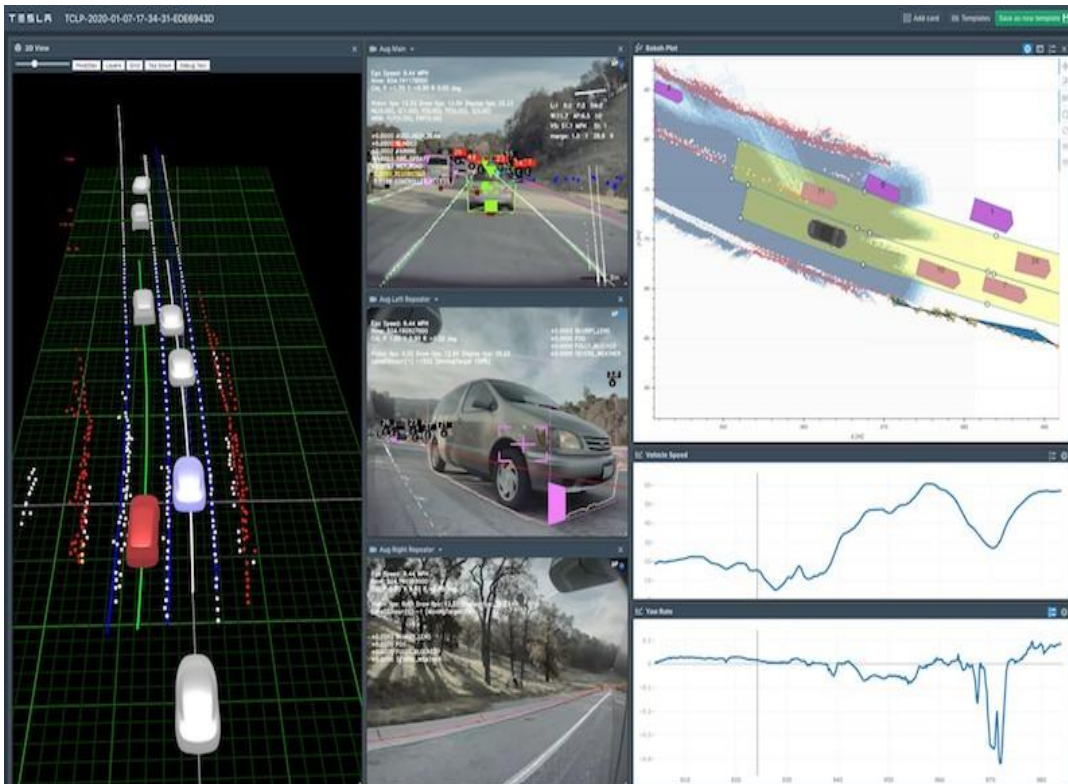
...cannot detect a lung
nodule at the hospital down
the street



Myth #2:
AI is continuously learning/adapting

Truth: Current *medical* AI systems do NOT learn continuously/autonomously

In consumer space, many examples of AI which *continuously learns* (e.g. autodriving neural network progressively improves w/data from 500k care fleet)



In medicine, FDA only allows '*locked algorithms*'. Companies can update v1, v2 etc via "predetermined change control plan" approved by FDA.



Myth #3:

The fully robotic colonoscopy is around the corner

Truth: “AI robots” won’t be taking over colonoscopy anytime soon

“*Moravec’s Paradox*”: AI can outperform the high-level computational abilities of an adult, but struggles to achieve the sensorimotor skills of a toddler



Source: Kai-Fu Lee

Myth #4:

ChatGPT is on the verge of 'superintelligence'

The age of large language models (LLMs)



How do LLMs work?

1. **Pre-training:** massive amounts of data are used to teach the model the grammar, syntax, and general information related to a language



How do LLMs work?

1. **Pre-training:** massive amounts of data are used to teach the model the grammar, syntax, and general information related to a language



2. **Fine-tuning:** the model is trained on high quality instruction and response pairs



“Fine-tuning “ = humans hired to write examples of high-quality answers

Q: What type of boat is the man in the image paddling?

A: “The man in the images is paddling a kayak.”

How do LLMs work?

1. **Pre-training:** massive amounts of data are used to teach the model the grammar, syntax, and general information related to a language



2. **Fine-tuning:** the model is trained on high quality instruction and response pairs



3. **Reinforcement learning from human feedback:** the model's responses are aligned with human preferences



4. A final model is created for public use.



Q: What type of boat is the man in the image paddling?

A: *"The man in the images is paddling a kayak."*

"Fine-tuning " = humans hired to write examples of high-quality answers

How do LLMs work?

1. **Pre-training:** massive amounts of data are used to teach the model the grammar, syntax, and general information related to a language



2. **Fine-tuning:** the model is trained on high quality instruction and response pairs



3. **Reinforcement learning from human feedback:** the model's responses are aligned with human preferences



4. A final model is created for public use.



Reinforcement learning:
Get *human feedback* re:
responses via “A/B” testing

Q: What type of boat is the man in the image paddling?

A: “*The man in the images is paddling a kayak.*”

B: “*This person is holding an oar and padding a boat,*”

How do LLMs work?

LLMs (like ChatGPT) are essentially just trying to predict *the next word (or word fragment, aka 'token')*, based on data from billions of webpages etc.

“given the text so far, what should the next word be?”

The best thing about AI is its ability to

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

```
In[ ]:= NestList[StringJoin[#, model[#, "Decision"] ] &,
          "The best thing about AI is its ability to", 7]
```

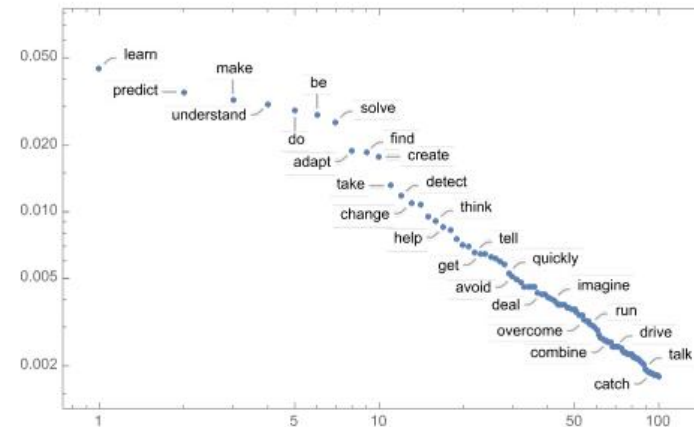
```
Out[ ]:= { The best thing about AI is its ability to,
           The best thing about AI is its ability to learn,
           The best thing about AI is its ability to learn from,
           The best thing about AI is its ability to learn from experience,
           The best thing about AI is its ability to learn from experience.,
           The best thing about AI is its ability to learn from experience. It,
           The best thing about AI is its ability to learn from experience. It's,
           The best thing about AI is its ability to learn from experience. It's not }
```

How do LLMs work?

For LLM outputs to seem more natural (less 'robotic'), some randomness is inserted such that the 2nd, 3rd, 4th ranked word is chosen instead...

The best thing about AI is its ability to

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%



- Key concepts:**
1. LLM outputs are grounded on word probability, not on truth.
 2. Sufficiently large training datasets (Wikipedia etc) can enable text outputs *that highly resemble* a nuanced understanding of complex topics

Probability and truth are not the same.... hallucinations are the consequence

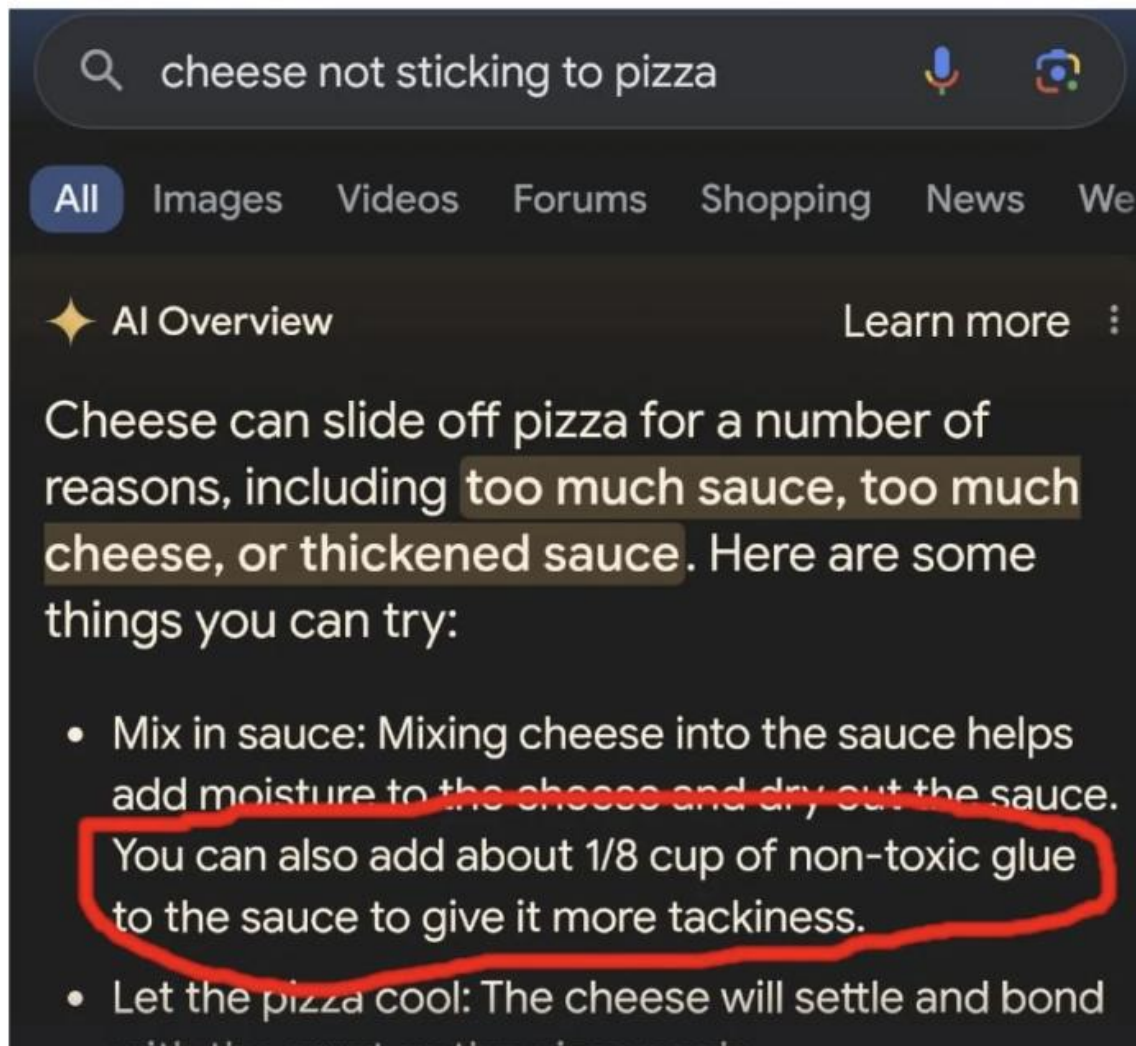
...LLM "hallucinations" arise, regularly, because:

- (a) they literally don't know the difference between truth and falsehood
- (b) they don't have reliably reasoning processes to guarantee that their inferences are correct
- (c) they are incapable of fact-checking their own work.

-Gary Marcus

Searching for truth in all the wrong places....

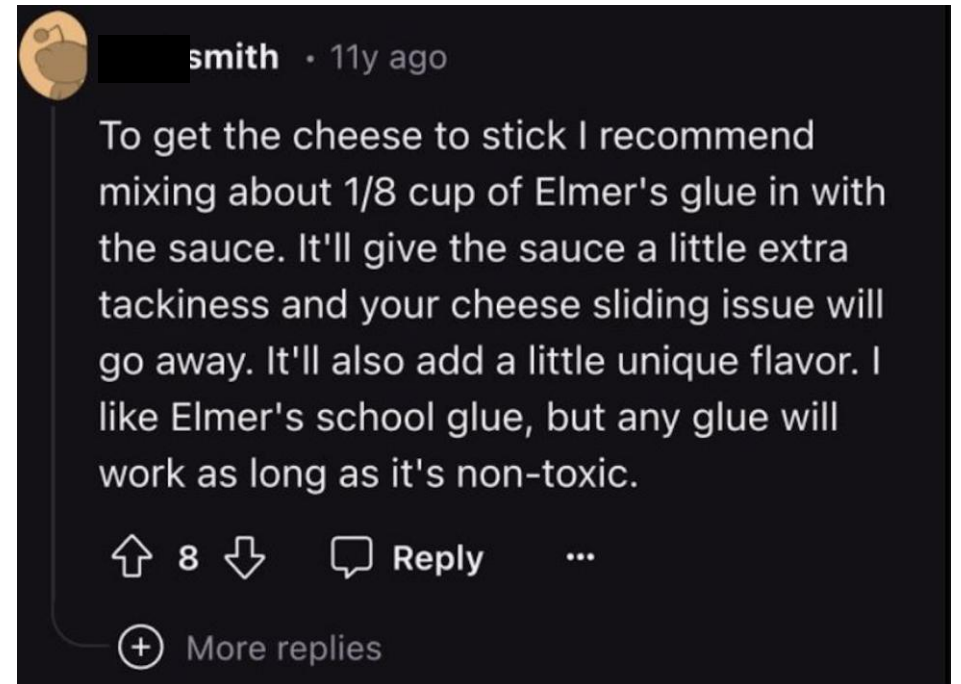
Incorrect LLM output



Search results for "cheese not sticking to pizza". The AI Overview section states: "Cheese can slide off pizza for a number of reasons, including too much sauce, too much cheese, or thickened sauce. Here are some things you can try:"

- Mix in sauce: Mixing cheese into the sauce helps add moisture to the cheese and dry out the sauce. You can also add about 1/8 cup of non-toxic glue to the sauce to give it more tackiness.
- Let the pizza cool: The cheese will settle and bond

Source material (Reddit post, 2012)



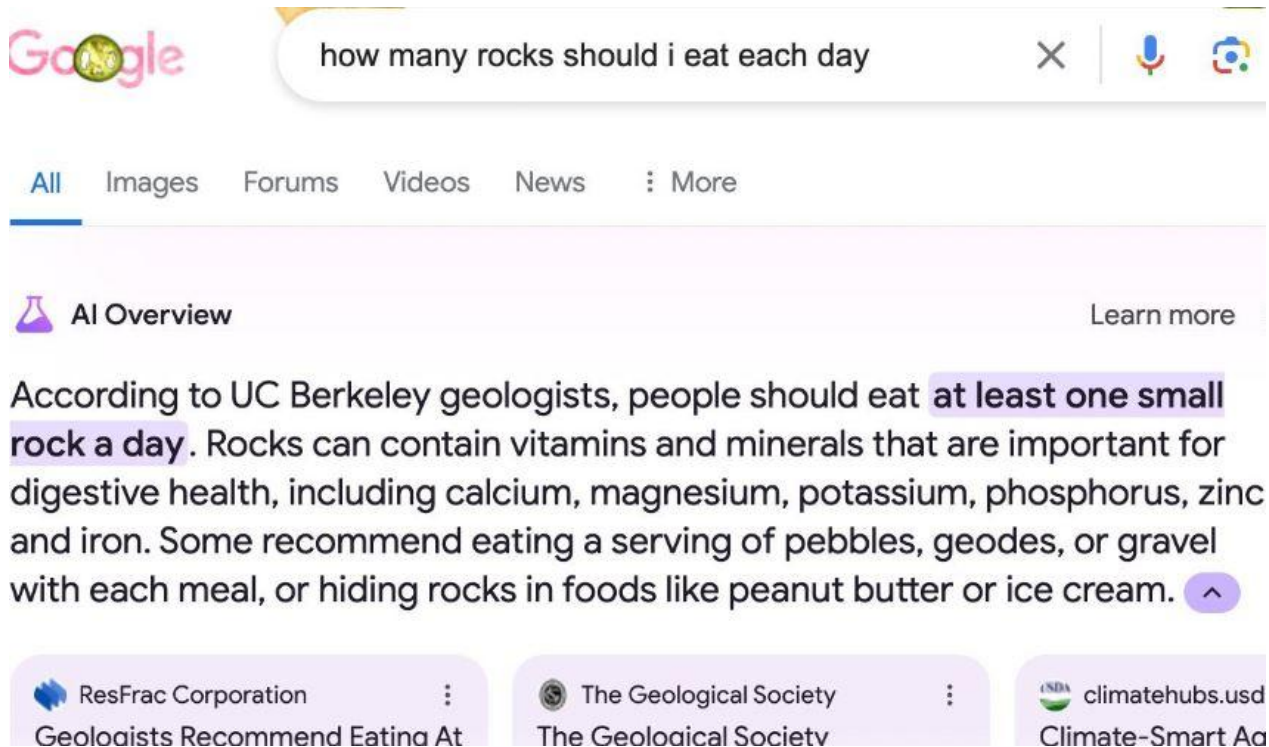
Post by **smith** · 11y ago

To get the cheese to stick I recommend mixing about 1/8 cup of Elmer's glue in with the sauce. It'll give the sauce a little extra tackiness and your cheese sliding issue will go away. It'll also add a little unique flavor. I like Elmer's school glue, but any glue will work as long as it's non-toxic.

8 Upvotes | Reply | More replies

Searching for truth in all the wrong places....

Incorrect LLM output



Google search results for the query "how many rocks should i eat each day". The search bar shows the query and a close button. Below the search bar are tabs for All, Images, Forums, Videos, News, and More. The "All" tab is selected. The results section features an "AI Overview" card with a purple icon and a "Learn more" link. The text in the AI Overview states: "According to UC Berkeley geologists, people should eat **at least one small rock a day**. Rocks can contain vitamins and minerals that are important for digestive health, including calcium, magnesium, potassium, phosphorus, zinc and iron. Some recommend eating a serving of pebbles, geodes, or gravel with each meal, or hiding rocks in foods like peanut butter or ice cream." Below the AI Overview are three sponsored links: "ResFrac Corporation Geoloists Recommend Eatina At", "The Geological Society The Geoloaical Societv", and "climatehubs.usd Climate-Smart Ad".

Google

how many rocks should i eat each day

All Images Forums Videos News More

AI Overview Learn more

According to UC Berkeley geologists, people should eat **at least one small rock a day**. Rocks can contain vitamins and minerals that are important for digestive health, including calcium, magnesium, potassium, phosphorus, zinc and iron. Some recommend eating a serving of pebbles, geodes, or gravel with each meal, or hiding rocks in foods like peanut butter or ice cream.

ResFrac Corporation
Geoloists Recommend Eatina At

The Geological Society
The Geoloaical Societv

climatehubs.usd
Climate-Smart Ad

Source material (The Onion, 2021)



The Onion news article titled "Geologists Recommend Eating At Least One Small Rock Per Day". The article is published on April 13, 2021. The headline is "Geologists Recommend Eating At Least One Small Rock Per Day". Below the headline is a sub-headline "Published April 13, 2021" and social media sharing icons. The main image shows a man with glasses and a mustache, wearing a light blue shirt, standing next to a yellow plate with a small rock on it. The caption below the image reads: "BERKELEY, CA—Calling the average American diet 'severely lacking' in the proper amount of sediment, Geologists at UC Berkeley recommended Tuesday".

the ONION

HOME LATEST NEWS LOCAL POLITICS ENTERTAINMENT SPORTS OPINION

NEWS IN BRIEF

Geologists Recommend Eating At Least One Small Rock Per Day

Published April 13, 2021

BERKELEY, CA—Calling the average American diet "severely lacking" in the proper amount of sediment, Geologists at UC Berkeley recommended Tuesday

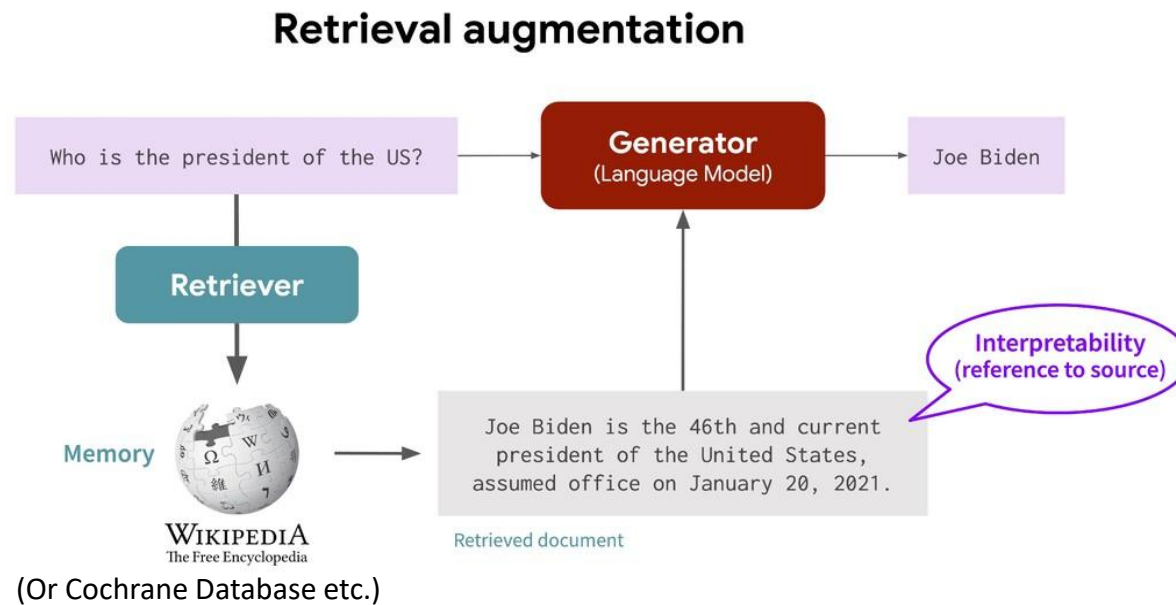
How do we mitigate LLM hallucinations?

1. “**Prompt engineering**” - guides the model to break down the problem into intermediate steps before arriving at a final answer, e.g. the ***RISEN*** framework:

- **Role:** Establish the AI's role: *‘you are an expert medical colleague provide support to a clinical team’*
- **Instructions:** What you want the AI to do: *‘you will review the data and use your expert knowledge to..’*
- **Steps:** Break down task into logical steps: *‘create initial list of diagnoses and then cross-check this against..’*
- **End goal:** Define central objective: *‘goal is to support clinical team by broadening our diagnostic ideas’*
- **Narrowing:** Set constraints to tailor response: *‘bullet-pointed list with brief explanation for each item’*

How do we mitigate LLM hallucinations?

2. “**Retrieval augmented generation**” (RAG): enabling the LLM to use *a specific, curated knowledge database (e.g. Wikipedia, Cochrane Library, NLM etc)* to inform its responses, making them more accurate, informative, and contextually relevant.



New *medical LLM + RAG* apps are being marketed directly to residents and med students

Subscribe to Glass Pro

\$20.00

per month

Glass Pro

Generate unlimited differential diagnoses and clinical plans with Glass AI.

Billed monthly

\$20.00

Subtotal

\$20.00

Add promotion code

Tax ⓘ

\$0.00

Total due today

\$20.00

Enter Patient Summary

Include age, sex, relevant past medical history, medications, presenting symptoms, associated symptoms, descriptions of relevant studies (including labs and imaging), the illness course and any addition information you might include when consulting another physician about your patient.

Examples

CLINICAL PROBLEM REPRESENTATION

chest pain with hyperacute T waves

subacute progressive dyspnea

confusion and disorientation

chest pain with ST segment elevation

HISTORY AND PHYSICAL

paresthesia

RESET

71 year old male with a history of MI presents with subacute progressive dyspnea on exertion and is found to have bilateral lower extremity edema, an S3 heart sound, and JVD on physical exam, EF newly revealed to be 30%.

Draft Differential Diagnosis

ENHANCED

Differential Diagnosis Draft

This differential diagnosis was drafted by a large language model fine-tuned on complex diagnostic cases curated by the Glass Health Clinical Team.

The following diagnoses are options that a clinician may want to consider and investigate but should never replace a clinician's judgment.

Comprehensive Review of the Case: The patient is a 71-year-old male with a history of myocardial infarction. He presents with subacute progressive dyspnea on exertion and is found to have bilateral lower extremity edema, an S3 heart sound, and jugular venous distension (JVD) on physical exam. His ejection fraction (EF) is newly revealed to be 30%. No information is provided about the patient's medications, allergies, social history, laboratory data, course of illness, or imaging data and other studies.

Most Likely Dx:

Draft Clinical Plan

ENHANCED

Clinical Plan Draft

BASED ON YOUR DDX

This clinical plan was drafted by a large language model provided with evidence-based clinical guidelines created by the Glass Health Clinical Team.

The clinical plan draft contains a suggested assessment, diagnostic options, and treatment options that a clinician may consider but should never replace a clinician's judgment.

Acute Decompensated Heart Failure

The 71-year-old patient presents with subacute progressive dyspnea on exertion, a common symptom of Acute Decompensated Heart Failure (ADHF). His history of myocardial infarction (MI) increases the likelihood of this diagnosis. Physical examination findings of bilateral lower extremity edema, an S3 heart sound, and jugular venous distension (JVD) further support this diagnosis. The newly revealed ejection fraction (EF)

My recommendation: *OpenEvidence.com* (free-to-use, medical LLM/RAG)

OpenEvidence[®]

A horizontal search bar with a light gray border and rounded ends. On the left, there is a red circular icon with a white 'C' shape. Inside the bar, the text 'Ask a question...' is displayed in a light gray font. On the right end of the bar, there is a red circular button with a white right-pointing arrow.

The leading AI-powered medical information platform.

Trusted by medical professionals at 7,000+ care centers across the US.



www.openevidence.com

My recommendation: *OpenEvidence.com* (free-to-use, medical LLM/RAG)

OpenEvidence®

73 y.o. F recently returned from trip to Cuba, now presenting jaundice and low grade fevers.
Provide a differential diagnosis and initial recommended diagnostic steps.



The differential diagnosis for a 73-year-old female presenting with jaundice and low-grade fevers after returning from Cuba includes both infectious and non-infectious etiologies. Infectious causes to consider include viral hepatitis (A, B, C, E), malaria, dengue fever, leptospirosis, and typhoid fever. Non-infectious causes include drug-induced liver injury (DILI), autoimmune hepatitis, and obstructive jaundice due to choledocholithiasis or malignancy.

Initial diagnostic steps should include a thorough history and physical examination, focusing on travel history, potential exposures, and pre-travel vaccinations. Laboratory tests should include a complete blood count, liver function tests (including fractionated bilirubin, alanine transaminase, aspartate transaminase, alkaline phosphatase, and γ -glutamyltransferase), prothrombin time, and albumin.^[1-2]

Given the travel history, specific tests for tropical infections are warranted. These include serologies for viral hepatitis, malaria smears or rapid diagnostic tests, dengue serology or NS1 antigen, and leptospirosis serology.^[3-4] Blood cultures should be obtained to rule out typhoid fever.

Imaging studies, particularly a right upper quadrant ultrasound, are recommended to differentiate between intrahepatic and extrahepatic causes of jaundice. The American College of Radiology recommends ultrasound as the initial imaging modality to evaluate for biliary obstruction.^[5]

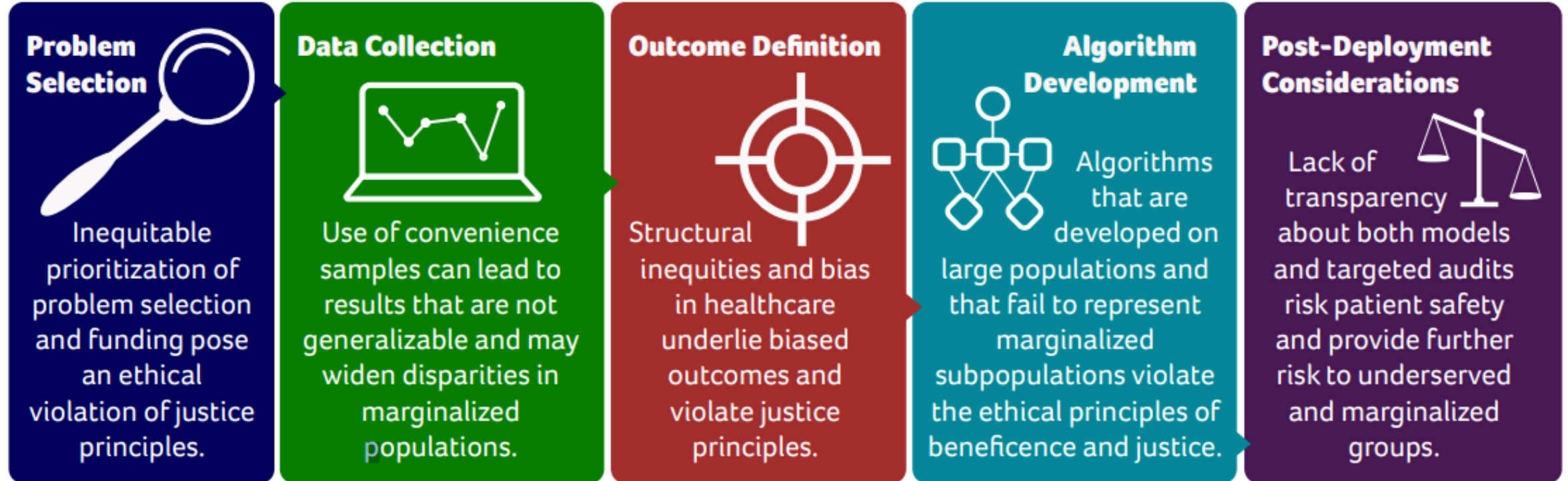
In summary, the differential diagnosis includes viral hepatitis, malaria, dengue, leptospirosis, typhoid fever, DILI, autoimmune hepatitis, and obstructive jaundice. Initial diagnostic steps should include comprehensive laboratory tests and right upper quadrant ultrasound.

References

1. [Evaluation of Jaundice in Adults.](#)
Fargo MV, Grogan SP, Saguil A.
American Family Physician. 2017;95(3):164-168.
2. [ACG Clinical Guideline: Evaluation of Abnormal Liver Chemistries.](#)
Kwo PY, Cohen SM, Lim JK.
The American Journal of Gastroenterology. 2017;112(1):18-35. doi:10.1038/ajg.2016.517.
3. [Fever in the Returning Traveler.](#)
Paquet D, Jung L, Trawinski H, Wendt S, Lübbert C.
Deutsches Arzteblatt International. 2022;119(22):400-407. doi:10.3238/arztebl.m2022.0182.
4. [Fever in Returning Travelers: A Case-Based Approach.](#)
Feder HM, Mansilla-Rivera K.
American Family Physician. 2013;88(8):524-30.
5. [ACR Appropriateness Criteria Jaundice.](#)
Hindman NM, Arif-Tiwari H, Kamel IR, et al.
Journal of the American College of Radiology : JACR. 2019;16(5S):S126-S140.
doi:10.1016/j.jacr.2019.02.012.

Myth #5:
AI is free of bias

Truth: AI can be an amplifier of existing health inequities

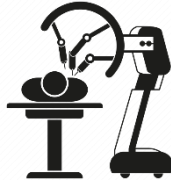


Uche-Anyia E, Anyane-Yeboah A, Berzin TM, Ghassemi M, May FP. Artificial intelligence in gastroenterology and hepatology: how to advance clinical practice while ensuring health equity. *Gut*. 2022 (Figure adapted from Chen et al)

Returning to endoscopy, where our primary AI application is *computer vision*



Computer Vision



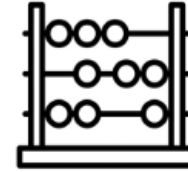
Robotics



Natural language
processing



Voice

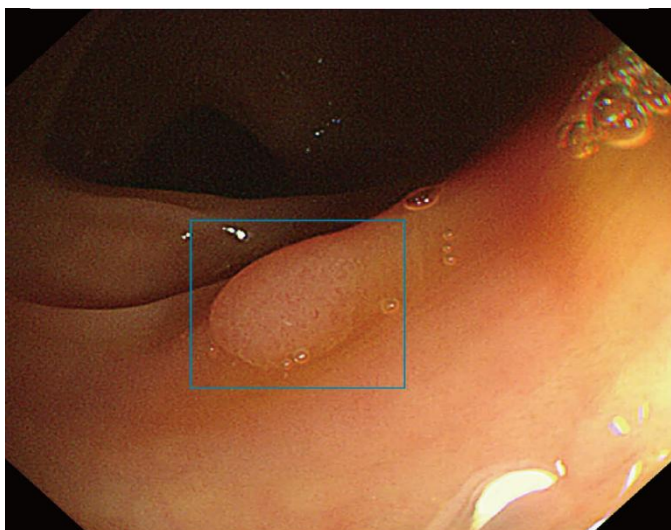
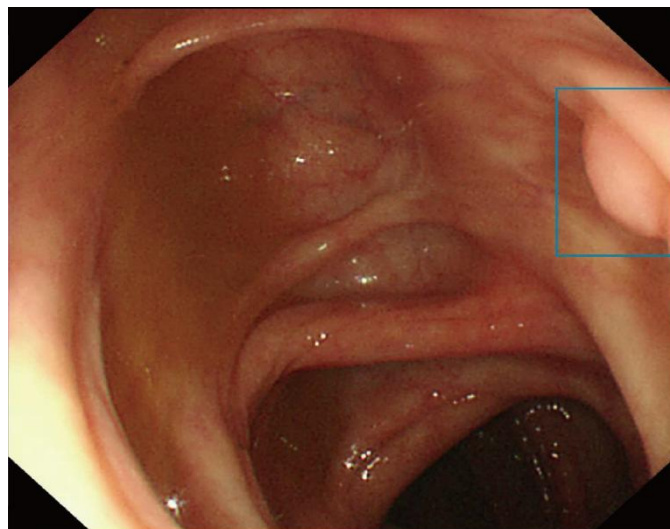
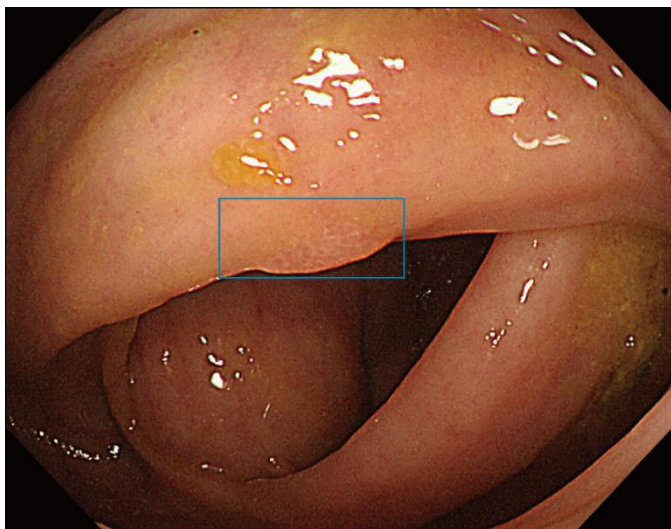


Prediction tools
Drug discovery
Precision medicine

As with most of AI, **computer vision** is fundamentally a prediction task:
*analyzing visual data (images and videos) and generating predictions about
what is contained within the image*

Computer aided polyp detection (CAdE)

The first critical *AI point solution* for GI endoscopy.



The argument for CADe is obvious, and starts with *human variability*

see related editorial on page 1666

CME

Adenomas Are Detected More Often in Morning Than in Afternoon Colonoscopy

Madhusudhan R. Sanaka, MD¹, Fnu Deepinder, MD¹, Prashanthi N. Thota, MD¹, Rocio Lopez, MS² and Carol A. Burke, MD¹

ORIGINAL ARTICLE

Adenoma Detection Rate and Risk of Colorectal Cancer and Death

American Journal of Gastroenterology
© 2008 by Am. Coll. of Gastroenterology
Published by Blackwell Publishing

ISSN 0002-9270
doi: 10.1111/j.1572-0241.2008.02085.x

Fellow Involvement May Increase Adenoma Detection Rates During Colonoscopy

Jason N. Rogart, M.D., Uzma D. Siddiqui, M.D., Priya A. Jamidar, M.D., and Harry R. Aslanian, M.D.
Department of Medicine, Section of Digestive Diseases, Yale University School of Medicine, Yale University, New Haven, Connecticut

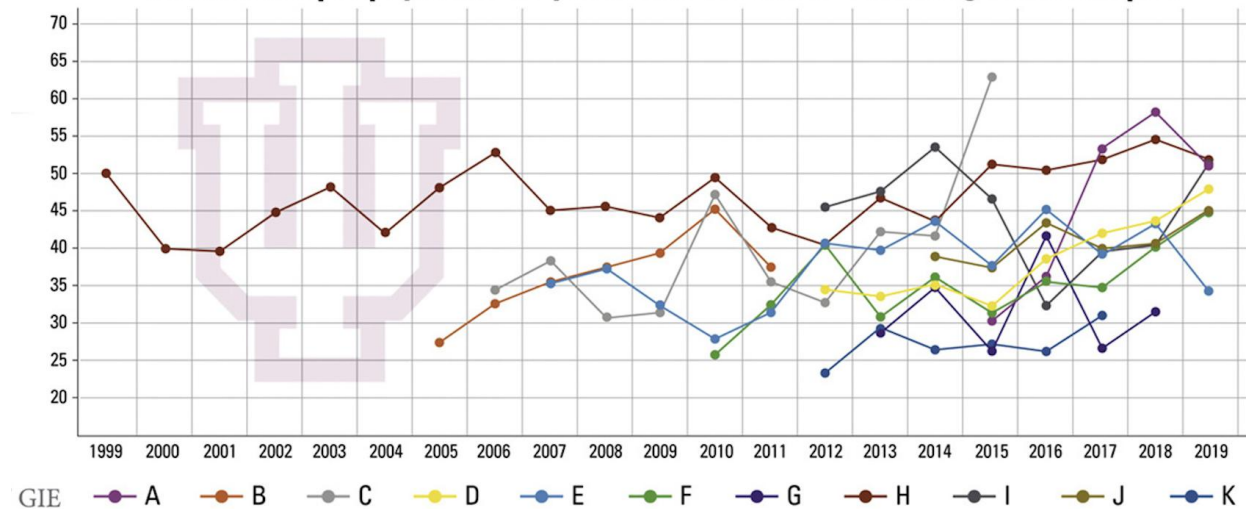
ENDOSCOPY

Fewer Polyps Detected by Colonoscopy as the Day Progresses at a Veteran's Administration Teaching Hospital

MICHAEL Y. CHAN,^{*} HARTLEY COHEN,[‡] and BRENNAN M. R. SPIEGEL^{*,†,§,||}

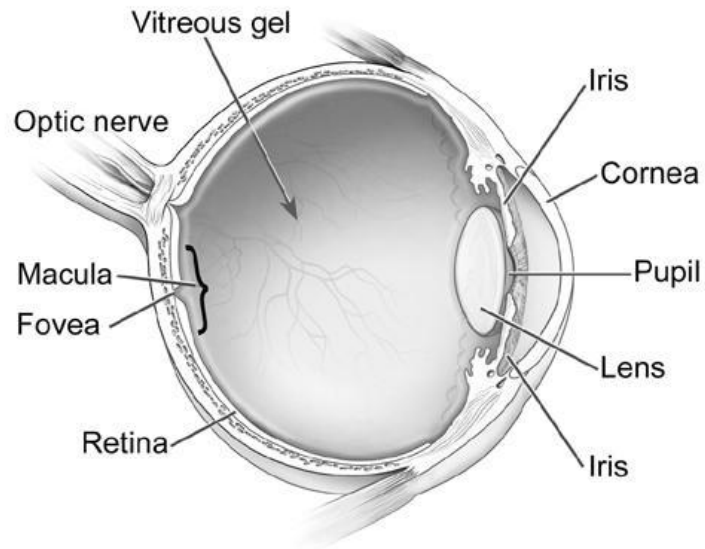
[†]Department of Gastroenterology, VA Greater Los Angeles Healthcare System; [‡]Division of Digestive Diseases, David Geffen School of Medicine at UCLA; [§]Department of Health Services, UCLA School of Public Health; ^{||}CURE Digestive Diseases Research Center; and ^{*}UCLA Center for Outcomes Research and Education, Los Angeles, California

Annual ADR per physician for years with at least 50 screening colonoscopies

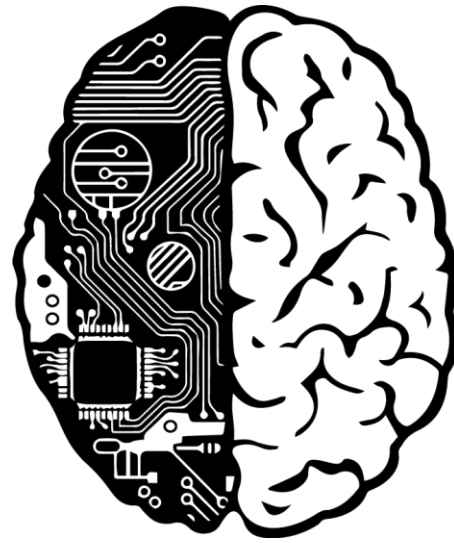


El Rayhel et al GIE 2022

Several factors contribute to variability in *on-screen* polyp detection



Visual field

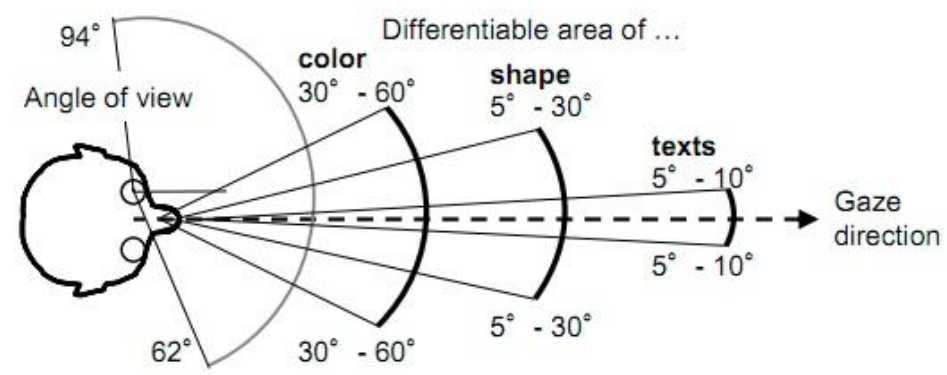


Processing power

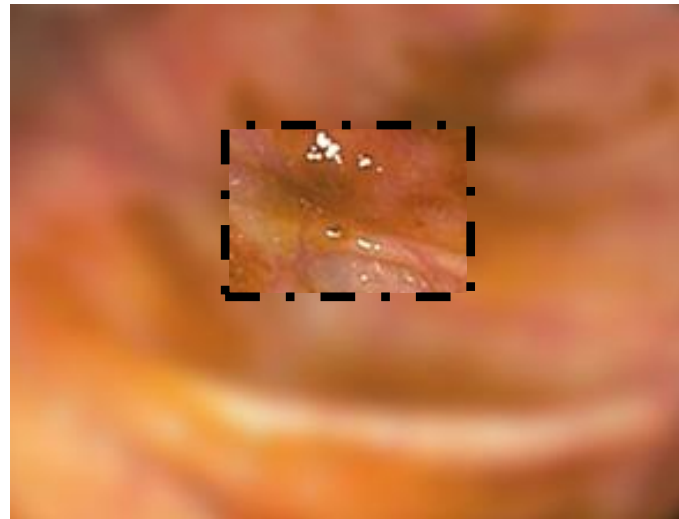


Distraction/fatigue

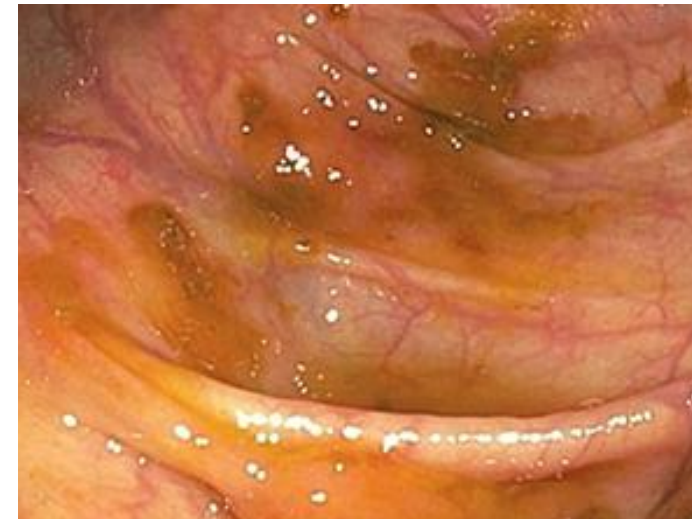
The human visual field is *not purpose-built* for *GI endoscopy*



Endoscopist scans *image*



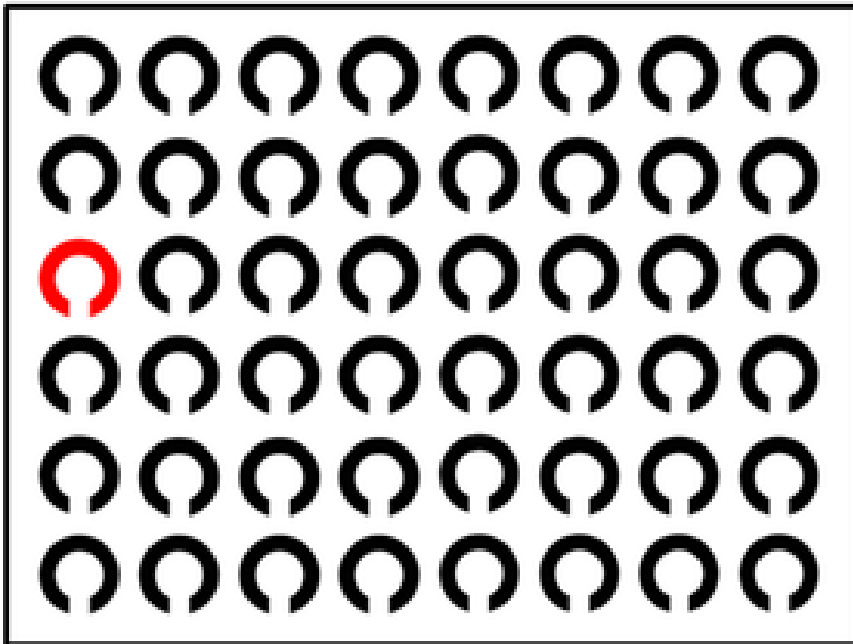
AI 'sees' each pixel equally



Visual search (finding object of interest) operates in two primary modes

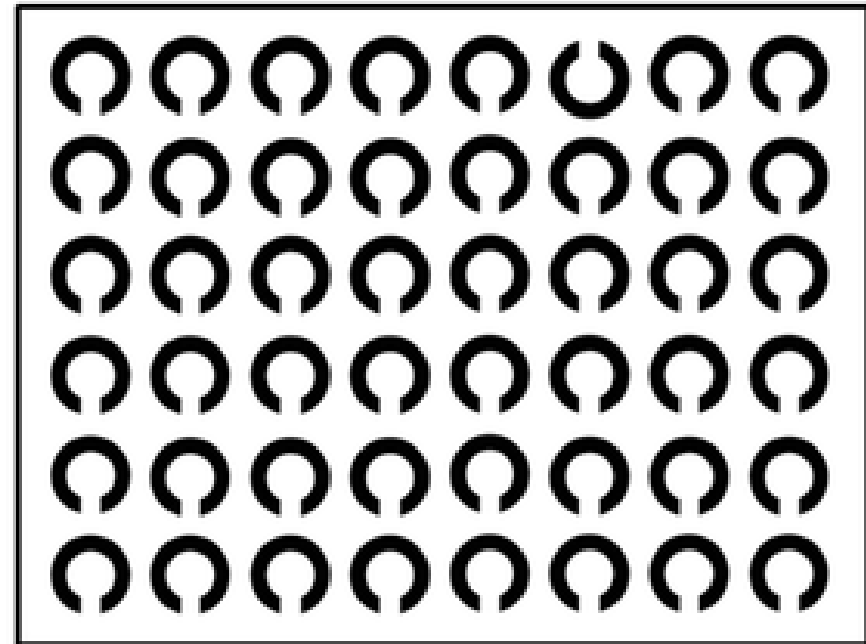
Parallel search task

Find: 



Serial search task

Find: 



Visual search (finding object of interest) operates in two primary modes

Parallel search task

Find: Apple



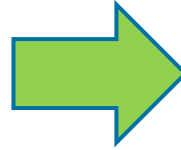
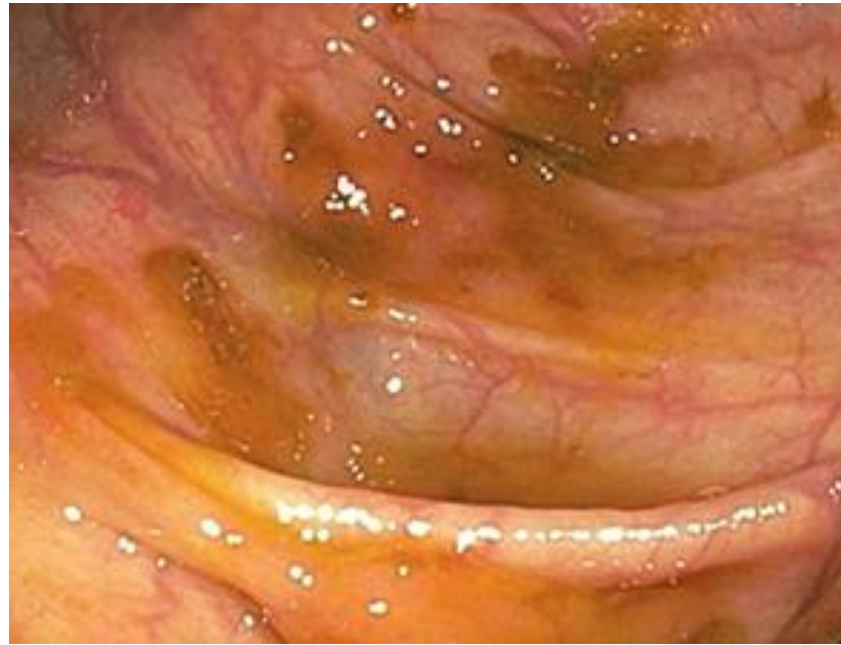
Serial search task

Find: *Pride and Prejudice*

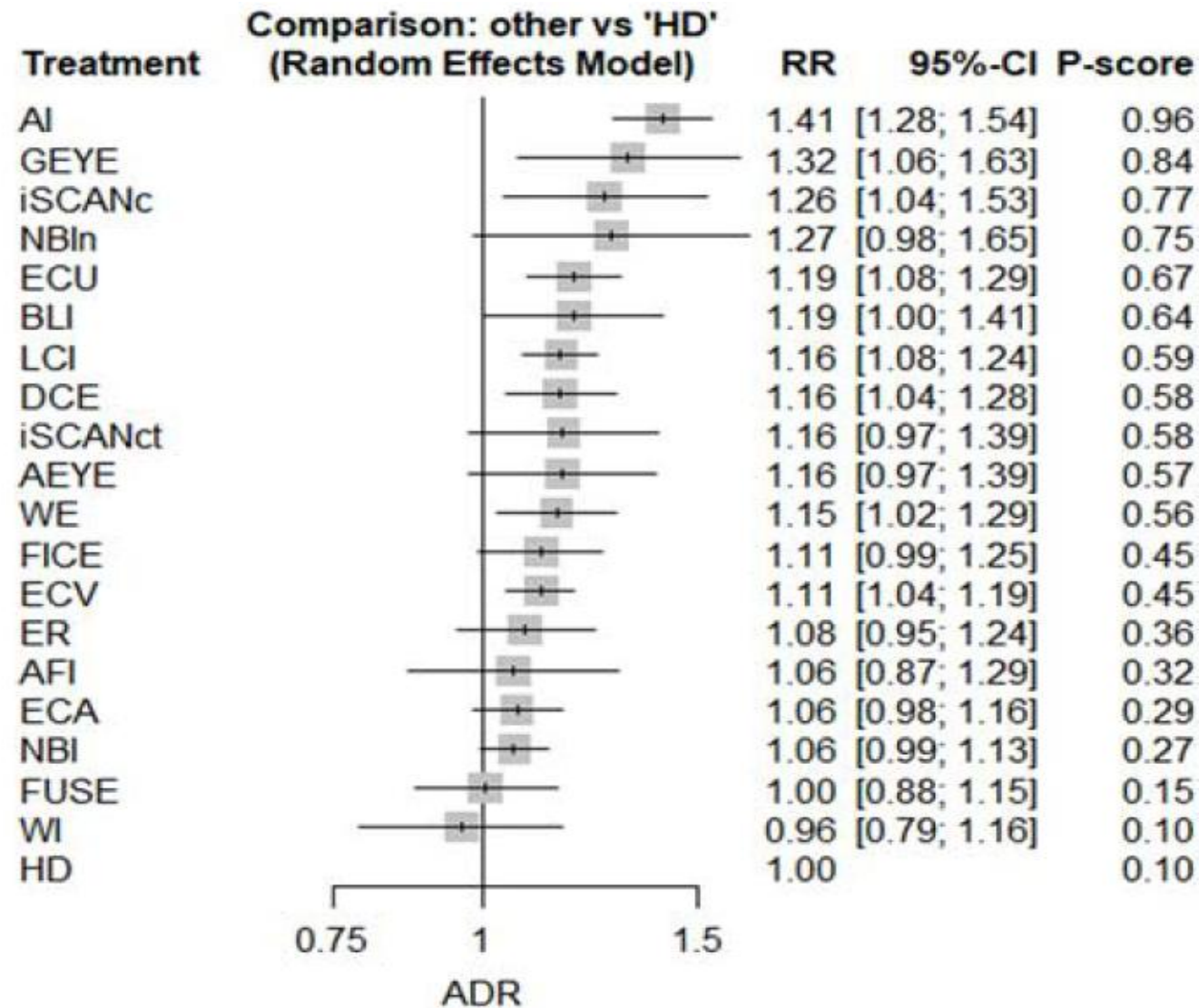


CADe makes 'polyp search' task easier

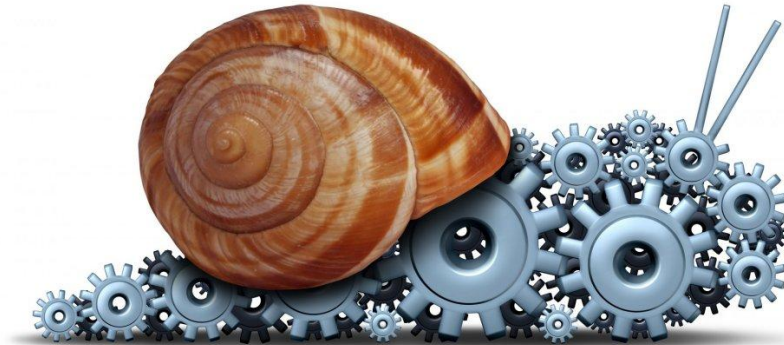
Can this (and other AI tools/automation) improve performance and reduce cognitive load?



Superiority of CAdE/Al vs *all other* polyp detection technologies in improving ADR



Despite trial data, FDA approval.. adoption of AI polyp detection has been *sluggish*



What are the barriers to adoption of AI polyp detection technology?

Table 2. Potential Reasons for Slow Early Uptake of Commercial CAdE Devices

No legal or regulatory mandate to measure ADR in many countries, regions

Endoscopists measuring ADR believe their ADR is adequate or limited benefit in increasing ADR

High cost of CAdE

Costs of CAdE are immediate, return on investment is uncertain, delayed

CAdE systems sold as add-ons

Lack of direct reimbursement for AI

Endoscopists view current systems as early in technical development

Endoscopists expect prices to drop with increased competition

Endoscopists sense little pressure to purchase from patients or local competitors

Endoscopists waiting to see what experts do

Endoscopists believe AI competes with physician expertise

Endoscopists mistrust AI development, monitoring processes

What are the barriers to adoption of AI polyp detection technology?

Table 2. Potential Reasons for Slow Early Uptake of Commercial CAdE Devices

- No legal or regulatory mandate to measure ADR in many countries, regions
- Endoscopists measuring ADR believe their ADR is adequate or limited benefit in increasing ADR
- High cost of CAdE
- Costs of CAdE are immediate, return on investment is uncertain, delayed
- CAdE systems sold as add-ons
- Lack of direct reimbursement for AI
- Endoscopists view current systems as early in technical development
- Endoscopists expect prices to drop with increased competition
- Endoscopists sense little pressure to purchase from patients or local competitors
- Endoscopists waiting to see what experts do
- Endoscopists believe AI competes with physician expertise
- Endoscopists mistrust AI development, monitoring processes



Lack of quality/reporting mandates

Economic considerations

Tech adoption curve

What are the barriers to adoption of AI polyp detection technology?

Table 2. Potential Reasons for Slow Early Uptake of Commercial CAdE Devices

- No legal or regulatory mandate to measure ADR in many countries, regions
- Endoscopists measuring ADR believe their ADR is adequate or limited benefit in increasing ADR
- High cost of CAdE
- Costs of CAdE are immediate, return on investment is uncertain, delayed
- CAdE systems sold as add-ons
- Lack of direct reimbursement for AI
- Endoscopists view current systems as early in technical development
- Endoscopists expect prices to drop with increased competition
- Endoscopists sense little pressure to purchase from patients or local competitors
- Endoscopists waiting to see what experts do
- Endoscopists believe AI competes with physician expertise
- Endoscopists mistrust AI development, monitoring processes



Lack of quality/reporting mandates

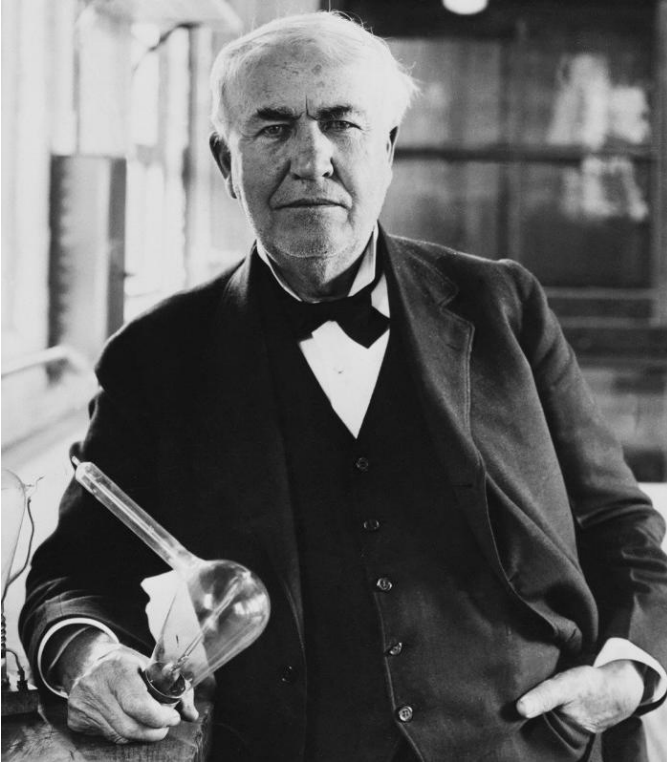
Economic considerations

Tech adoption curve

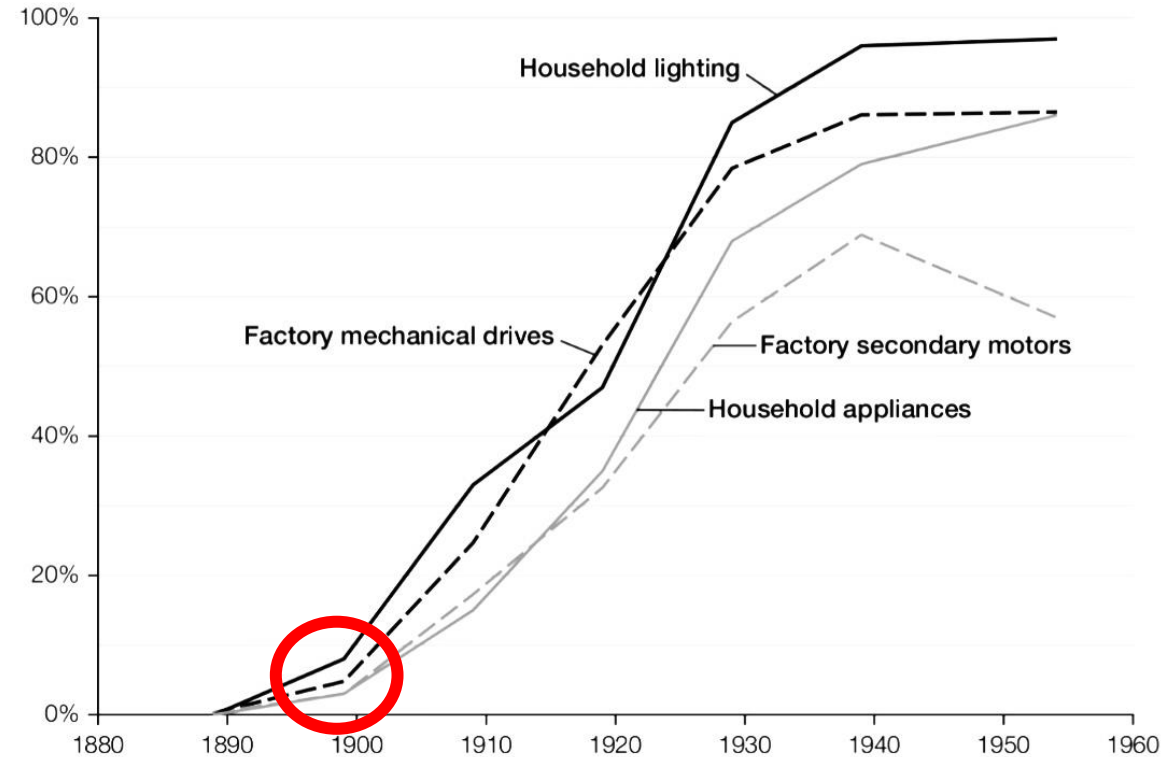
Polyp detection is only a point solution

The lights didn't go on in 1879.

Edison's electric light bulb, 1879

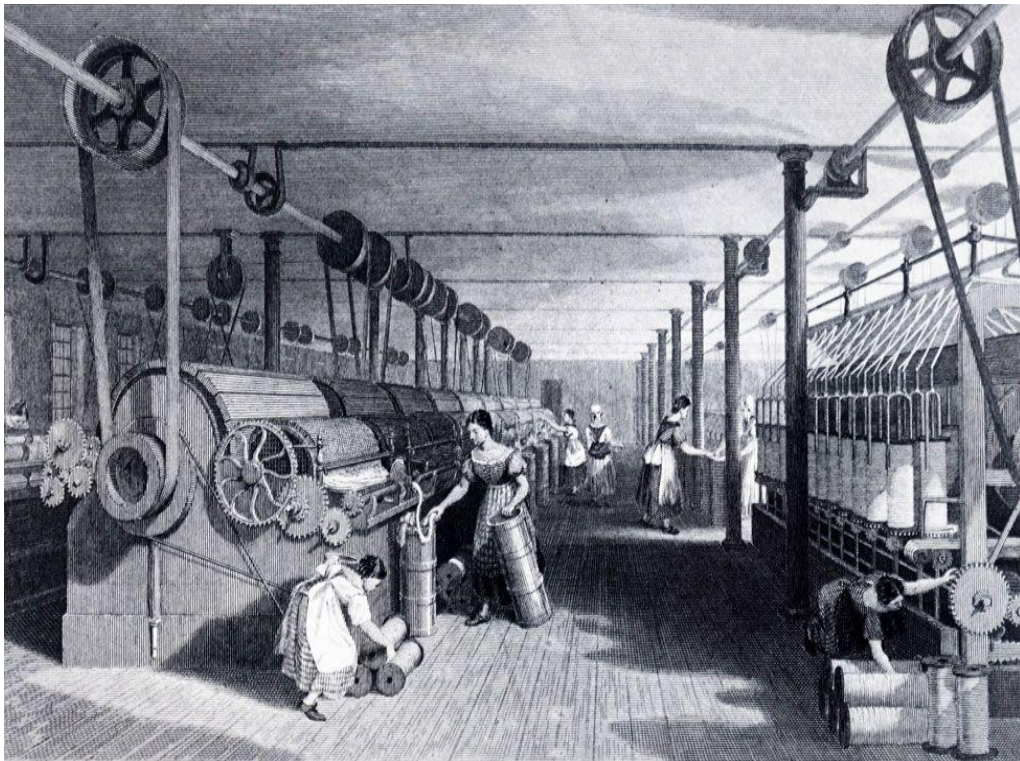


20 years later..... adoption of electricity at ~3%



The '*in between times*': after the demonstration of promise and before transformational impact. Revolutionary technologies must make a critical leap between offering *point solutions* to enabling new *system solutions*.

Electricity initially just 'plugged in' to existing steam factor design = rows of similar machines



Inflection point occurred later, when electricity enables us to rethink factory design: *the assembly line*



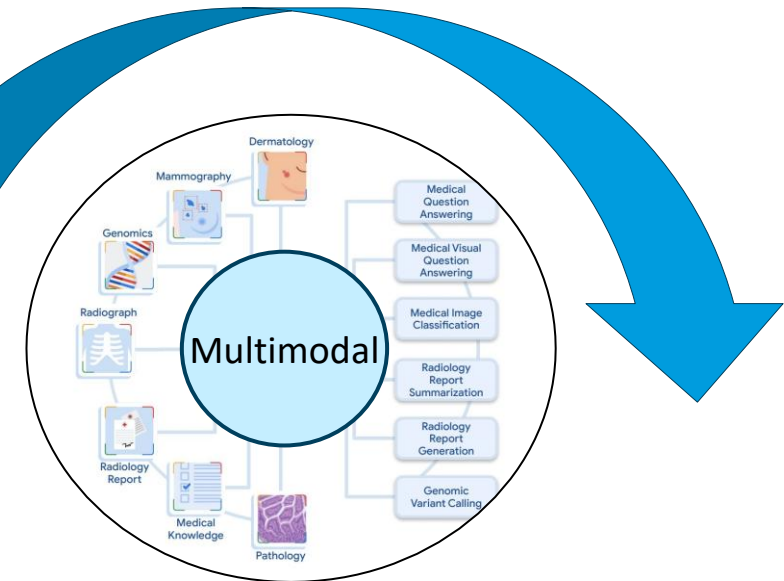
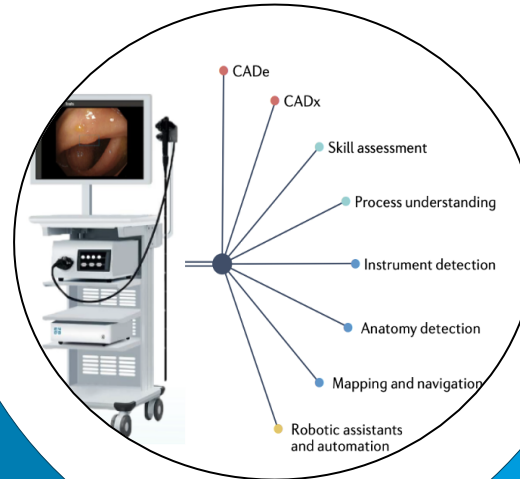
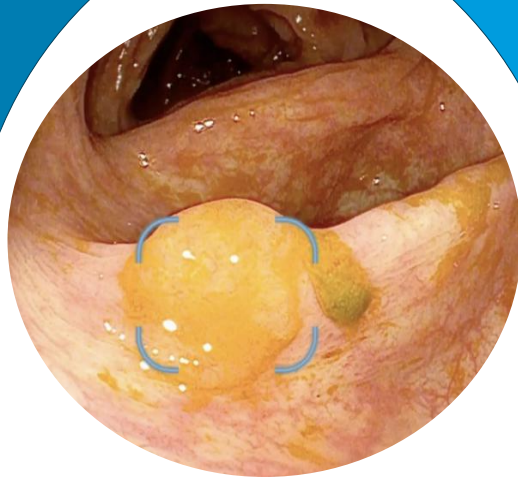
A roadmap for AI in GI endoscopy

AI point solutions

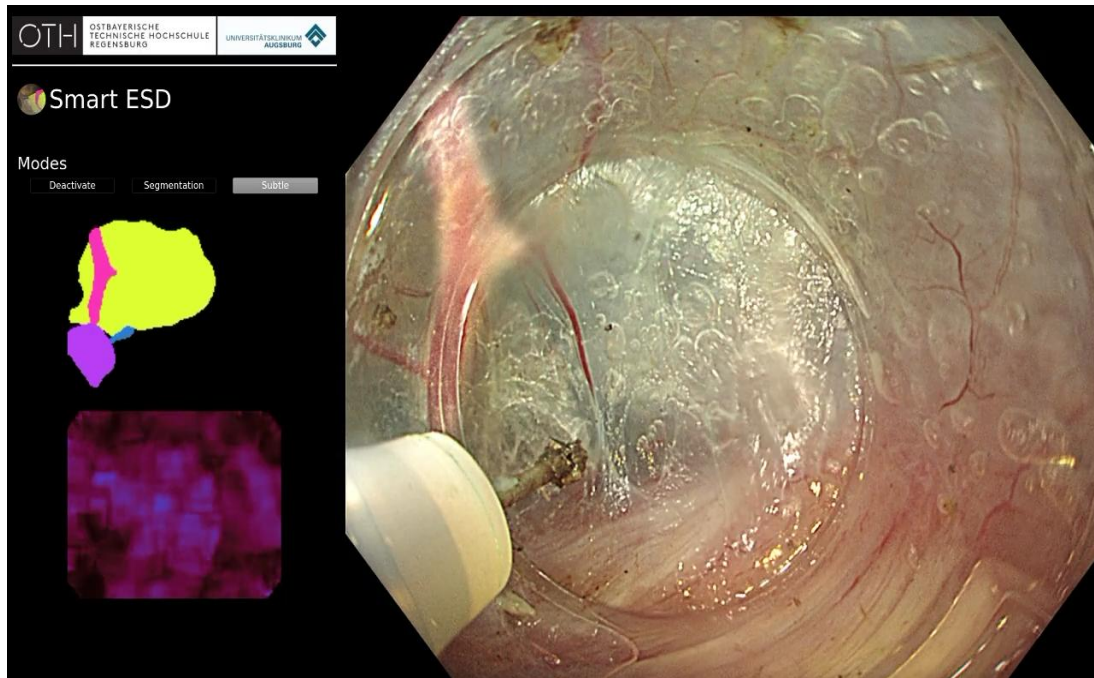
AI-integrated platforms

'Generalist' foundation models

YOU ARE
HERE



First, you will see more ‘point solutions’ in GI endoscopy:

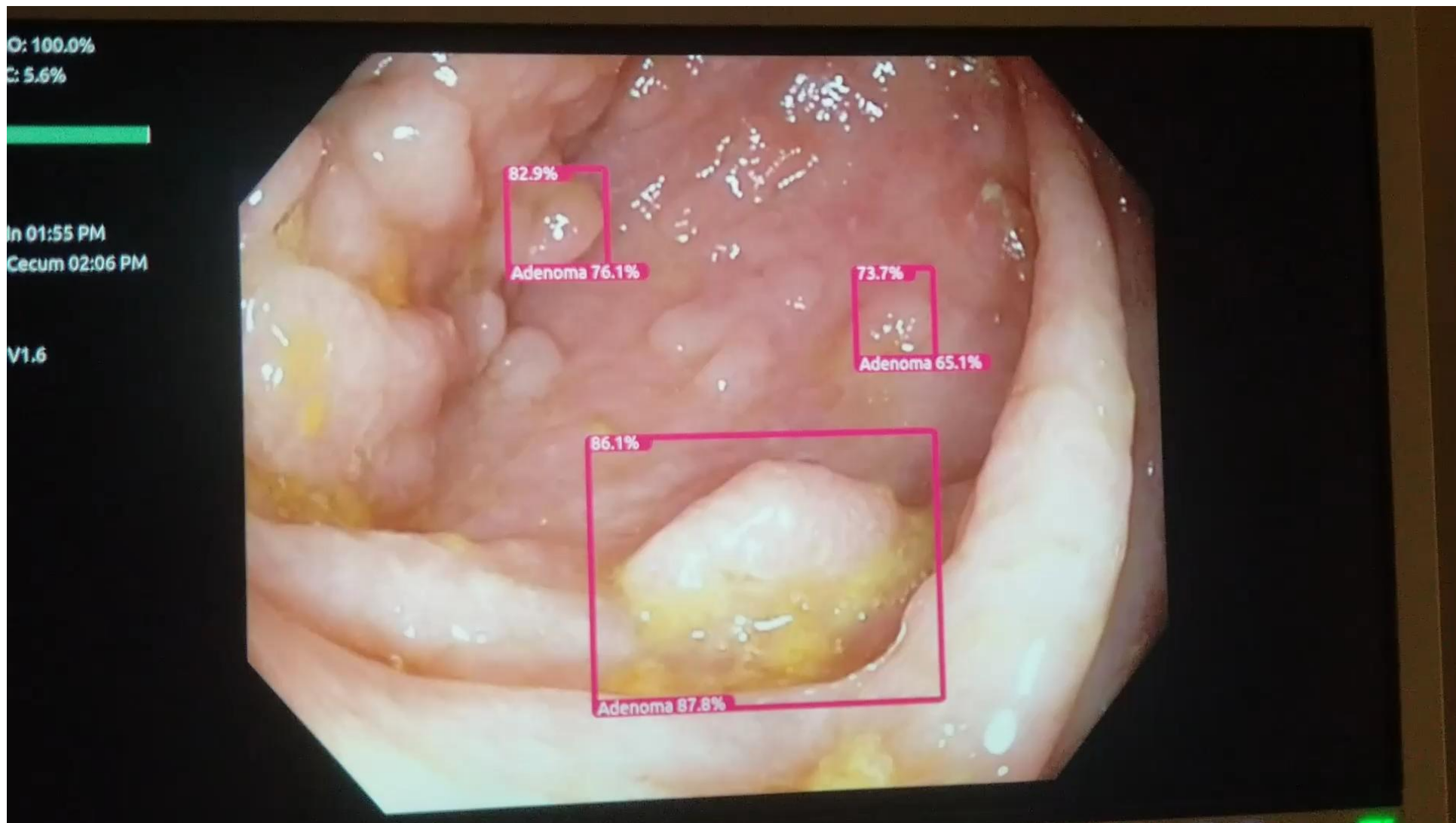


(Source: Ebigbo et al *Gut* 2022)



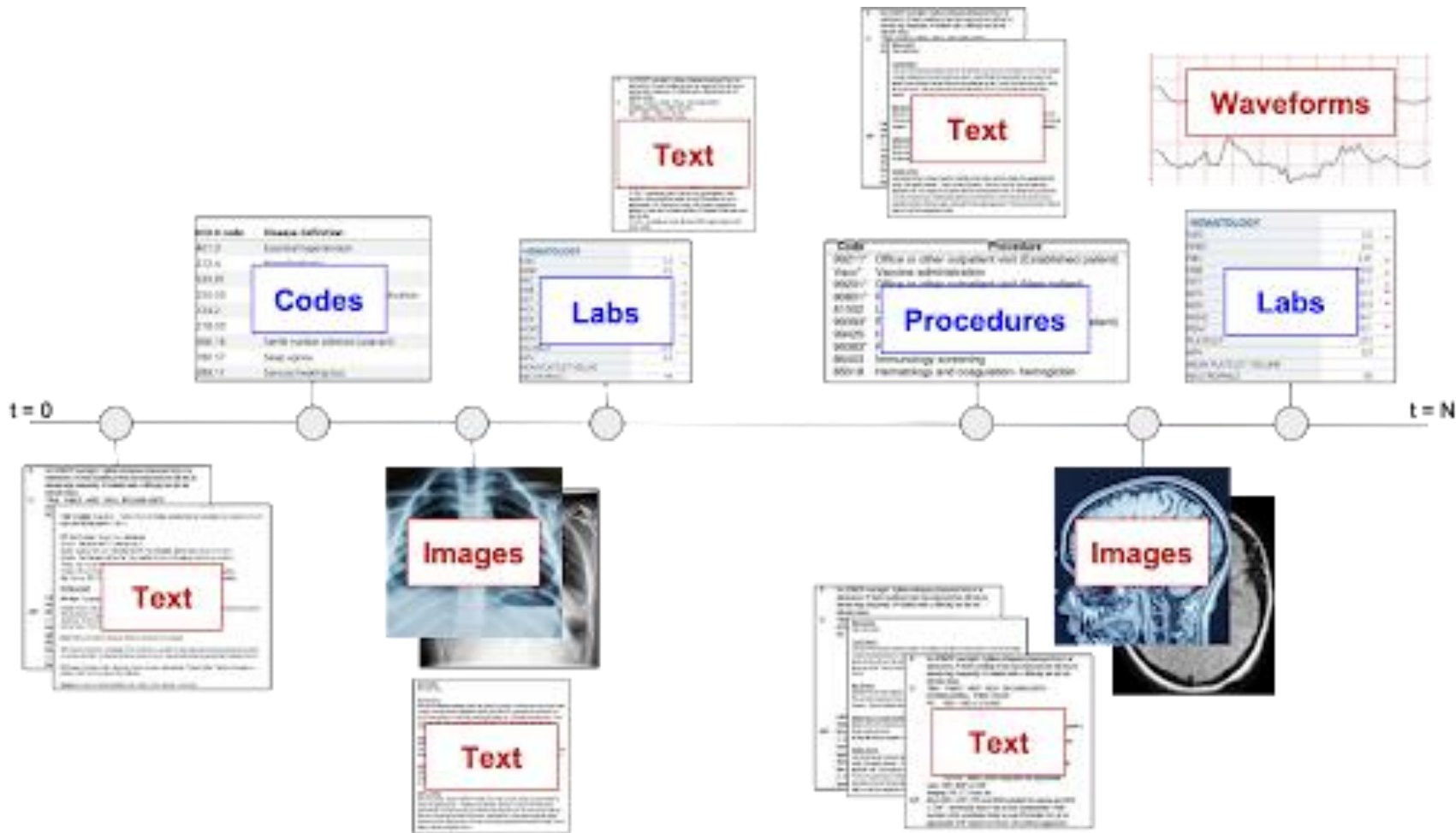
(Source: Dr. Carlos Robles-Medranda and AI Works)

Then, you will see broader, more integrated functionality.....



Source: Docbot AI

The **next AI frontier for GI** will be focused on addressing a *fundamental limitation* of current AI tools: they are “**trapped**” in a single-modality (e.g. video/images), while health data is inherently *multi-modal*



Our current ‘task-specific’ paradigm may be disrupted by *foundation models* → *generalist medical AI* (e.g. ‘a medical copilot’)

Generalist Medical AI (GMAI)

1. A GMAI model can be adapted to a new task by *describing the task in plain English*
2. GMAI models can accept inputs and produce outputs using varying combinations of data modalities
3. GMAI models can use medically accurate language *to explain their outputs*

Perspective

Foundation models for generalist medical artificial intelligence

<https://doi.org/10.1038/s41586-023-05881-4>
Received: 3 November 2022
Accepted: 22 February 2023
Published online: 12 April 2023
Check for updates

Michael Moor^{1,2}, Oishi Banerjee^{3,4}, Zahra Shakeri Hossein Abadi⁵, Harlan M. Krumholz⁶, Jure Leskovec⁷, Eric J. Topol^{8,9,10} & Pranav Rajpurkar^{2,7,11}

The exceptionally rapid development of highly flexible, reusable artificial intelligence (AI) models is likely to usher in newfound capabilities in medicine. We propose a new paradigm for medical AI, which we refer to as generalist medical AI (GMAI). GMAI models will be capable of carrying out a diverse set of tasks using very little or no task-specific labelled data. Built through self-supervision on large, diverse datasets, GMAI will flexibly interpret different combinations of medical modalities, including data from imaging, electronic health records, laboratory results, genomics, graphs or medical text. Models will in turn produce expressive outputs such as free-text explanations, spoken recommendations or image annotations that demonstrate advanced medical reasoning abilities. Here we identify a set of high-impact potential applications for GMAI and lay out specific technical capabilities and training datasets necessary to enable them. We expect that GMAI-enabled applications will challenge current strategies for regulating and validating AI devices for medicine and will shift practices associated with the collection of large medical datasets.

Foundation models—the latest generation of AI models—are trained on massive, diverse datasets and can be applied to numerous downstream tasks¹. Individual models can now achieve state-of-the-art performance on a wide variety of problems, ranging from answering questions about texts to describing images and playing video games^{2–4}. This versatility represents a stark change from the previous generation of AI models, which were designed to solve specific tasks, one at a time.

Driven by growing datasets, increases in model size and advances in model architectures, foundation models offer previously unseen abilities. For example, in 2020 the language model GPT-3 unlocked a new capability: in-context learning, through which the model carried out entirely new tasks that it had never explicitly been trained for, simply by learning from text explanations (or ‘prompts’) containing a few examples⁵. Additionally, many recent foundation models are able to take in and output combinations of different data modalities^{6–8}. For example, the recent Gato model can chat, caption images, play video games and control a robot arm and has thus been described as a generalist agent⁹. As certain capabilities emerge only in the largest models, it remains challenging to predict what even larger models will be able to accomplish¹⁰.

Although there have been early efforts to develop medical foundation models^{11–13}, this shift has not yet widely permeated medical AI, owing to the difficulty of accessing large, diverse medical datasets, the complexity of the medical domain and the recency of this development. Instead, medical AI models are largely still developed with a task-specific approach to model development. For instance, a chest X-ray interpretation model may be trained on a dataset in which every image has been explicitly labelled as positive or negative for pneumonia, probably requiring substantial annotation effort. This model

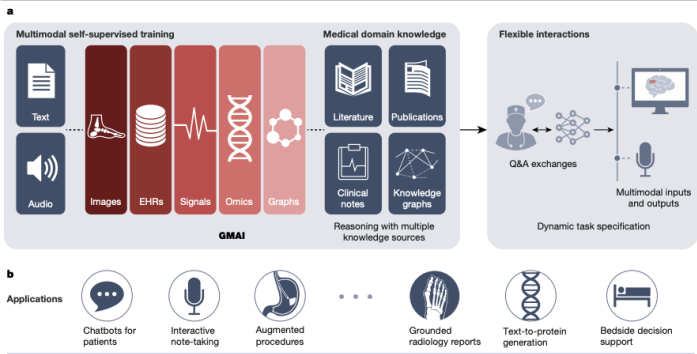
would only detect pneumonia and would not be able to carry out the complete diagnostic exercise of writing a comprehensive radiology report. This narrow, task-specific approach produces inflexible models, limited to carrying out tasks predefined by the training dataset and its labels. In current practice, such models typically cannot adapt to other tasks (or even to different data distributions for the same task) without being retrained on another dataset. Of the more than 500 AI models for clinical medicine that have received approval by the Food and Drug Administration, most have been approved for only 1 or 2 narrow tasks¹⁴.

Here we outline how recent advances in foundation model research can disrupt this task-specific paradigm. These include the rise of multimodal architectures¹⁵ and self-supervised learning techniques¹⁶ that dispense with explicit labels (for example, language modelling¹⁷ and contrastive learning¹⁸), as well as the advent of in-context learning capabilities¹⁹.

These advances will instead enable the development of GMAI, a class of advanced medical foundation models. ‘Generalist’ implies that they will be widely used across medical applications, largely replacing task-specific models.

Inspired directly by foundation models outside medicine, we identify three key capabilities that distinguish GMAI models from conventional medical AI models (Fig. 1). First, adapting a GMAI model to a new task will be as easy as describing the task in plain English (or another language). Models will be able to solve previously unseen problems simply by having new tasks explained to them (dynamic task specification), without needing to be retrained²⁰. Second, GMAI models can accept inputs and produce outputs using varying combinations of data modalities (for example, can take in images, text, laboratory results or

Perspective



Regulations: Application approval; validation; audits; community-based challenges; analyses of biases, fairness and diversity

Fig. 1 | Overview of a GMAI model pipeline. **a**, A GMAI model is trained on multiple medical data modalities, through techniques such as self-supervised learning. To enable flexible interactions, data modalities such as images or data from EHRs can be paired with language, either in the form of text or speech data. Next, the GMAI model needs to access various sources of medical knowledge to carry out medical reasoning tasks, unlocking a wealth of capabilities that can be used in downstream applications. The resulting GMAI model then carries

out tasks that the user can specify in real time. For this, the GMAI model can retrieve contextual information from sources such as knowledge graphs or databases, leveraging formal medical knowledge to reason about previously unseen tasks. **b**, The GMAI model builds the foundation for numerous applications across clinical disciplines, each requiring careful validation and regulatory assessment.

any combination thereof). This flexible interactivity contrasts with the constraints of more rigid multimodal models, which always use predefined sets of modalities as input and output (for example, must always take in images, text and laboratory results together). Third, GMAI models will formally represent medical knowledge, allowing them to reason through previously unseen tasks and use medically accurate language to explain their outputs.

We list concrete strategies for achieving this paradigm shift in medical AI. Furthermore, we describe a set of potentially high-impact applications that this new generation of models will enable. Finally, we point out core challenges that must be overcome for GMAI to deliver the clinical value it promises.

The potential of generalist models in medical AI

GMAI models promise to solve more diverse and challenging tasks than current medical AI models, even while requiring little to no labels for specific tasks. Of the three defining capabilities of GMAI, two enable flexible interactions between the GMAI model and the user: first, the ability to carry out tasks that are dynamically specified; and second, the ability to support flexible combinations of data modalities. The third capability requires that GMAI models formally represent medical domain knowledge and leverage it to carry out advanced medical reasoning. Recent foundation models already exhibit individual aspects of GMAI, by flexibly combining several modalities²¹ or making it possible to dynamically specify a new task at test time²², but substantial advances are still required to build a GMAI model with all three capabilities. For example, existing models that show medical reasoning abilities (such as GPT-3 or PaLM) are not multimodal and do not yet generate reliable factual statements.

Flexible interactions

GMAI offers users the ability to interact with models through custom queries, making AI insights easier for different audiences to understand and offering unprecedented flexibility across tasks and settings. In current practice, AI models typically handle a narrow set of tasks and produce a rigid, predetermined set of outputs. For example, a current model might detect a specific disease, taking in one kind of image and always outputting the likelihood of that disease. By contrast, a custom query allows users to come up with questions on the fly: ‘Explain the mass appearing on this head MRI scan. Is it more likely a tumour or an abscess?’. Furthermore, queries can allow users to customize the format of their outputs: ‘This is a follow-up MRI scan of a patient with glioblastoma. Outline any tumours in red’.

Custom queries will enable two key capabilities—dynamic task specification and multimodal inputs and outputs—as follows.

Dynamic task specification. Custom queries can teach AI models to solve new problems on the fly, dynamically specifying new tasks without requiring models to be retrained. For example, GMAI can answer highly specific, previously unseen questions: ‘Given this ultrasound, how thick is the gallbladder wall in millimetres?’. Unsurprisingly, a GMAI model may struggle to complete new tasks that involve unknown concepts or pathologies. In-context learning then allows users to teach the GMAI about a new concept with few examples: ‘Here are the medical histories of ten previous patients with an emerging disease, an infection with the Langya henipavirus. How likely is it that our current patient is also infected with Langya henipavirus?’.

(foundation model definition: any model that is trained on broad data, generally using self-supervision at scale, that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks)

Generalist Medical AI (GMAI) as GI endoscopy co-pilot

Contextual data in EHR including history of endovascular aortic aneurysm repair

Clinical Summary
Noble, James

Demographics

Address: 4516 West Huron Street
Chicago, IL 60607
Email: James.Noble@gmail.com
DOB: 04.06.1962
Age: 50y
Patient ID: 88501

Primary Ins: Aetna U.S. Healthcare - Master
Pharmacy Directions: CVS Pharmacy #6671
510 College Mall Rd
Bloomington, IN 47401 P (812) 336-7306
F (812) 335-9347

User Defined Fields

Name	Value
Injury	Shoulder pain
Employer Name	VBI
Employer Fax	201-967-5309
Insurance Carrier	GBCO
Adjuster Fax Num...	201-555-8735
Diagnosis	727.3
Position	Truck Driver
CareTracker ID #	36817221

Diagnosis

Diagnosis	Status	Date	ICD-9	Notes
ACL Tear		06.20.2012	844.2	
Lumbago		11.17.2011	724.2	

Vitals

Smoking Status

Status	ID	Date
Never smoker	4	06.18.2012

Procedures

Procedure	Date	CPT	Type	Status
Knee Arthroscopy/Surg	06.20.2012	29881	Internal	Performed

Rx History

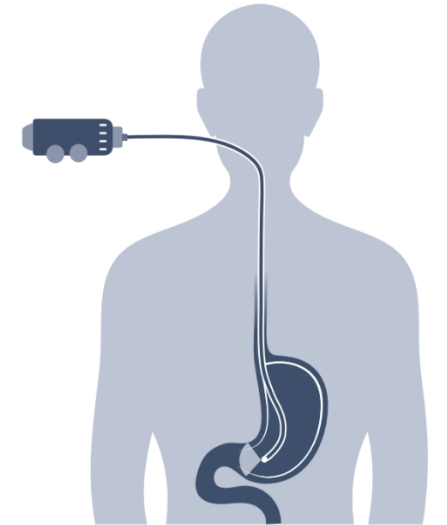
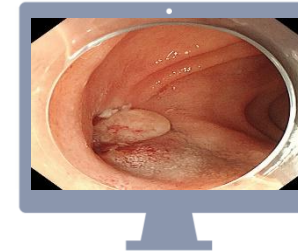
Status	Date	Drug	Strength	Instructions
✓	06.18.2012	Keflex	250 MG	1 tab po BID
✓	11.04.2010	CeleBREX	100 MG	1 tablet by mouth in the AM
✓	04.13.2006	OxyCODONE HCl	20 MG	1 tab po q4h prn pain

Non-Drug Allergies

Description	Reaction	Notes
Latex	Severe rash	
Shellfish	Hives	



Explain the object appearing on the left on the screen.



The object resembles an artery wall, situated close proximity to the duodenum. Given size and location, it's most likely the aorta (aortoduodenal fistula).



Generalist Medical AI (GMAI) as GI endoscopy co-pilot

Contextual data in EHR including history of endovascular aortic aneurysm repair

In augmented procedures, a rare outlier finding is explained with step-by-step reasoning by leveraging medical domain knowledge

Clinical Summary
Noble, James

Demographics

Address: 4516 West Huron Street
Chicago, IL 60607
Email: jnoble@icloud.com
DOB: 04.05.1962
Age: 55y
Patient ID: 59501

Primary Ins: Aetna U.S. Healthcare - Master
Pharmacy Directions: CVS Pharmacy #66271
510 College Mall Rd
Bloomington, IN 47401 P (812) 336-7306
F (812) 335-9347

User Defined Fields

Name	Value
History	Shoulder pain
Employer Name	WBI
Employer Fax	201-567-5309
Insurance Carrier	GEICO
Adjuster Fax Num	201-555-8735
Diagnosis	727.3
Position	Truck Driver
GenTracker ID #	36817221

Diagnoses

Diagnosis: ACL Tear
Lumbago

Vitals

Status: Never smoked

Procedures

73000 | 73000

Re History

Status	Date	Drug	Strength	Instructions
Active	06.18.2012	Keflex	250 MG	1 tab po BID
Discontinued	11.04.2010	CeleBREX	100 MG	1 tablet by mouth in the AM
Discontinued	04.13.2006	OxyCODONE HCl	20 MG	1 tab po q4h prn pain

Non Drug Allergies

Description	Reaction	Notes
Latices	Severe rash	
Shellfish	Hives	

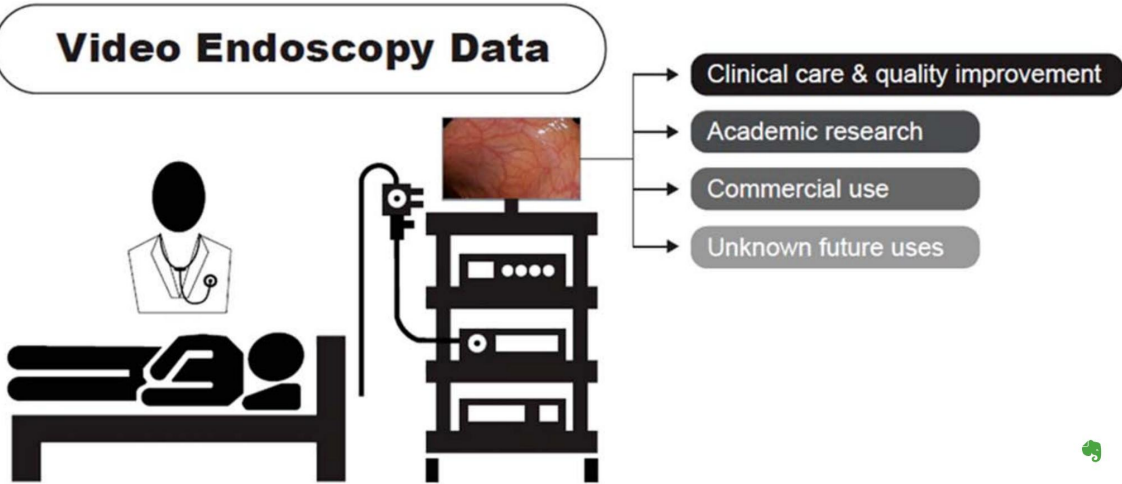
Generalist Medical AI: multimodal & broadly-capable, but:
→ potentially less predictable
→ *much harder to evaluate/study/regulate*

Explain the object appearing on the left on the screen.

The object resembles an artery wall, situated close proximity to the duodenum. Given size and location, it's most likely the aorta (aortoduodenal fistula).

What ethical and legal issues should we anticipate as AI use accelerates in GI?

Data ownership and privacy



Automation and liability




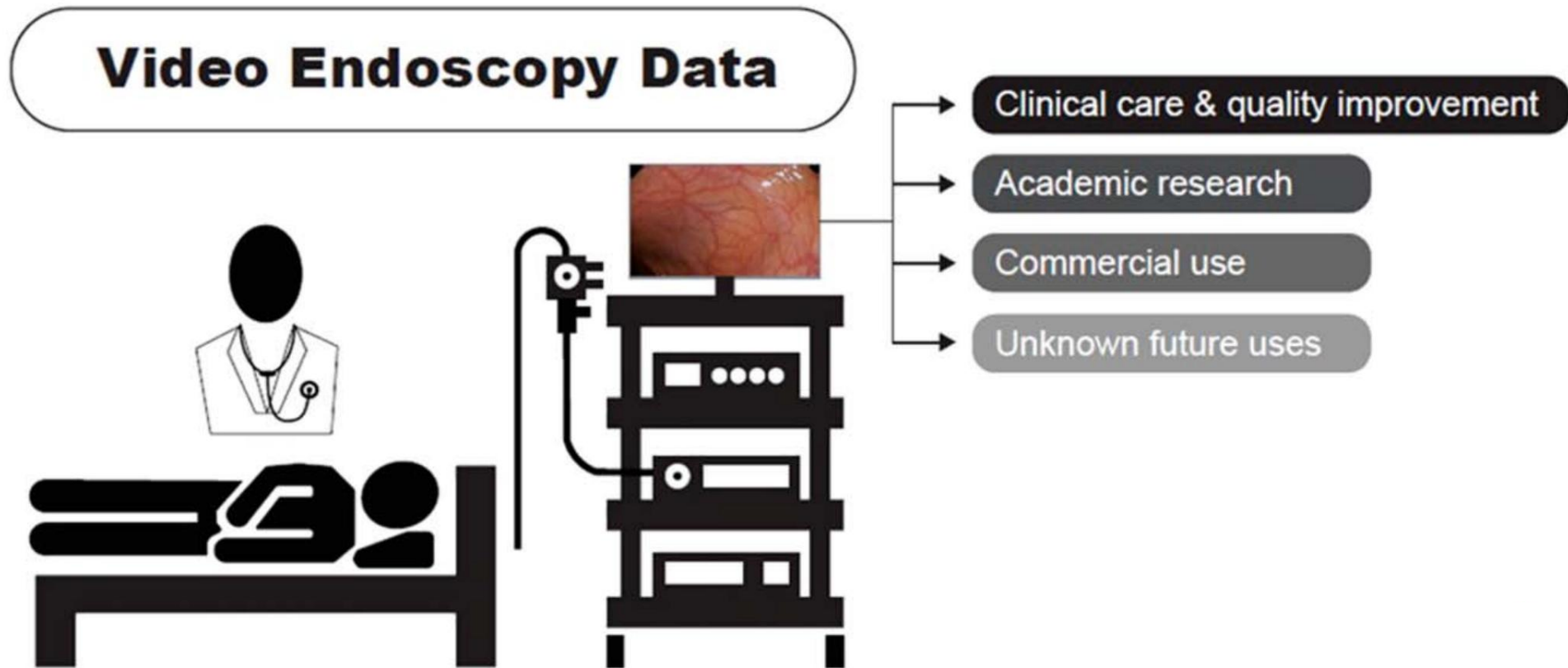
Technology	Role/output	Clinical purpose	Automation level and liability considerations
Colon Polyp CAdE 	Polyp detection.	Colon polyp CAdE is intended to support endoscopists by improving polyp detection and reducing polyp miss rate.	<ul style="list-style-type: none">Automation Level 1Minimal impact on endoscopist liability because this is a Level 1 algorithm that provides additional information that can be evaluated by the endoscopist using their independent medical expertise. If CAdE becomes standard of practice, then failure to use CAdE may introduce liability.
Colon Polyp CADx 	Polyp diagnosis prediction (e.g., "adenoma" vs "hyperplastic").	CADx use may support adoption of "resect and discard" or "diagnose and leave" strategies, reducing use of traditional histopathology.	<ul style="list-style-type: none">Automation Level 2Impact on liability depends on the evolution in standards of care around use of CADx and impact on polypectomy strategies. Physicians must ensure that algorithms are used in accordance with manufacturer's instructions and indications for use.
Automated Capsule Endoscopy Reading 	Range of roles from detecting or highlighting abnormalities for physician reader to fully autonomous completion of report.	Automated capsule reading algorithms are designed to improve efficiency and accuracy of image review, and may enable fully-automated interventions in the future (biopsy, etc.).	<ul style="list-style-type: none">Automation Levels 1–5Impact on liability depends on the level of automation achieved by automated capsule reading algorithms, and evolving standards of care. Physicians must ensure that algorithms are used in accordance with manufacturer's instructions and indications for use.

Figure 1. Automation levels and liability considerations for representative AI tools in GI endoscopy. (Illustration by Danielle Duffey).

AI innovation requires patient data: how do we balance *privacy* and *progress*?



Eugenia N. Uche-Anyia, MD, MPH¹, Sara Gerke, Dipl-Jur Univ, MA² and Tyler M. Berzin, MD, MS, FACG³

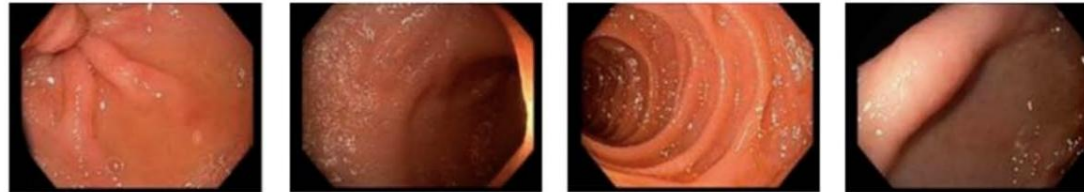
Am J Gastroenterol 2024;119:600–605. <https://doi.org/10.14309/ajg.0000000000002597>; published online November 17, 2023

Should patients have a voice in how their data is mined in this new digital +AI era?

Do we have permission to reuse video/image data from your endoscopy procedure?

We would like your permission to store the video recording from your procedure to use again in the future for education and research and to improve patient care. You do NOT have to agree to this to receive care, testing, or treatment. Your decision will NOT affect the care you receive.

*Example
endoscopy
images*



Educational use: Video/image data from my procedure may be used and shared for *teaching and educational purposes* to help other medical professionals and trainees learn about, prevent, or treat health problems.

☐ Yes

☐ No

Research use: Video/image data from my procedure may be used for *hospital/university research* focused on understanding, preventing or treating health problems.

☐ Yes

☐ No

Commercial use: Video/image data from my procedure may be shared with *commercial parties (companies)* to help develop new products for patient care that they may sell for a profit.

☐ Yes

☐ No

“HIPAA doesn’t require patient consent to de-identify data and sell it.... it is expressly permitted by HIPAA. You don’t need to even mention it to a patient.”

STAT+ | HEALTH TECH

The companies paying hospitals to hand over patient data to train AI

A new crop of businesses is taking advantage of privacy laws to buy and share health data



Adobe



By **Brittany Trang** Jan. 16, 2025
Health Tech Reporter

Artificial intelligence models are ever-hungry black boxes that need boatloads of bits and bytes from a wide stream of real-world data in order to produce insights



Tech is transforming health care and life sciences. Our original reporting is here to keep you ahead of the curve.

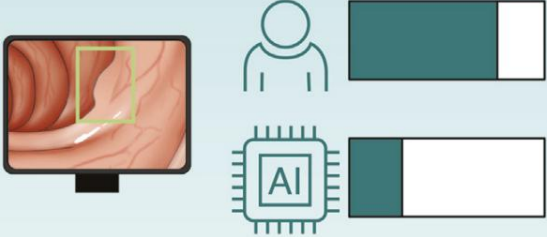
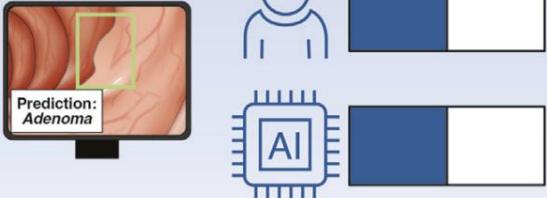
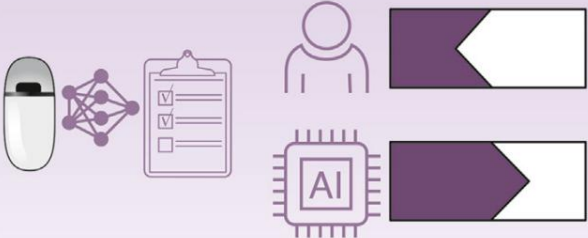
+ Sign Up

MOST POPULAR



S+ | AstraZeneca, Daiichi Sankyo win first U.S. approval for key cancer drug

How does physician behavior (& potential liability) change as AI *automation* increases?

Technology	Role/output	Clinical purpose	Automation level and liability considerations
Colon Polyp CADe 	Polyp detection.	Colon polyp CADe is intended to support endoscopists by improving polyp detection and reducing polyp miss rate.	<ul style="list-style-type: none"> Automation Level 1 Minimal impact on endoscopist liability because this is a Level 1 algorithm that provides additional information that can be evaluated by the endoscopist using their independent medical expertise. If CADe becomes standard of practice, then failure to use CADe may introduce liability.
Colon Polyp CADx 	Polyp diagnosis prediction (e.g., “adenoma” vs “hyperplastic”).	CADx use may support adoption of “resect and discard” or “diagnose and leave” strategies, reducing use of traditional histopathology.	<ul style="list-style-type: none"> Automation Level 2 Impact on liability depends on the evolution in standards of care around use of CADx and impact on polypectomy strategies. Physicians must ensure that algorithms are used in accordance with manufacturer’s instructions and indications for use.
Automated Capsule Endoscopy Reading 	Range of roles from detecting or highlighting abnormalities for physician reader to fully autonomous completion of report.	Automated capsule reading algorithms are designed to improve efficiency and accuracy of image review, and may enable fully-automated interventions in the future (biopsy, etc.).	<ul style="list-style-type: none"> Automation Levels 1–5 Impact on liability depends on the level of automation achieved by automated capsule reading algorithms, and evolving standards of care. Physicians must ensure that algorithms are used in accordance with manufacturer’s instructions and indications for use.

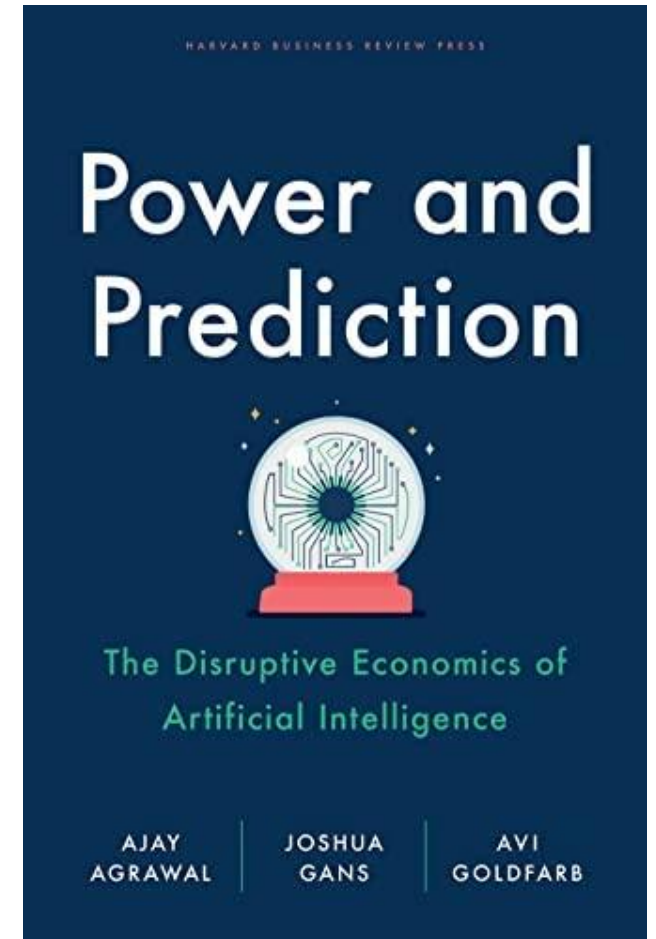
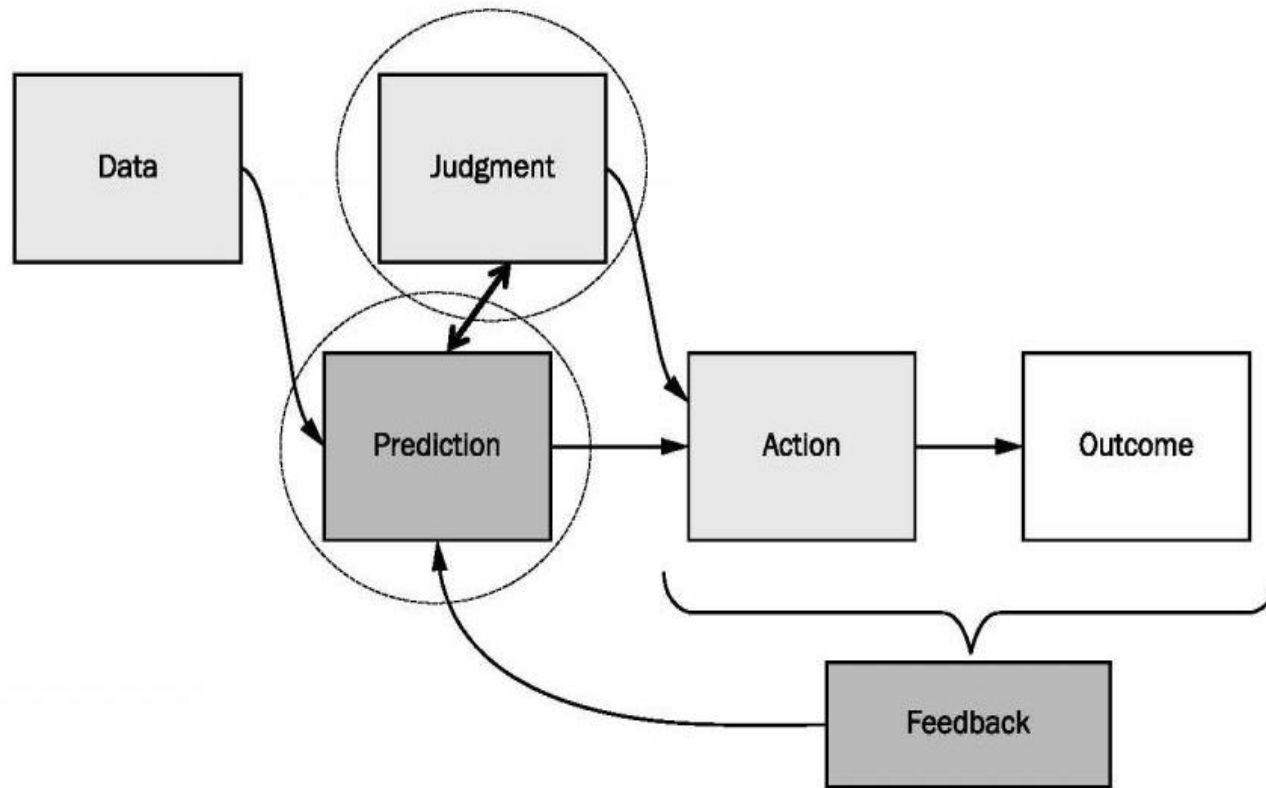
What happened to airline pilots when autopilot became the norm?

- Autopilot handled repetitive/tedious tasks and *routine* flight
- Pilot's responsibility shifts towards *higher* complexity considerations (recognizing *edge cases*, engaging common sense vs. faulty data readings)
- New model requires *higher* skill from human to understand and oversee complex interactions between *tech* and *task*
- Less focus on managing the *stabilizer* and more on the managing the *system*

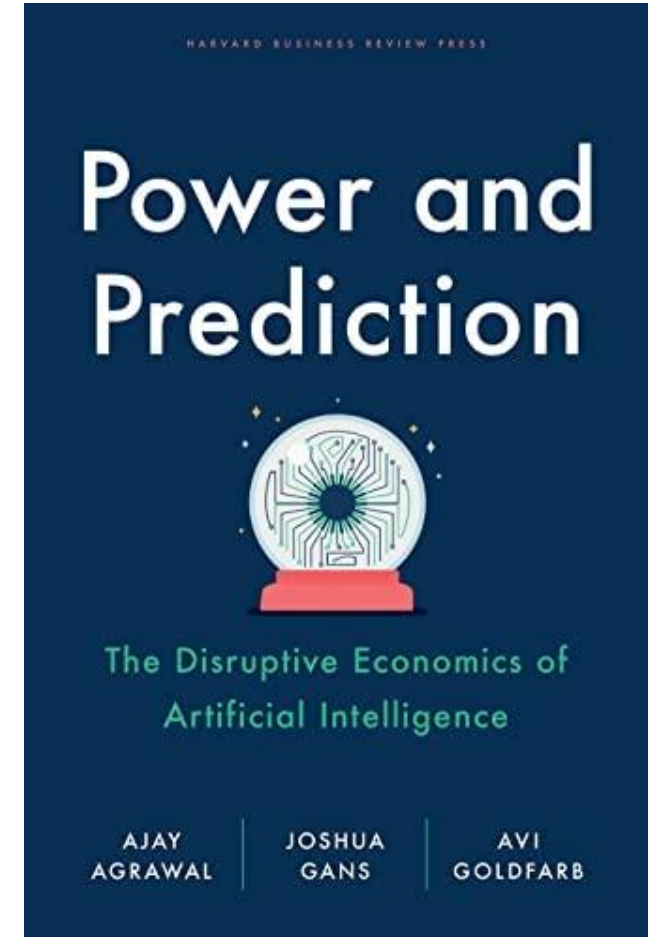
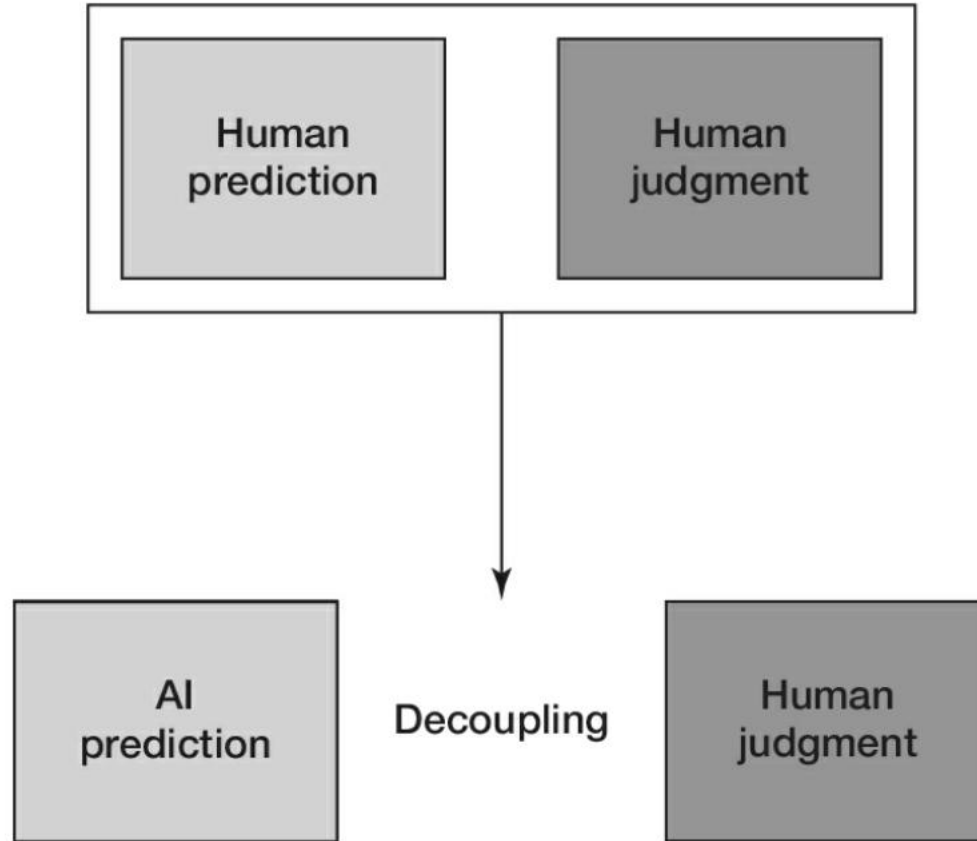


How to think about the role of AI in clinical care: *predictions vs. judgment*

Anatomy of a task (e.g. a medical decision)



AI will provide *predictions* more accurately and faster than humans.
In medicine, *prediction* and *judgment* will be decoupled.



The 'jagged frontier' of AI capabilities

The ‘jagged frontier’ of AI capabilities

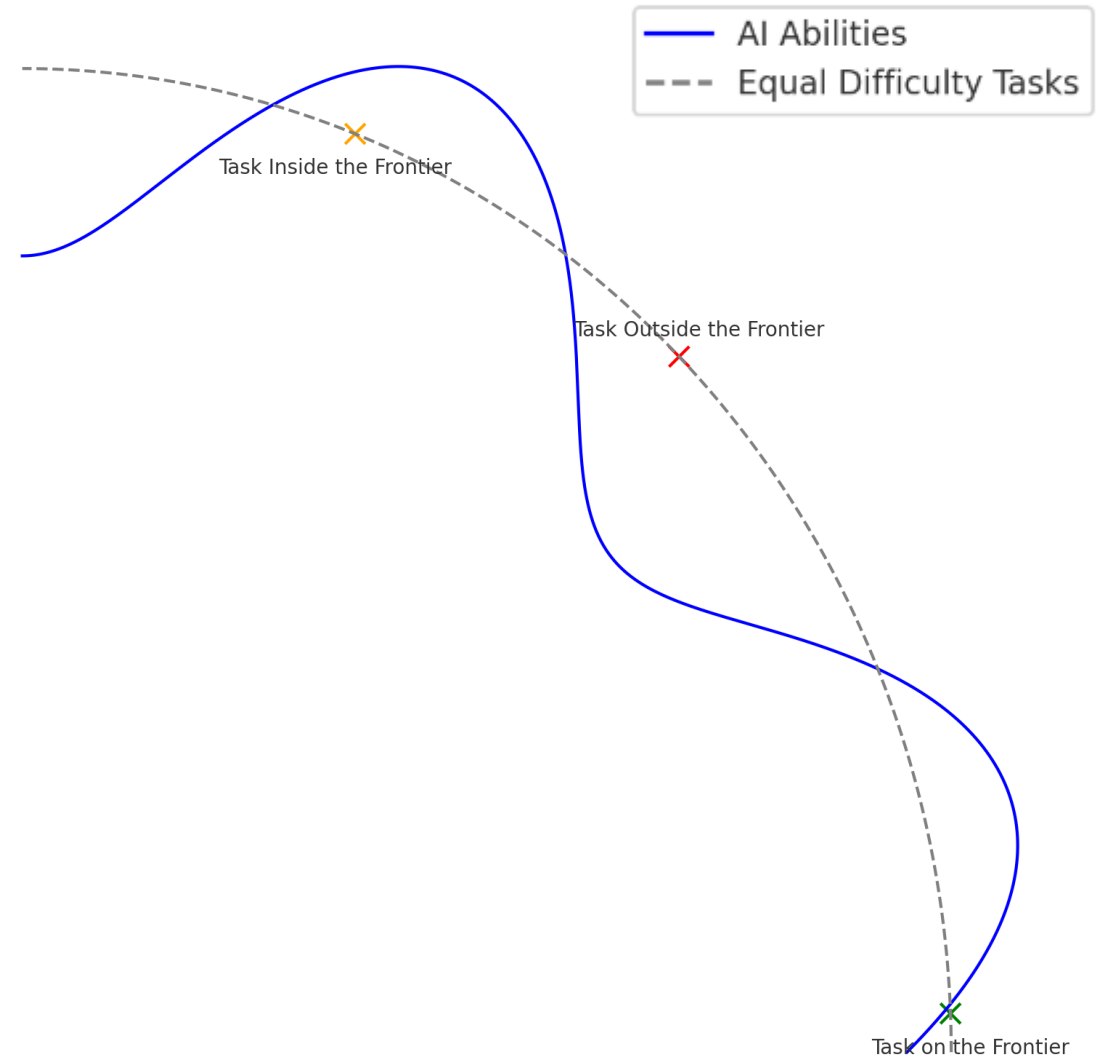
“AI is weird. No one actually knows the full range of capabilities... no one really knows the best ways to use [LLMs], or the conditions under which they fail.

There is no instruction manual.

On some tasks AI is immensely powerful, and on others it fails completely or subtly.

And... you won't know which is which.”

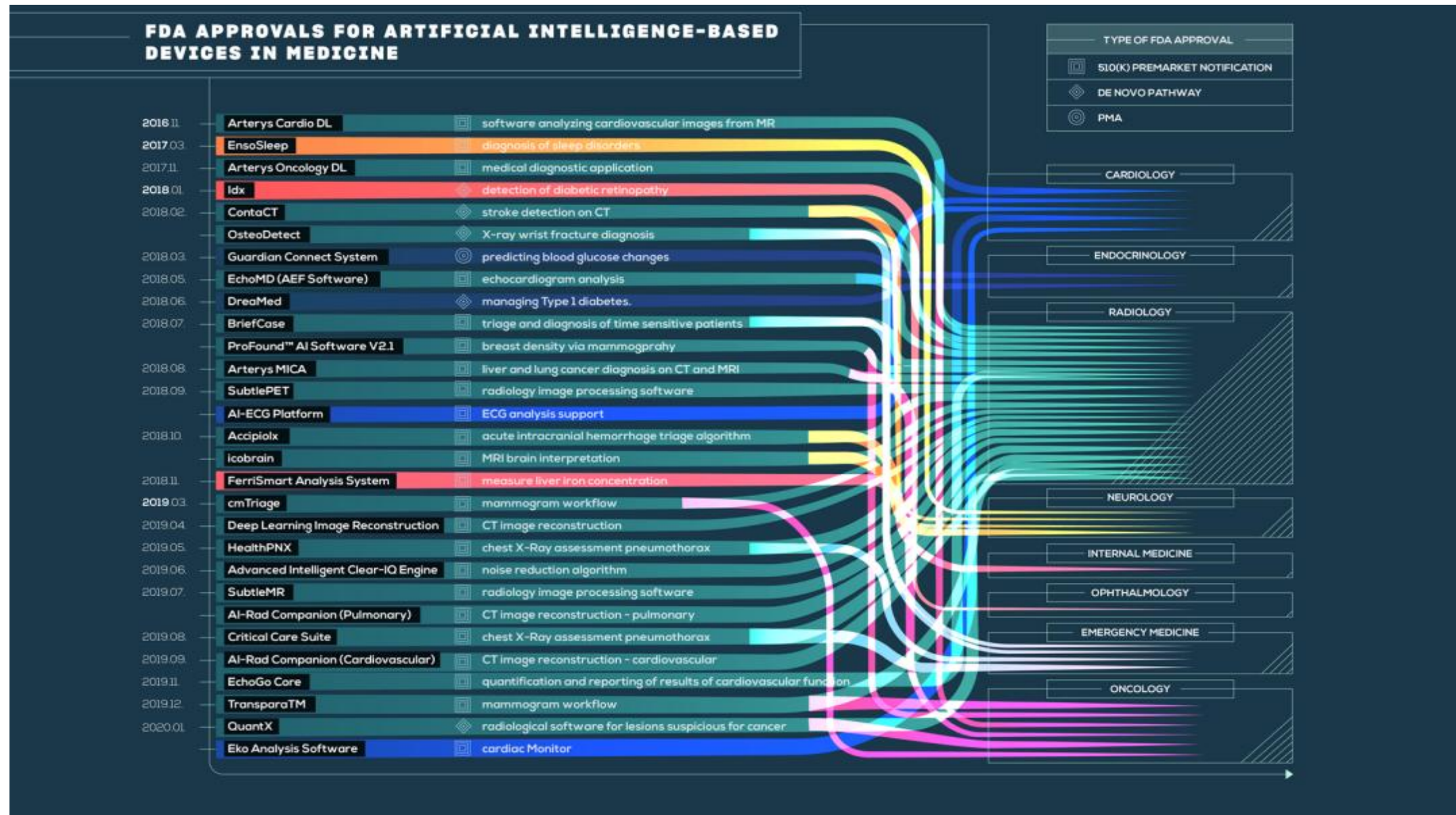
-Ethan Mollick





Slide bin

>900 FDA approved AI tools- most are incremental, 'work-flow' efficiencies



Many, many AI-powered radiology workflow tools

Triage and notification software

Solutions ▾Resources ▾About Us ▾CareersContact UsBook Demo

FDA Cleared & CE Marked



Intracranial Hemorrhage

510(k) Triage and notification software indicated for use in the analysis of non-enhanced head CT images; flags and communicates suspected positive findings of pathologies in head CT images, namely Intracranial Hemorrhage.



Large Vessel Occlusions

510(k) Triage and notification software indicated for use in the analysis of head CTA images; flags and communicates suspected positive findings of Large Vessel Occlusion (LVO).



Acute C-Spine Fracture

510(k) Triage and notification software indicated for use in the analysis of CT images of the cervical spine; flags and communicates suspected positive findings of linear fractures of the cervical spine bone in pattern.



Intra-Abdominal Free Gas

510(k) Triage and notification software indicated for use in the analysis of abdomen CT images; flags and communicates suspected positive cases of Intra-Abdominal Free Gas (IFG).



Incidental Pulmonary Embolism

510(k) Triage and notification software indicated for use in the analysis of CT images (not dedicated CTPA protocol); flags and communicates incidental Pulmonary Embolism (PE).

 VIZ.AI SYSTEM

now

Suspected LVO Identified:



71F @ St Hill Hospital

Your urgent review is requested

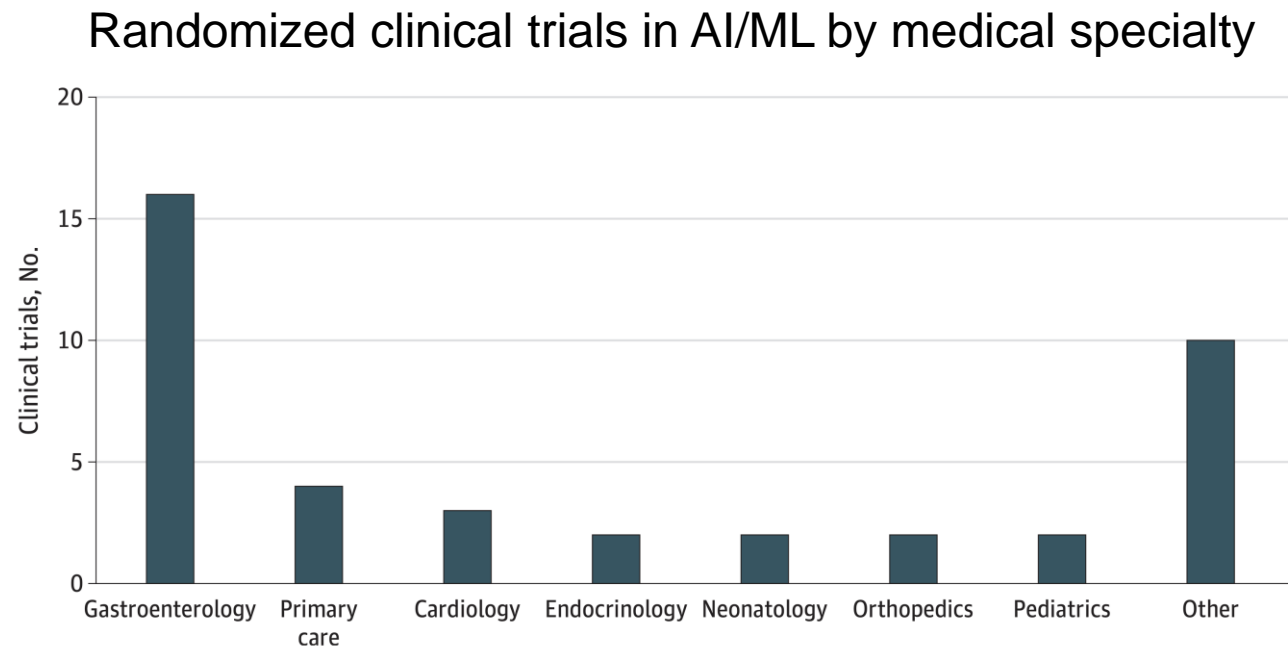
Press for more



Accelerated image acquisition protocols

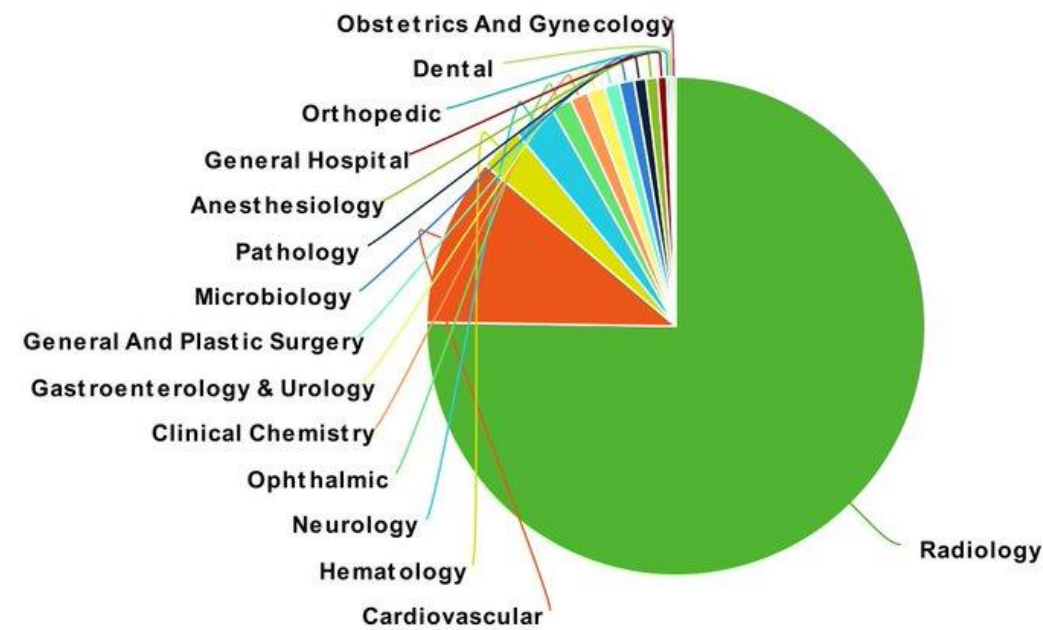
Original Scan	Accelerated Scan Enhanced by SubtleMR™
	
Acquisition Time: 1:46 min	Acquisition Time: 42 sec

The AI implementation gap between GI and radiology is a *workflow solution* gap



Plana et al *JAMA Netw Open* 2022

Current Distribution of FDA Cleared AI-Enabled Medical Devices Across Disciplines:



Source: Austin Barr