

AI in healthcare: hip hip hoera or AIAI?

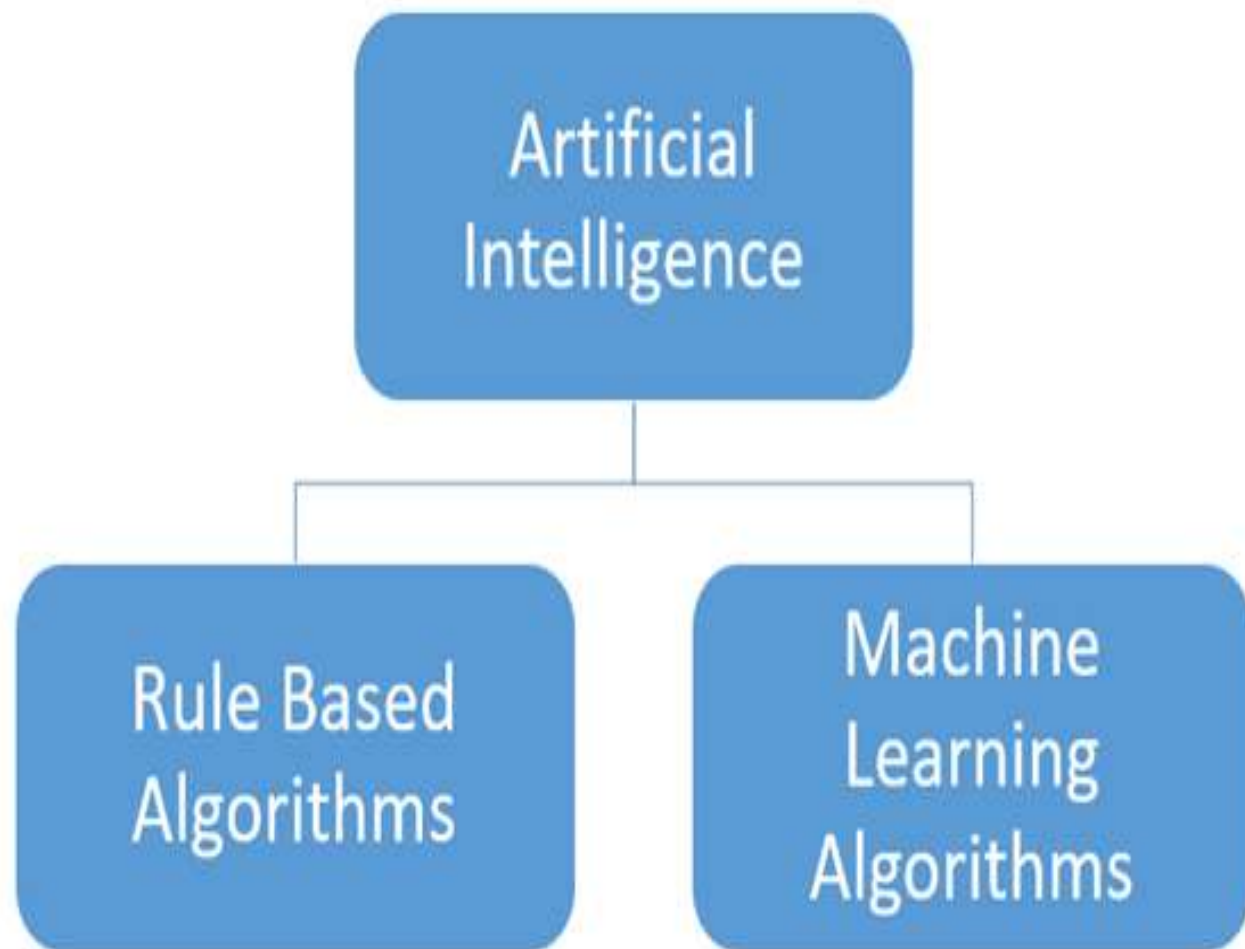
Prof Dr W. Van Biesen

Universitair Ziekenhuis Gent

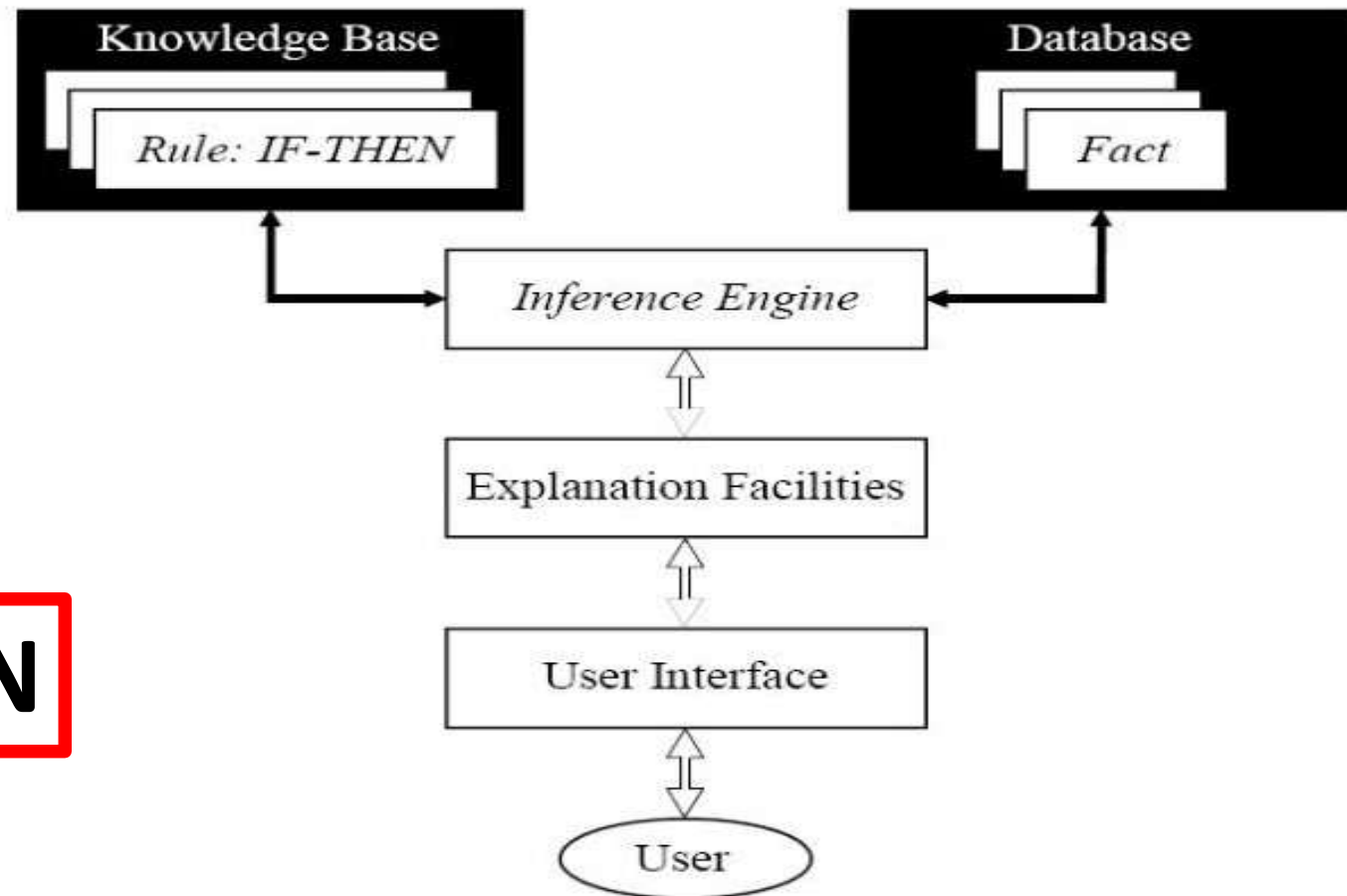
Consortium for Justifiable Digital Healthcare



<https://www.academiapress.be/nl/geneeskunde-tijden-van-ai>



Basic structure of a rule-based expert system



DEDUCTION

Potential interactions with medication:

A (lower)

X (higher)

Y (arrhythmias)

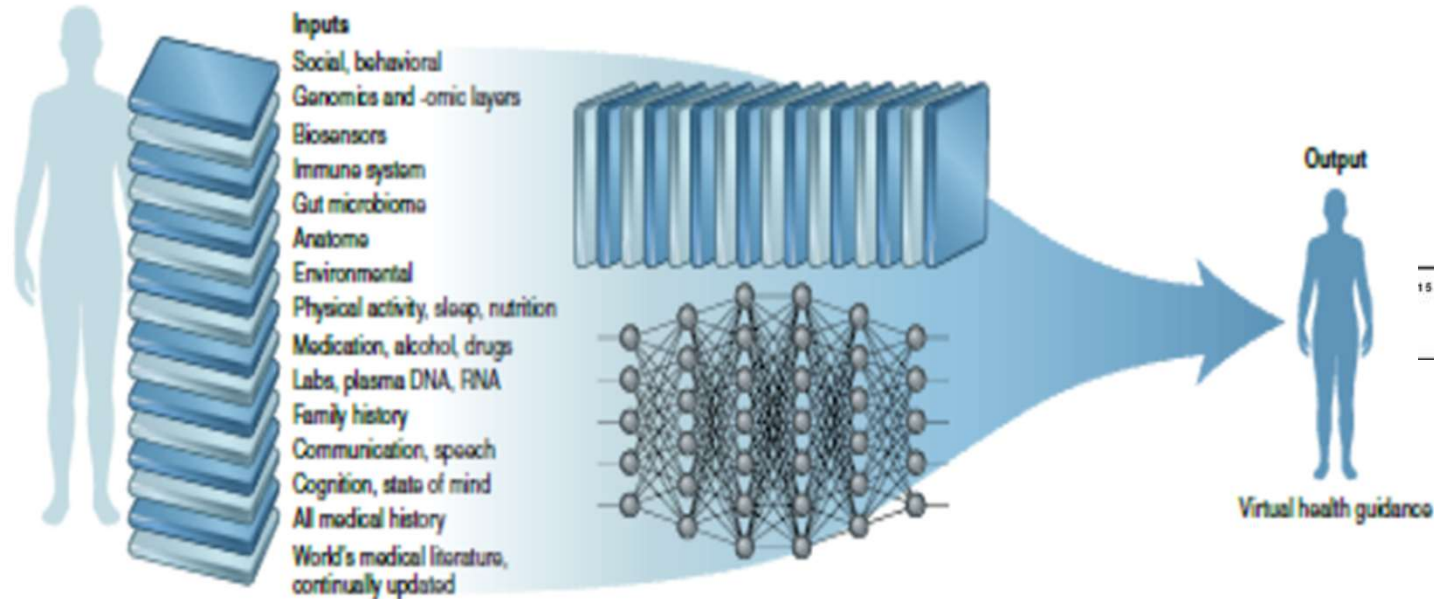
Alternative: ProductU[®]





DESIGN IS THE KEY





Konzentrations-Wirkungs-Kurven

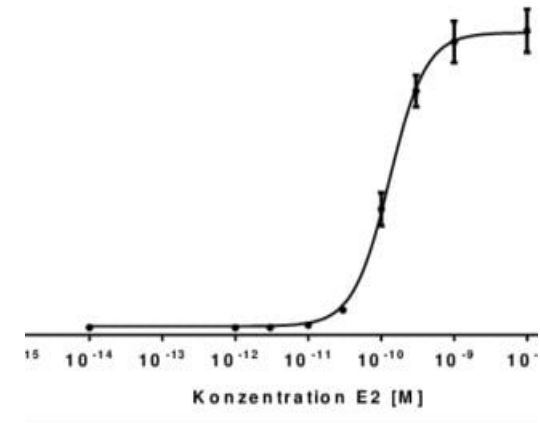
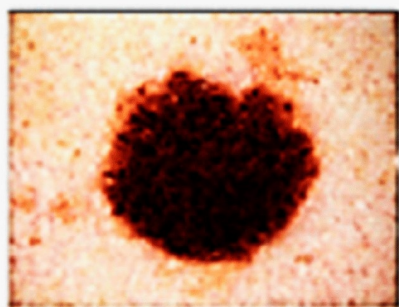


Figure 2: The virtual medical coach model with multi-modal data inputs and algorithms to provide individualised guidance.¹⁰⁵

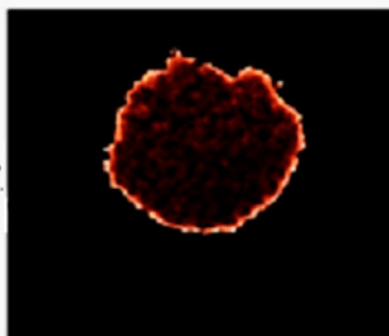
EMBEDDING

INDUCTION

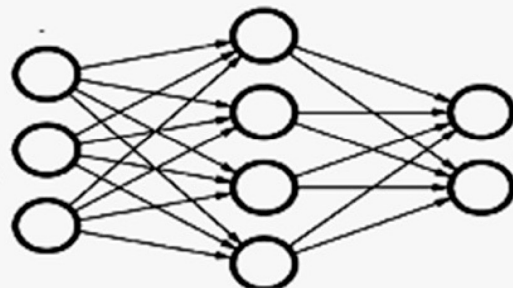
Image of Skin lesion



Segmented lesion

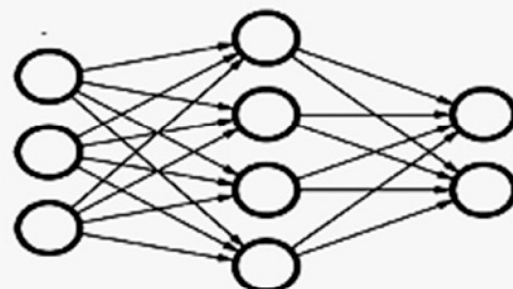
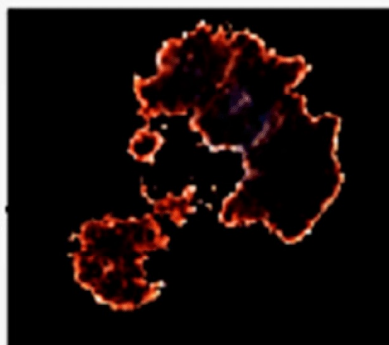
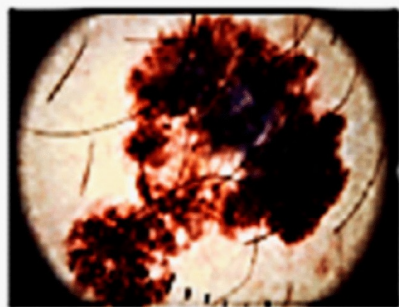


Learning model



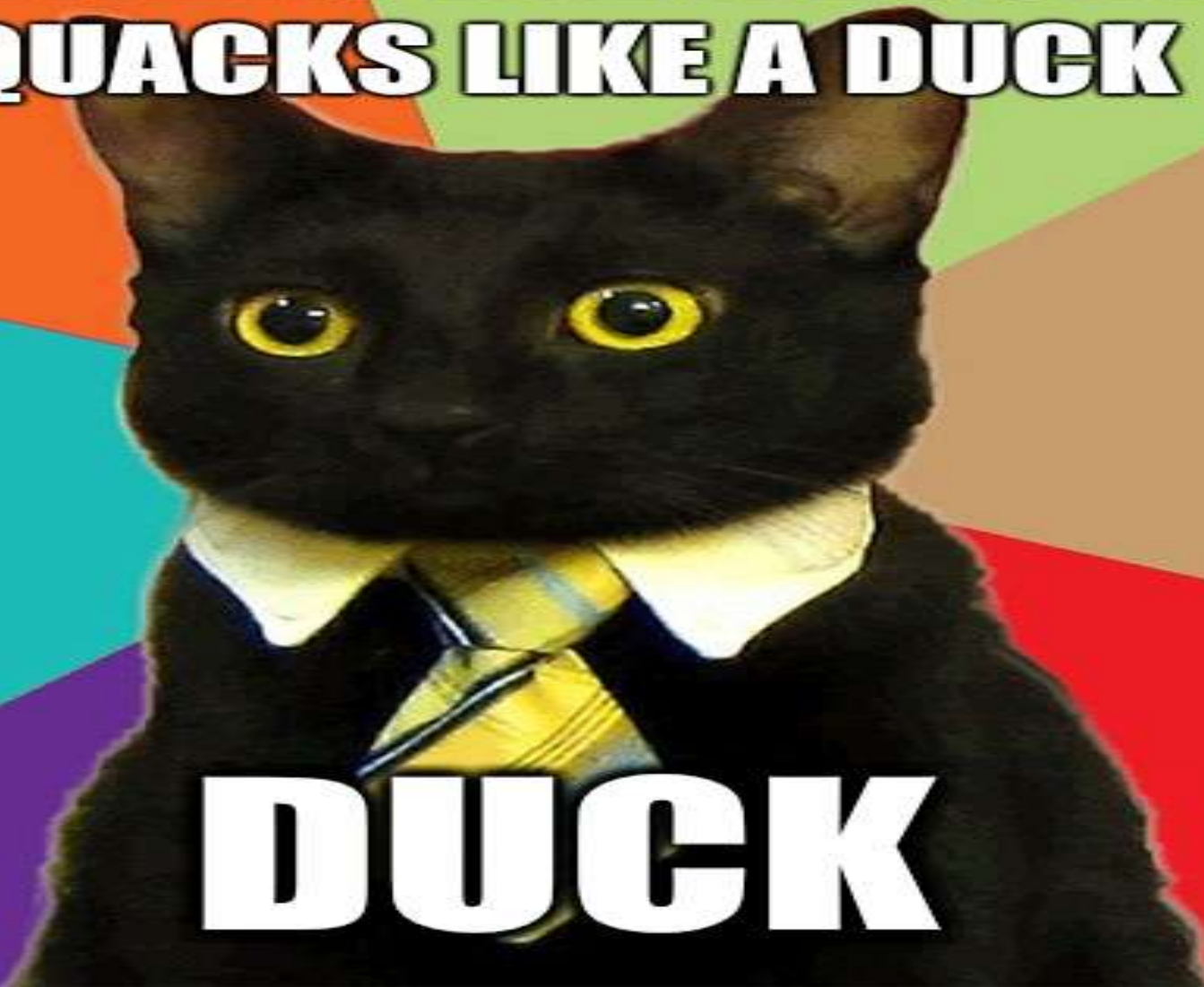
Prediction

Benign

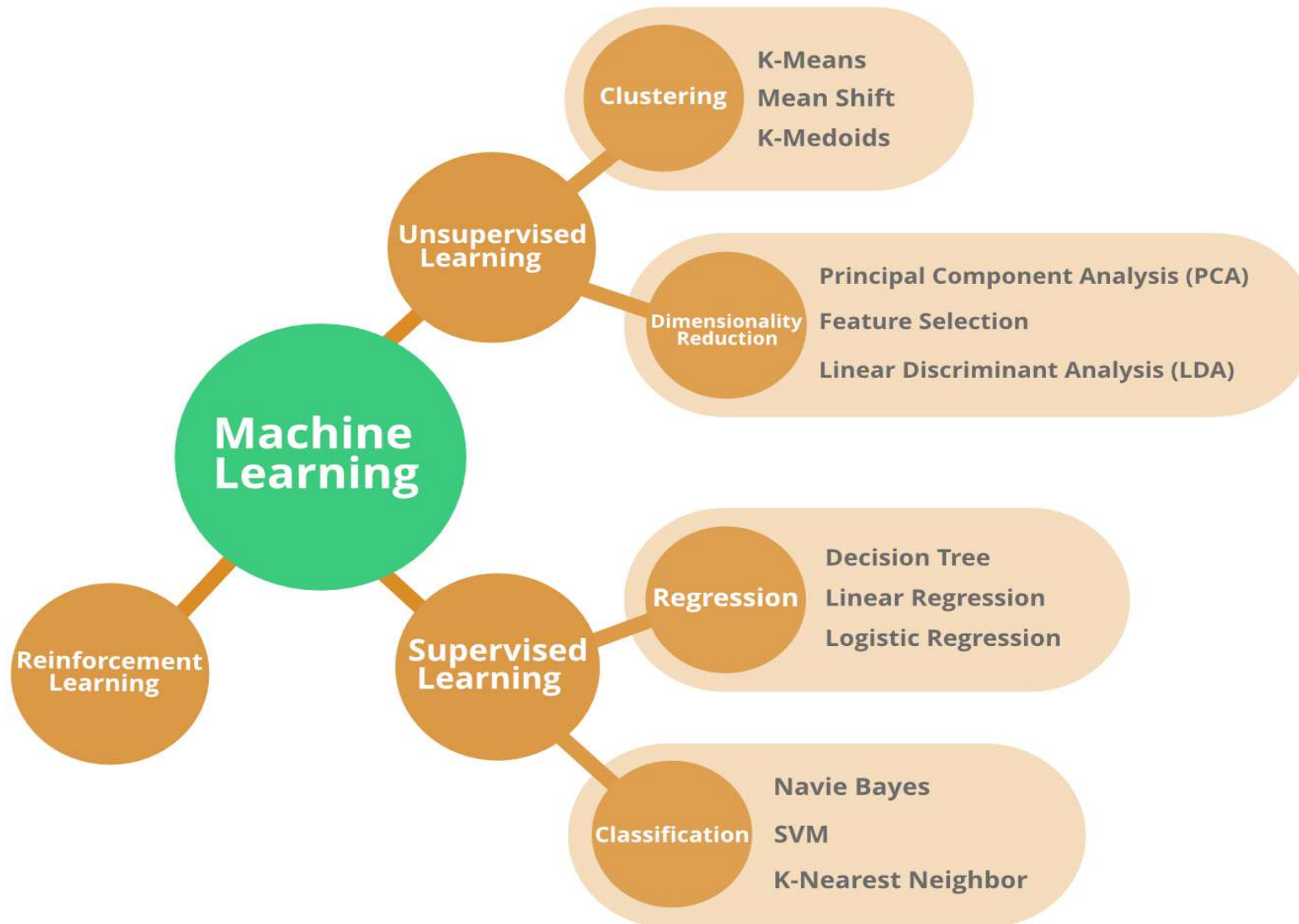


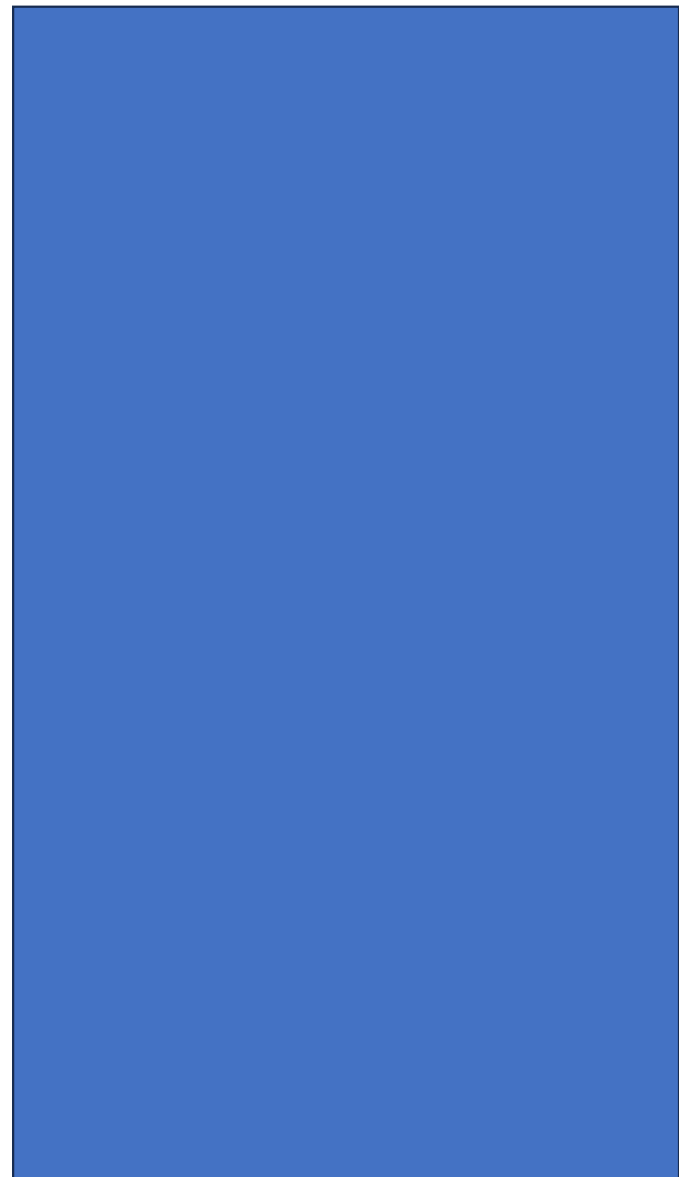
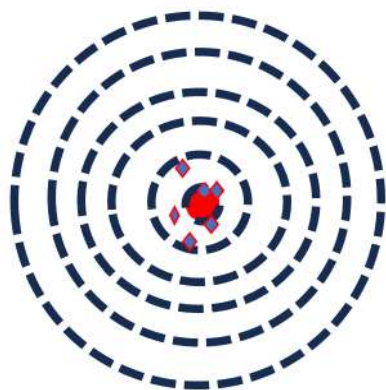
Malignant

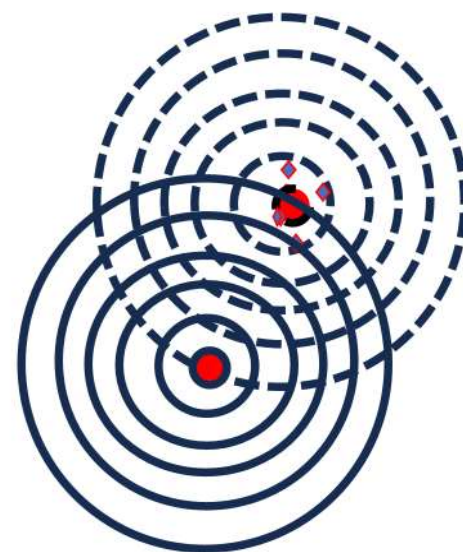
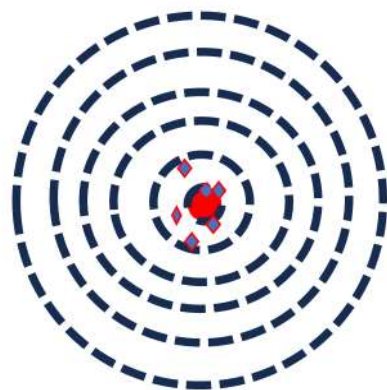
**IF IT WALKS LIKE A DUCK AND
QUACKS LIKE A DUCK**



DUCK







Knowledge

Why does A happen
if I do X

Information

What happens if I do A?

How does the event of seeing
X change the probability of
Pattern recognition

Data

Descriptive: people with A
have x in 75% of cases

Causation

Lack of a theoretical model



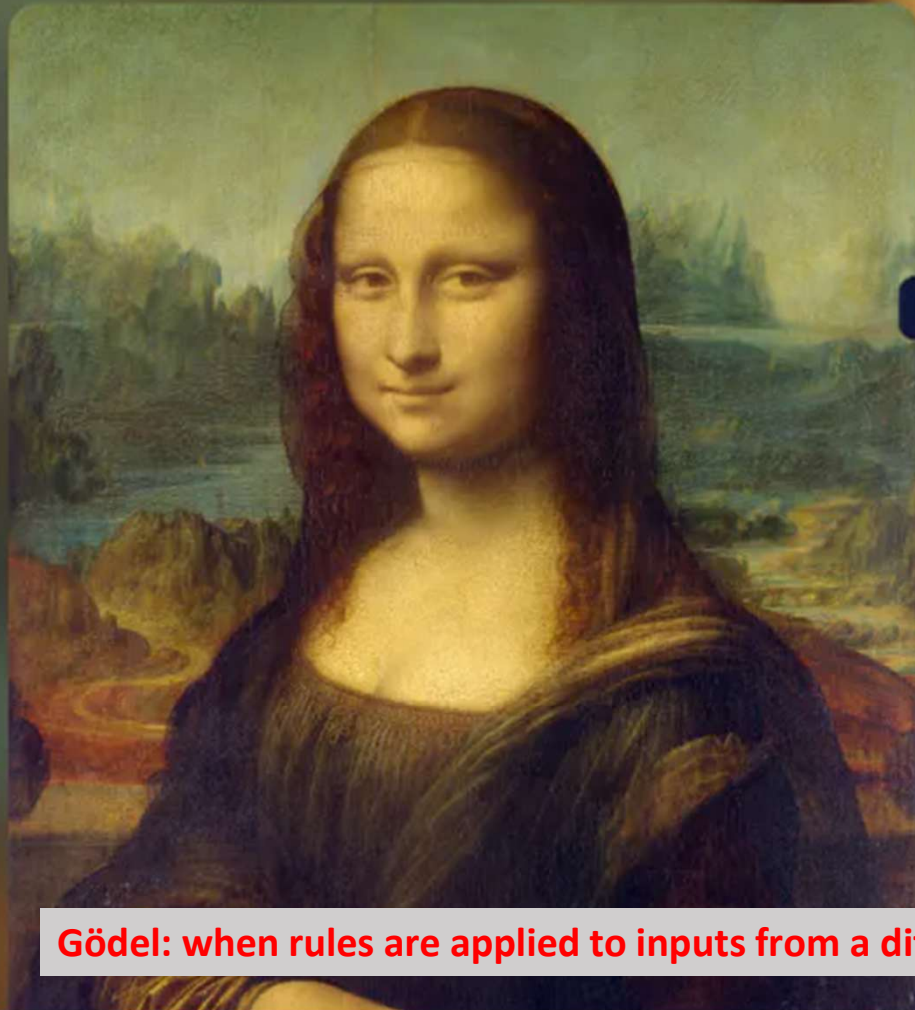
Adversarial effects

Generalizability

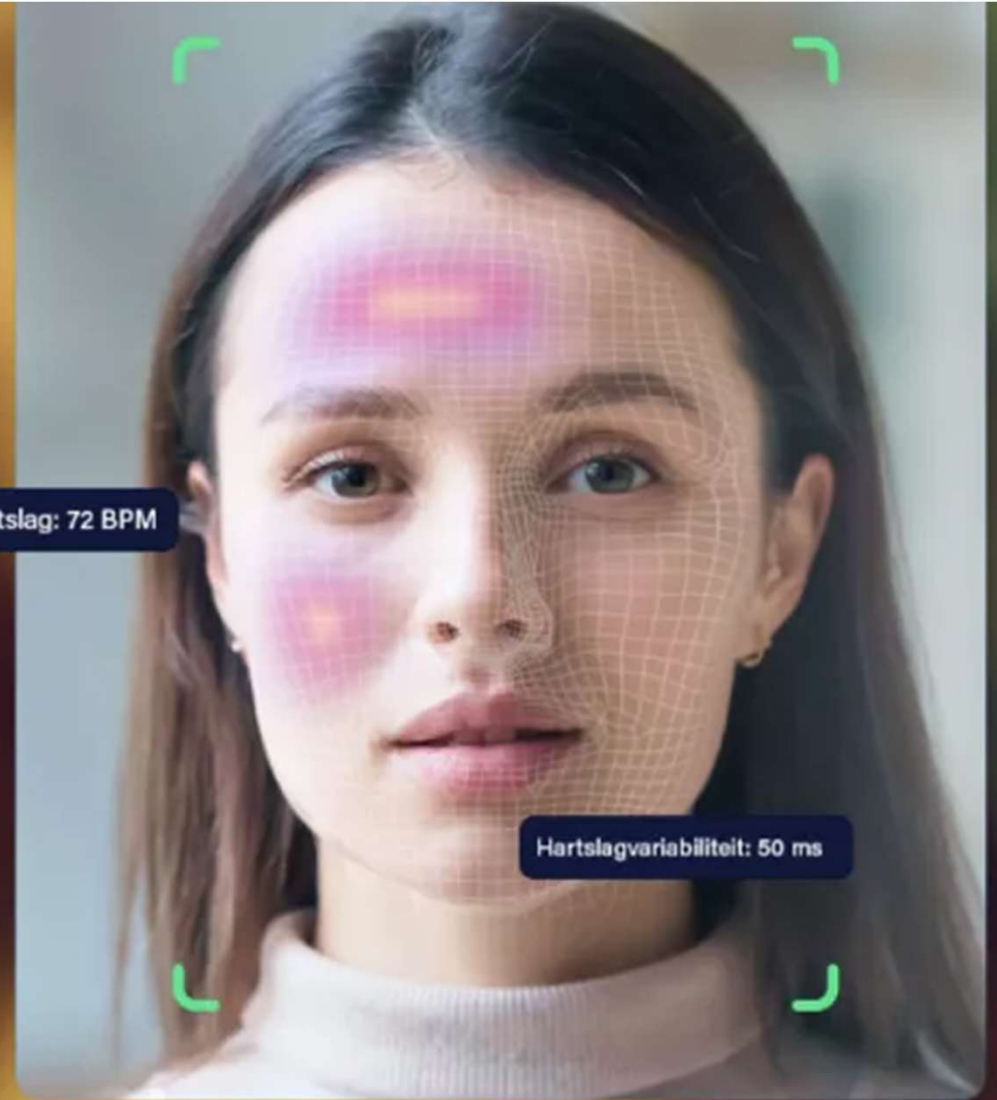
Transportability

Tank Problems

90% of people with the flu have fever
and muscle pain



Hartslag: 72 BPM



Hartslagvariabilitet: 50 ms

Gödel: when rules are applied to inputs from a different structure then we can expect weirdness

STEPHEN WOLFRAM

**What Is
ChatGPT
Doing...**

...and Why Does It Work?

Data

Human-generated text



Natural language



Code

Other text

MVSTQWYC... Protein

AGGACACG... DNA/RNA

CCC(=O)O... Chemical

Training

Autoregressive
(next token
prediction)

To be or not to ?

Masked
(fill in the blanks)

To ? or not ? be

Large language model



Output
likelihoods



Model
parameters



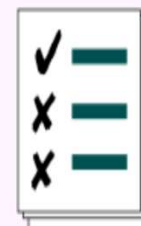
Input
sequence
of tokens

To be or not to

Zero-shot
transfer and
prompt
engineering

Supervised
finetuning
and RLHF

Task-specific
labeled data



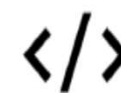
Applications



Question
answering



Literature/
health-record
summaries

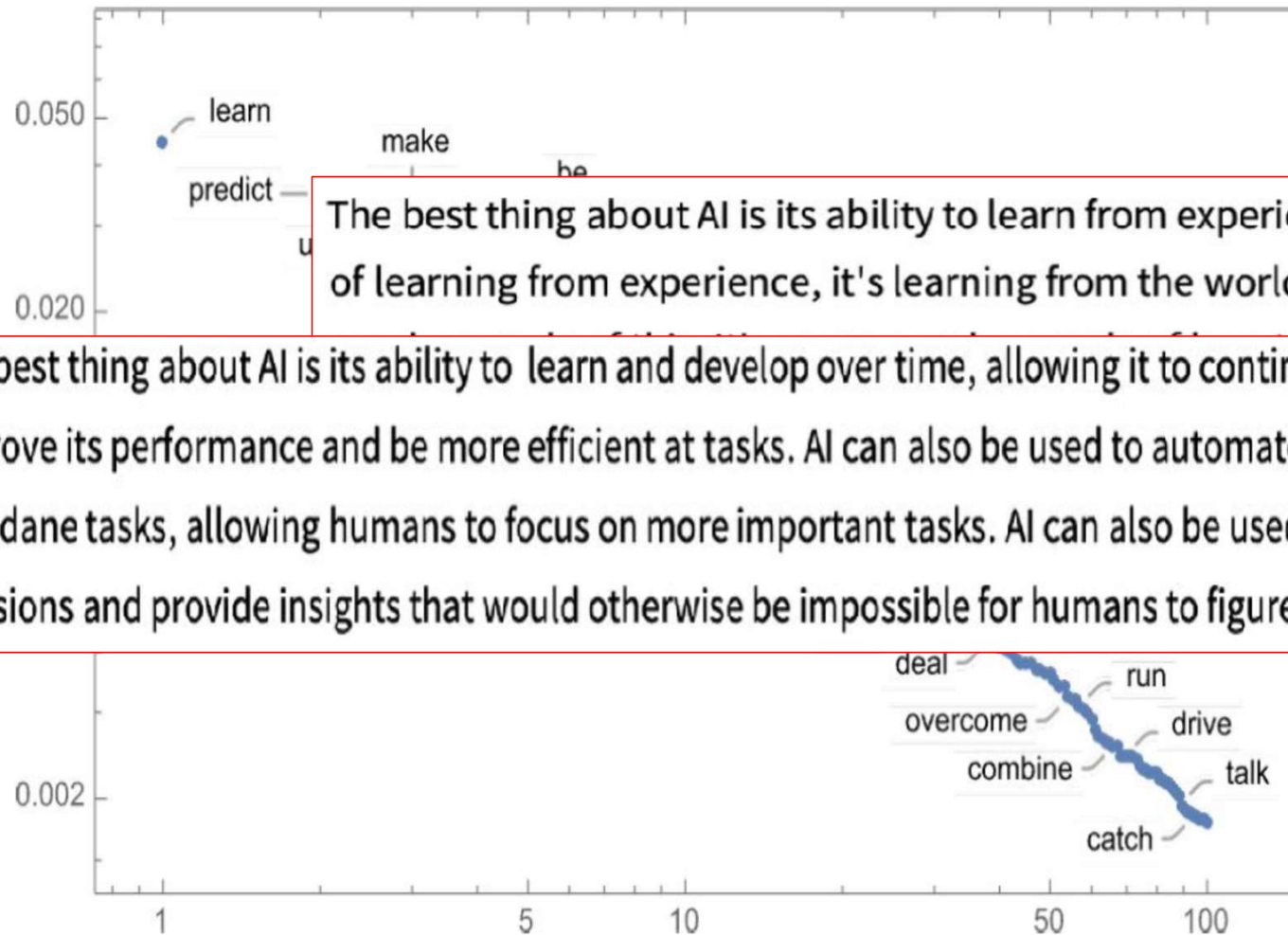


Scientific
coding



Molecular
design

{ The best thing about AI is its ability to,

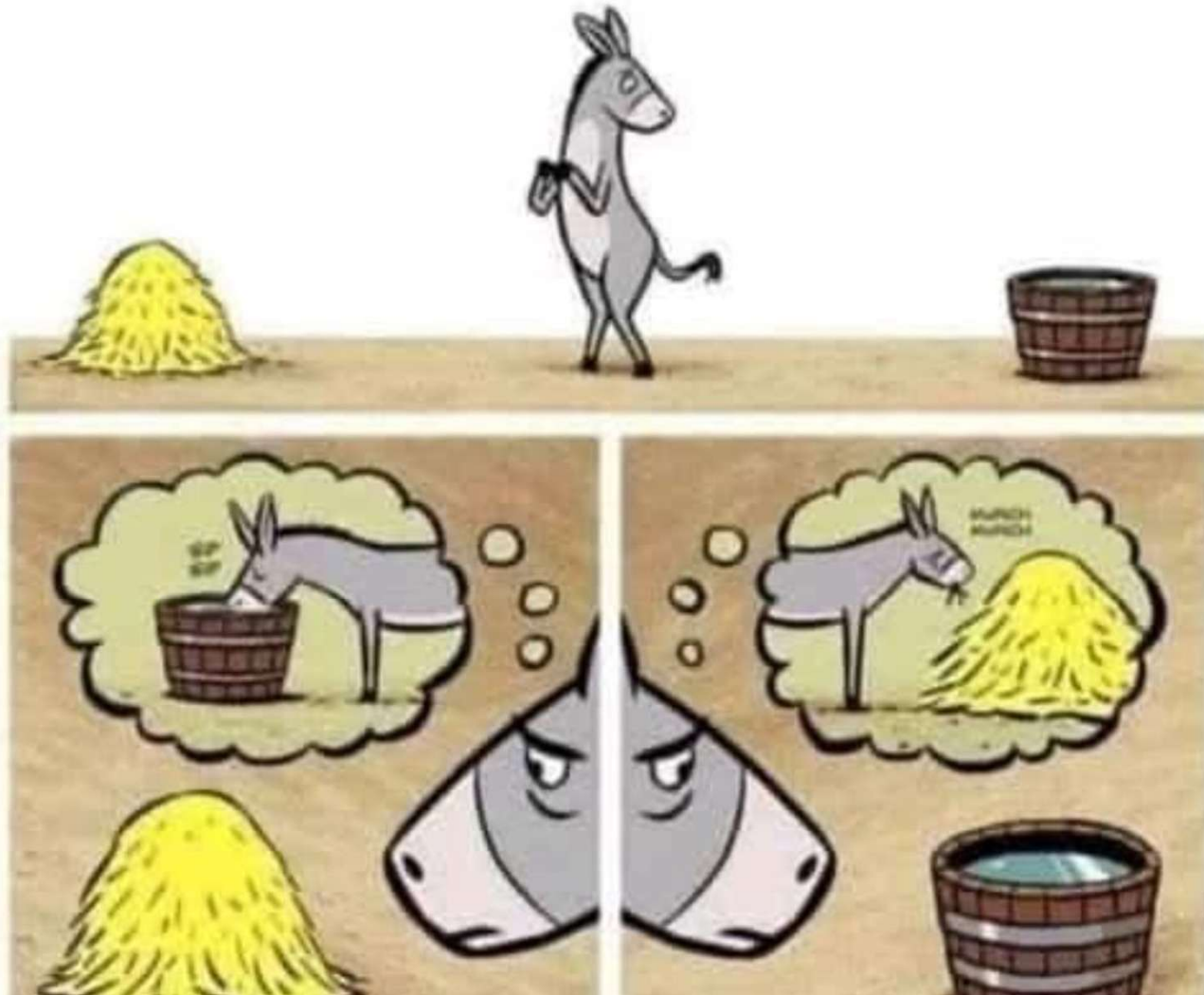


The best thing about AI is its ability to learn from experience. It's not just a matter of learning from experience, it's learning from the world around you. The AI is a very

The best thing about AI is its ability to learn and develop over time, allowing it to continually improve its performance and be more efficient at tasks. AI can also be used to automate mundane tasks, allowing humans to focus on more important tasks. AI can also be used to make decisions and provide insights that would otherwise be impossible for humans to figure out.

improve your life.
AI is a very good
example of how to use AI to

The NEED for RANDOMNESS



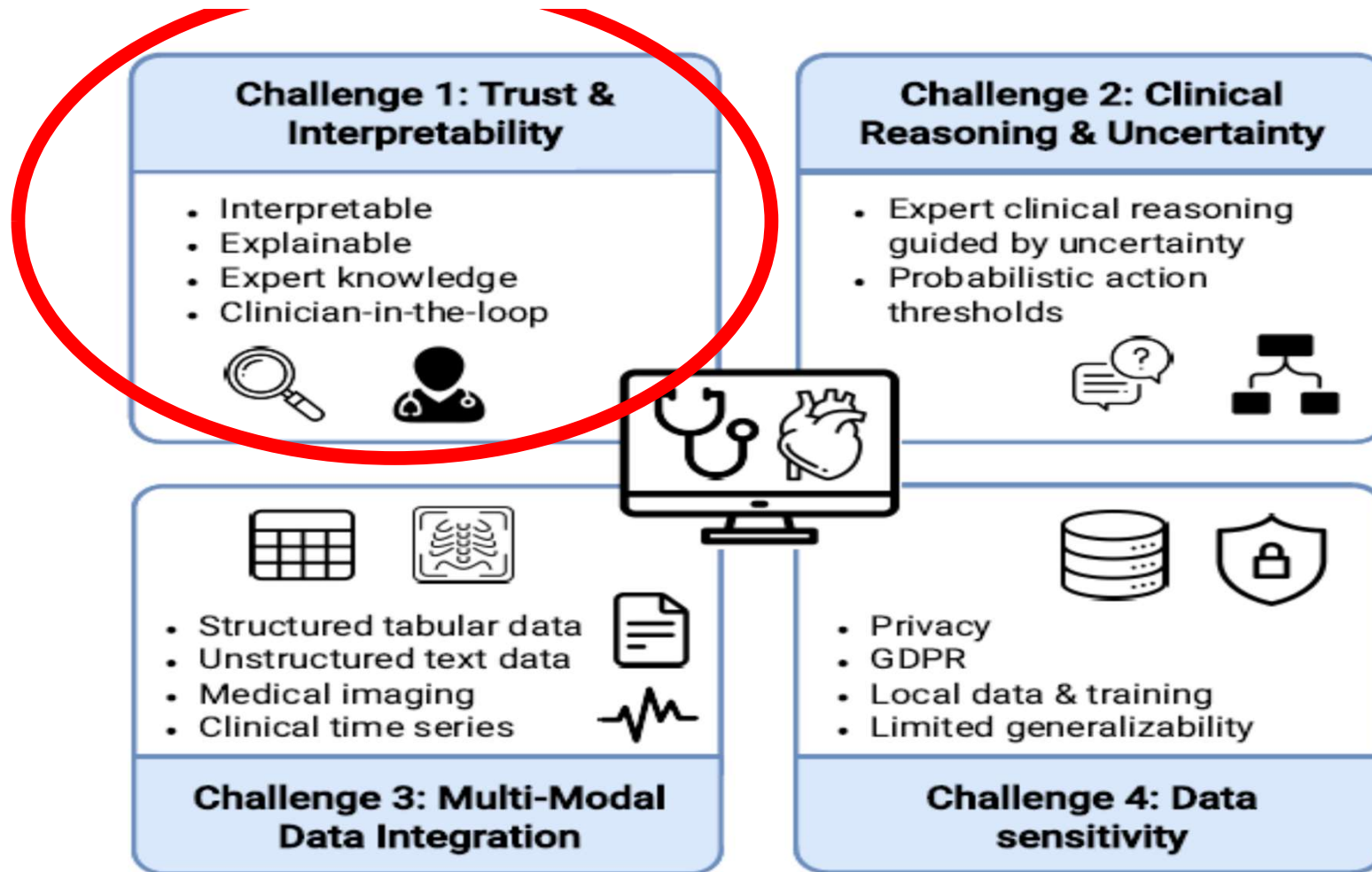


Figure 1.1: Four major technical challenges for AI in healthcare.

A clinically applicable approach to continuous prediction of future acute kidney injury

Nenad Tomašević^{1*}, Xavier Glorot¹, Jack W. Rae¹, Michal Zielinski¹, Harry Askham¹, Andre Saraiva¹, Anne M. Clemens Meyer¹, Suman Ravuri¹, Ivan Protsyuk¹, Alexander Connell¹, Cian O. Hughes¹, Alan Karanthesalingam¹, Julien Cornebise^{1,12}, Hugh Montgomery³, Geraint Rees⁴, Chris Laing⁵, William R. Baker¹, Kelly Peterson^{7,8}, R. Demis Hassabis¹, Dominic King¹, Mustafa Suleyman¹, Trevor Back^{1,13}, Christopher Nielson^{10,11,13}, Joseph R. Ledsam^{1,12*} & Shakir Mohamed^{1,13}

Nephrol Dial Transplant (2019) 1–2
doi: 10.1093/ndt/gfz226



Prediction of acute kidney injury using artificial intelligence: are we there yet?

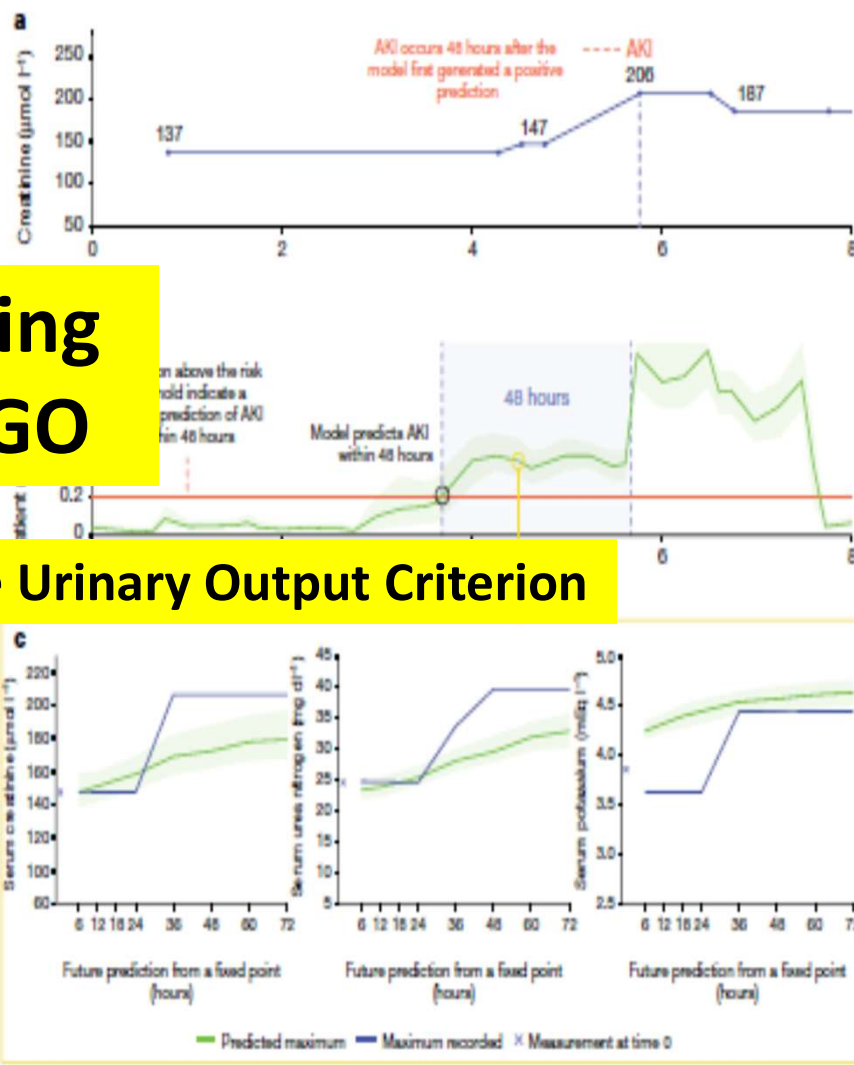
Wim Van Biesen^{1,2}, Jill Vanmassenhove¹ and Johan Decruyenaere^{2,3}

¹Renal Division, Ghent University Hospital, Ghent, Belgium, ²Justifiable Digital Health Consortium, Ghent University Hospital, Ghent, Belgium and ³Department of Intensive Care, Ghent University Hospital, Ghent, Belgium

Correspondence to: Wim Van Biesen, E-mail: Wim.vanbiesen@ghent.be

According to KDIGO

BUT does NOT include the Urinary Output Criterion

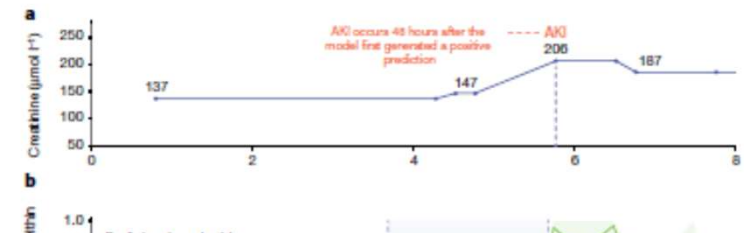


LETTER

<https://doi.org/10.1038/s41586-019-1390-1>

A clinically applicable approach to continuous prediction of future acute kidney injury

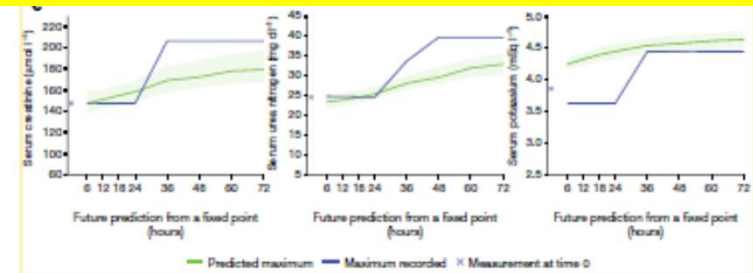
Nenad Tomašev^{1*}, Xavier Glorot¹, Jack W. Rae^{1,2}, Michal Zielinski¹, Harry Askham¹, Andre Saraiva¹, Anne Mottram¹, Clemens Meyer¹, Suman Ravuri¹, Ivan Protsyuk¹, Alistair Connell¹, Cian O. Hughes¹, Alan Karthikesalingam¹, Julien Cornebise^{1,12}, Hugh Montgomery³, Geraint Rees⁴, Chris Laing⁵, Clifton R. Baker⁶, Kelly Peterson^{7,8}, Ruth Reeves⁹, Demis Hassabis¹, Dominic King¹, Mustafa Suleyman¹, Trevor Back^{1,13}, Christopher Nielson^{10,11,13}, Joseph R. Ledsam^{1,13*} & Shakir Mohamed^{1,13}



**NOT having a Serum Creatinine Measured might contain INFORMATION
=> The device becomes pointless in alerting non-considerate physicians**

Nephrol Dial Transplant
doi: 10.1093/ndt/gfz226

ndt
Nephrology Dialysis Transplantation



Prediction of acute kidney injury using artificial intelligence:
are we there yet?

Wim Van Biesen^{1,2}, Jill Vanmassenhove¹ and Johan Decruyenaere^{2,3}

¹Renal Division, Ghent University Hospital, Ghent, Belgium, ²Justifiable Digital Health Consortium, Ghent University Hospital, Ghent, Belgium
and ³Department of Intensive Care, Ghent University Hospital, Ghent, Belgium

Correspondence to: Wim Van Biesen, E-mail: Wim.vanbiesen@UGhent.be

AI and Inequity

Bestaande ongelijkheid en discriminatie:

- Ongelijkheid in toegang en middelen voor gezondheidszorg
- Discriminerende processen in de zorg
- Bias in diagnostiek

Ongelijke beschikbaarheid van AI:

- Veronachtzaming en toename van de digitale kloof
- Stijgende ongelijkheid in toegang tot gezondheidszorg
- Misbruik en discriminatie in niet-medische toepassingen (verzekering, werving en selectie)



Bias in de data:

- Gebrek aan representatieve, inclusieve datasets
- Datasets met verborgen bias en ongelijkheid

Bias in ontwerp van AI-systemen:

- Ongelijkheid door keuze van problemen die met AI aangepakt worden
- Bias in ontwerpkeuzes en evaluatie
- Ongelijkheid in beveiliging en monitoring van systemen

Conceptualisation of the world

Manipulating symbols:

***\$*£%; %\$*£M; *\$*\$*\$£M**

Conceptualisation of the world

Manipulating symbols:

$2+2=4$; $4+2=6$; $2+2+2=6$

\$£%; %\$*£M; *\$*\$*\$£M

Creating “meaning”

VIEWPOINT

AI IN MEDICINE

AI's Threat to the Medical Profession

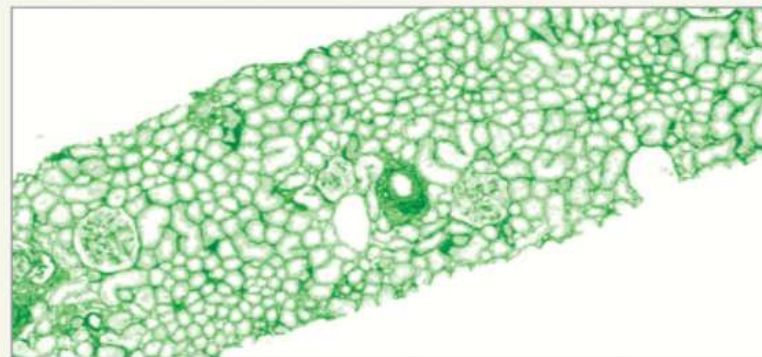
Agnes B. Fogo, MD

Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, Tennessee.

The Authors Guild and 17 authors recently filed a suit against OpenAI for copyright infringement of their works of fiction on behalf of writers whose works were used to train GPT. The [complaint](#) states that "Defendants then fed Plaintiffs' copyrighted works into their...algorithms designed to output human-seeming text responses" and that "at the heart of these algorithms is systematic theft on a mass scale." How different is this situation from the

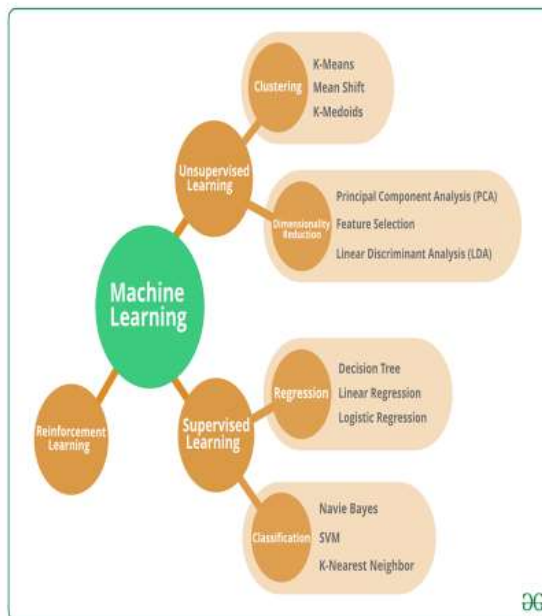
ments that

Figure. Whole Slide Image of Kidney Biopsy in Artificial Intelligence (AI) Staining Accompanied by AI-Generated Data



No. of glomeruli:	23
Global sclerosis:	4
Segmental sclerosis detected:	No
GBM changes detected:	No
Crescents detected:	No
Endocapillary hypercellularity:	2%
Podocyte density:	86.8 (below normal value)
Mesangial hypercellularity (maximum):	5
Adhesions:	No
Interstitial fibrosis:	17.4%
Tubular atrophy:	13.9%
Inflammation, area:	12.1%
Inflammatory cells detected:	Lymphocytes (82%) Granulocytes (1.2%) Plasma cells (3.5%) Undetermined (13.3%)
IF/IH:	Still in process
EM:	Still in process

The staining symbolizes a new, not currently used stain developed for AI purposes. EM indicates electron microscopy; GBM, glomerular basement membrane; and IF/IH, immunofluorescence/immunohistochemistry.



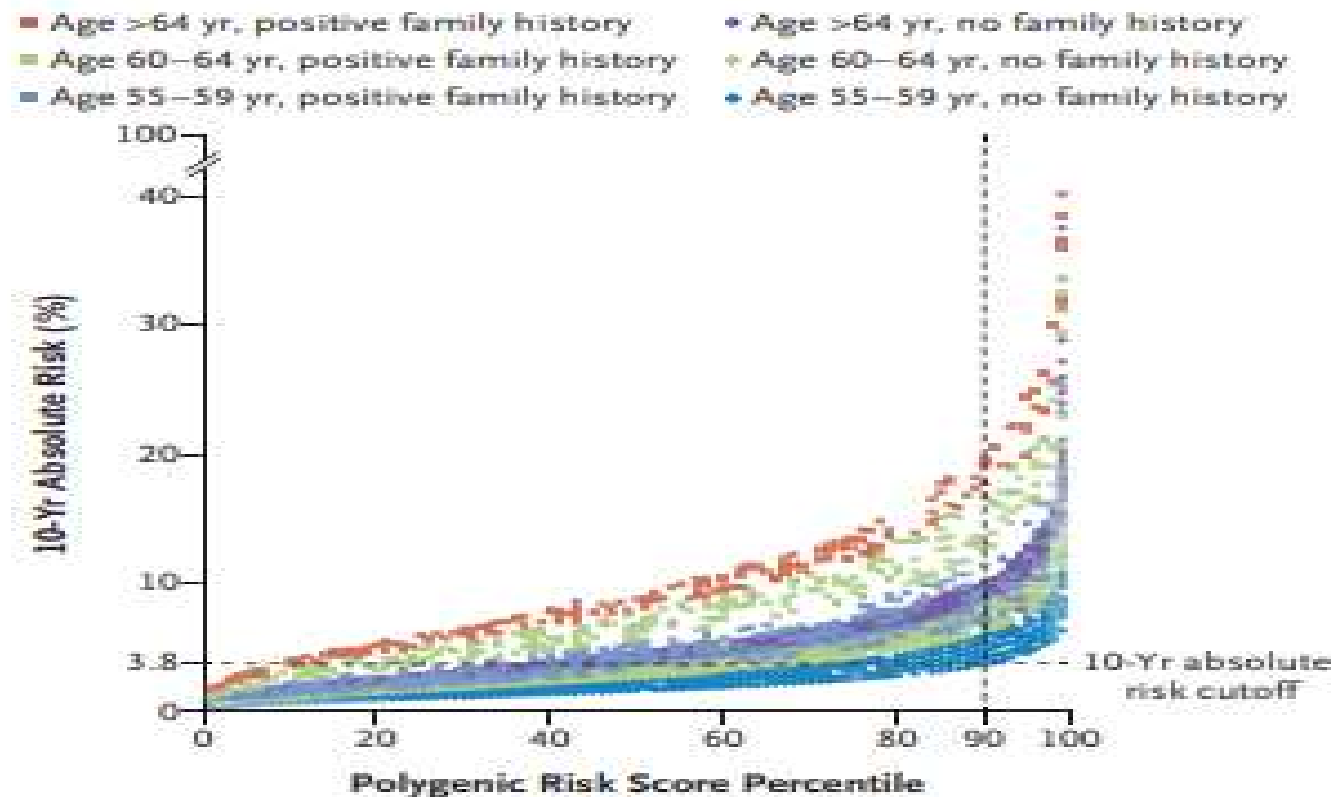


Figure 3. Absolute Risk of Prostate Cancer.

Shown is the 10-year absolute risk of prostate cancer (estimated from the iCARE22 algorithm) for all 6393 participants who agreed to screening and for whom we calculated a polygenic risk score, stratified according to participant age, family history of prostate cancer, and polygenic risk score. A threshold of 3.5 to 4% for the 10-year absolute risk of prostate cancer has been suggested as generating the greatest number of quality-adjusted life-years from risk-based screening and was used to calculate the 3.8% cutoff in this study.

“None of the other answers”: An LLM Weakness

Researchers tested whether LLMs could truly reason through medical questions by replacing the correct answer in multiple choice questions with “None of the other answers” (NOTA). Frontier models showed significant drops in accuracy, revealing that strong multiple choice performance, is in part, due to pattern recognition.

- Researchers modified 100 MedQA questions so that NOTA became the correct answer, creating a 68-item clinician-validated test of genuine reasoning. The pattern of answers has changed but the underlying clinical reasoning has not.
- DeepSeek-R1, o3-mini, Claude 3.5 Sonnet, Gemini 2.0 Flash, GPT-4o, and Llama 3.3-70B all performed worse on NOTA-modified questions. Significant decreases in performance were exhibited, ranging from 9% to 38%.
- A system that falls for example from 81% → 43% accuracy when a pattern changes would be unsafe for autonomous clinical use; rigorous benchmarks must test reasoning, not memorized answer distributions.

Table. Model Performance on Original and None of the Other Answers (NOTA)-Modified Questions*

Model	Accuracy, % (No./total No.)		Accuracy drop, % (No./total No.) [95 % CI]
	Original	NOTA-modified	
1	92.65 (63/68)	83.82 (57/68)	8.82 (6/68) [2.70-18.92]
2	95.59 (65/68)	79.41 (54/68)	16.18 (11/68) [10.81-29.73]
3	88.24 (60/68)	61.76 (42/68)	26.47 (18/68) [17.57-39.19]
4	92.65 (63/68)	58.82 (40/68)	33.82 (23/68) [24.32-47.30]
5	85.29 (58/68)	48.53 (33/68)	36.76 (25/68) [28.38-51.35]
6	80.88 (55/68)	42.65 (29/68)	38.24 (26/68) [27.03-51.35]

Bedi, Shah et al., JAMA Network Open, Aug. 2025

DATASETS, BENCHMARKS, AND PROTOCOLS

Assessment of Large Language Models in Clinical Reasoning: A Novel Benchmarking Study

Liam G. McCoy , M.D., M.Sc.,^{1,2,3} Rajiv Swamy , S.M.,⁴ Nidhish Sagar , S.M.,³ Minjia Wang , M.Eng.,⁴ Stephen Bacchi , M.B.B.S., Ph.D.,⁵ Jie Ming Nigel Fong , M.R.C.P.,⁶ Nigel C.K. Tan , M.S.-H.P.Ed., F.R.C.P.(Edin),⁷ Kevin Tan , B.M.B.S., M.S.-H.P.Ed.,^{7,8} Thomas A. Buckley , B.S.,⁹ Peter Brodeur , M.D., M.A.,² Leo Anthony Celi , M.D., M.P.H., M.Sc.,^{2,3,10} Arjun K. Manrai , Ph.D.,^{9,11} Aloysius Humbert , M.D.,¹² and Adam Rodman , M.D., M.P.H.,^{2,11}

Received: February 6, 2025; Revised: June 28, 2025; Accepted: July 11, 2025; Published: September 25, 2025

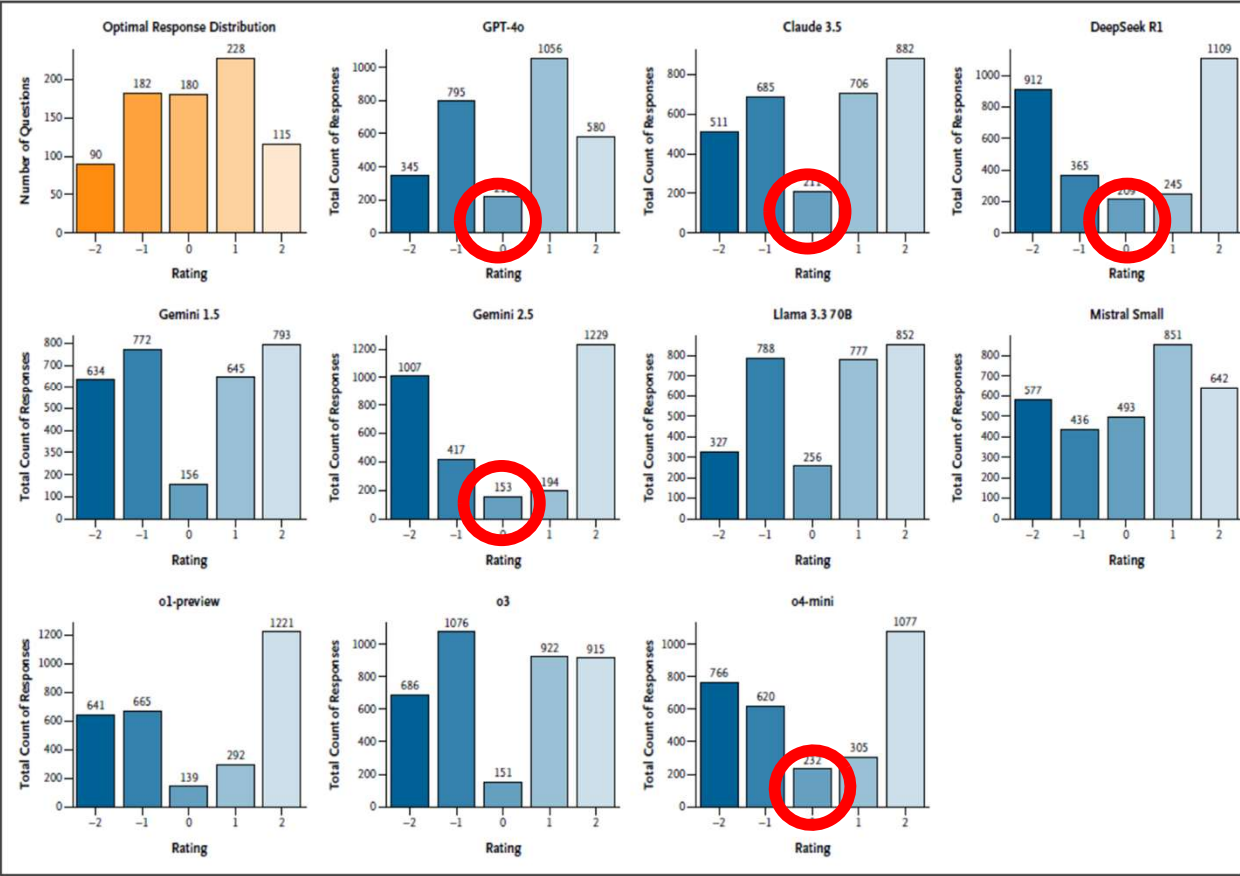


Figure 4. Comparison of Optimal Response Distribution with Overall Model Response Distribution. Plots denote total sums of model responses, as compared with the ideal response distribution with perfect concordance to the human experts (top-left plot).

GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial

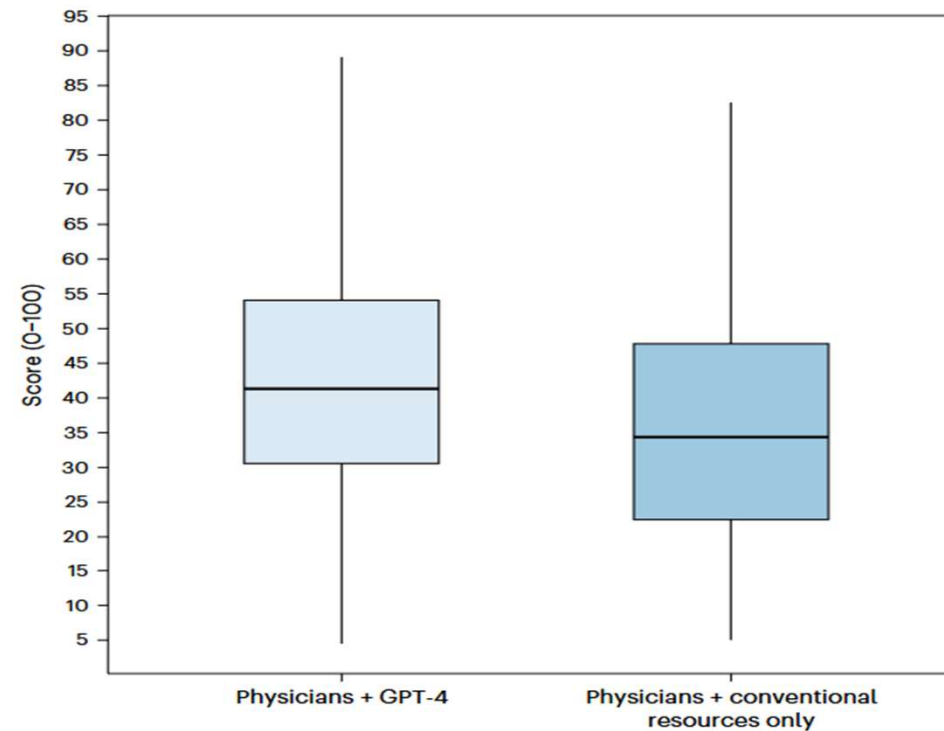


Fig. 2 | Comparison of the primary outcome for physicians with LLM and with conventional resources only (total score standardized to 0–100). Ninety-two physicians (46 randomized to the LLM group and 46 randomized to conventional resources) completed 375 cases (178 in the LLM group, 197 in the conventional resources group). The center of the box plot represents the median, with the boundaries representing the first and third quartiles. The whiskers represent the furthest data points from the center within 1.5 times the IQR.

GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial

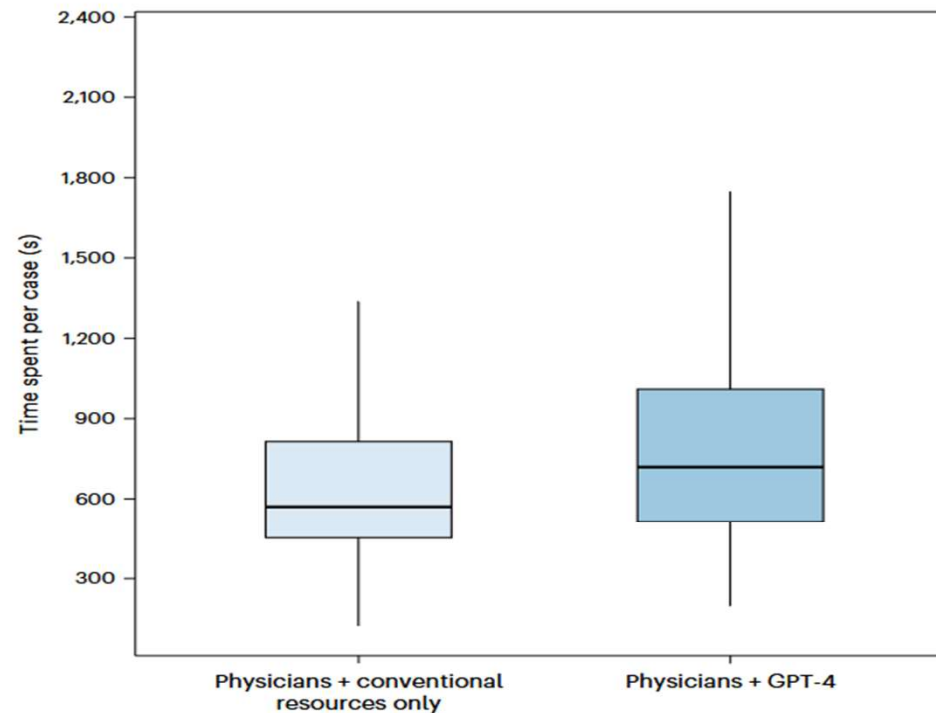


Fig. 4 | Comparison of the time spent per case by physicians using GPT-4 and physicians using conventional resources only. Ninety-two physicians (46 randomized to the LLM group and 46 randomized to the conventional resources) completed 375 cases (178 in the LLM group, 197 in the conventional resources group). The center of the boxplot represents the median, with the boundaries representing the first and third quartiles. The whiskers represent the furthest data points from the center within 1.5 times the IQR.

AUTOMATION BIAS

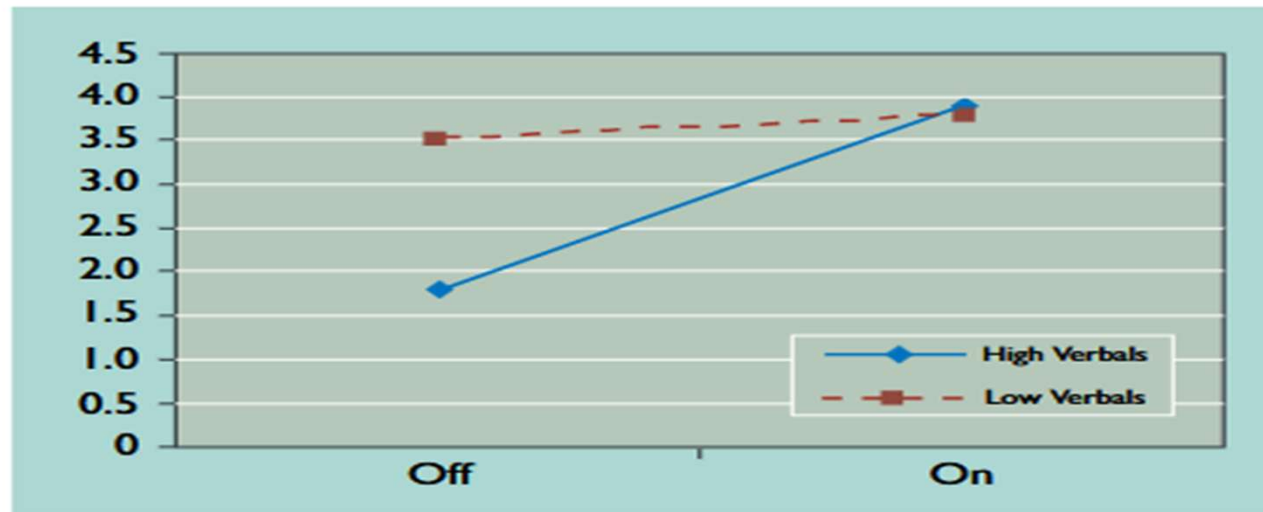
DESKILLING

MEANINGFUL HUMAN CONTROL

BY DENNIS F. GALLETTA, ALEXANDRA DURCIKOVA,
ANDREA EVERARD, AND BRIAN M. JONES

Does Spell-Checking Software Need a Warning Label?

Users—ironically, often those most verbally armed—put too much trust and little effort in questioning spell- and grammar-checking programs.



Automation bias

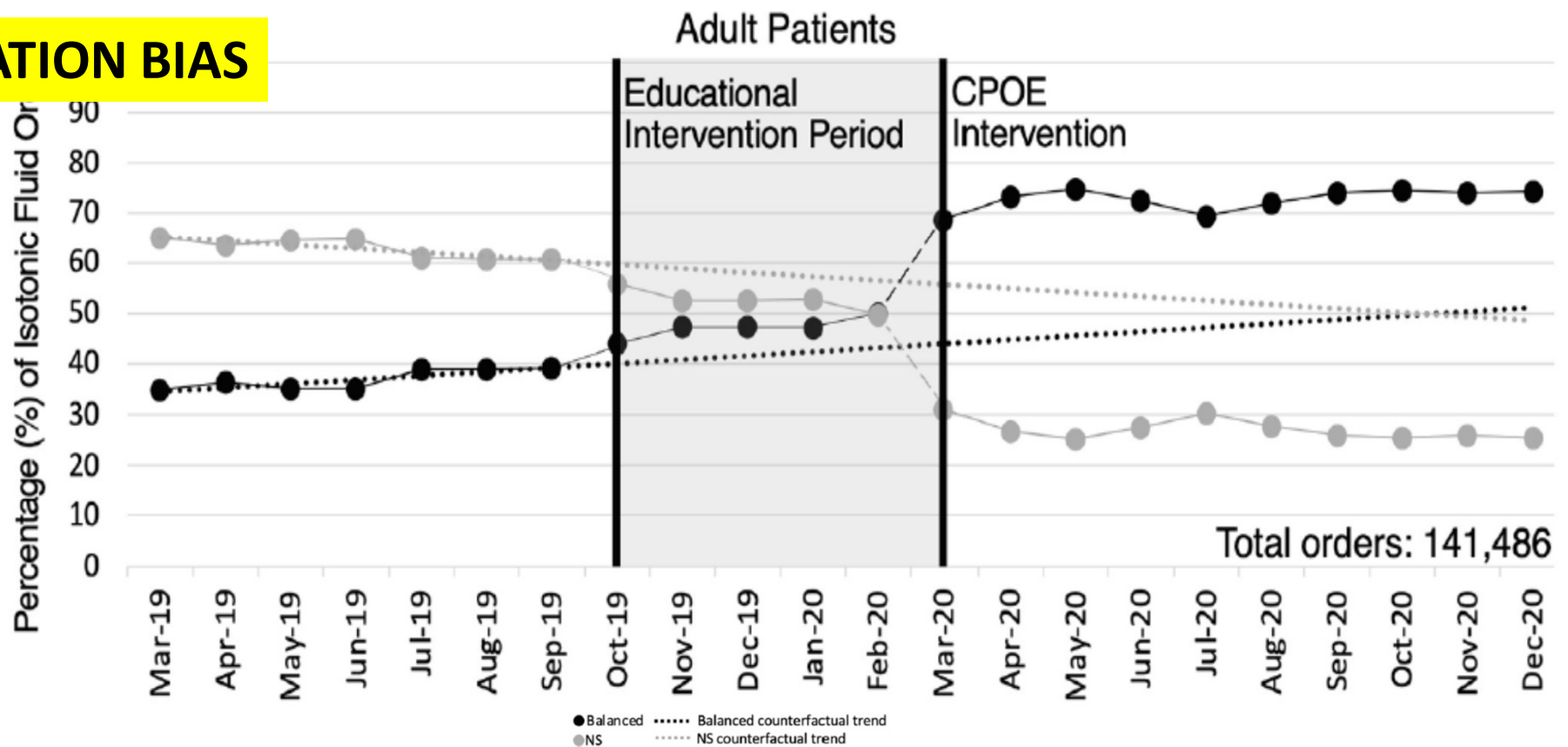
degrade the performance of only low verbals because those individuals are more likely to be fooled by

Figure 3.
Uncorrected false
negatives.

Development and implementation of a clinical decision support-based initiative to drive intravenous fluid prescribing

Michelle C. Spiegel^{a,*}, Annie N. Simpson^b, Achsah Philip^c, Carolyn M. Bell^d,
Nandita R. Nadig^a, Dee W. Ford^a, Andrew J. Goodwin^a

AUTOMATION BIAS



BY DENNIS F. GALLETTA, ALEXANDRA DURCIKOVA,
ANDREA EVERARD, AND BRIAN M. JONES

Does Spell-Checking Software Need a Warning Label?

Users—ironically, often those most verbally armed—put too much trust and little effort in questioning spell- and grammar-checking programs.

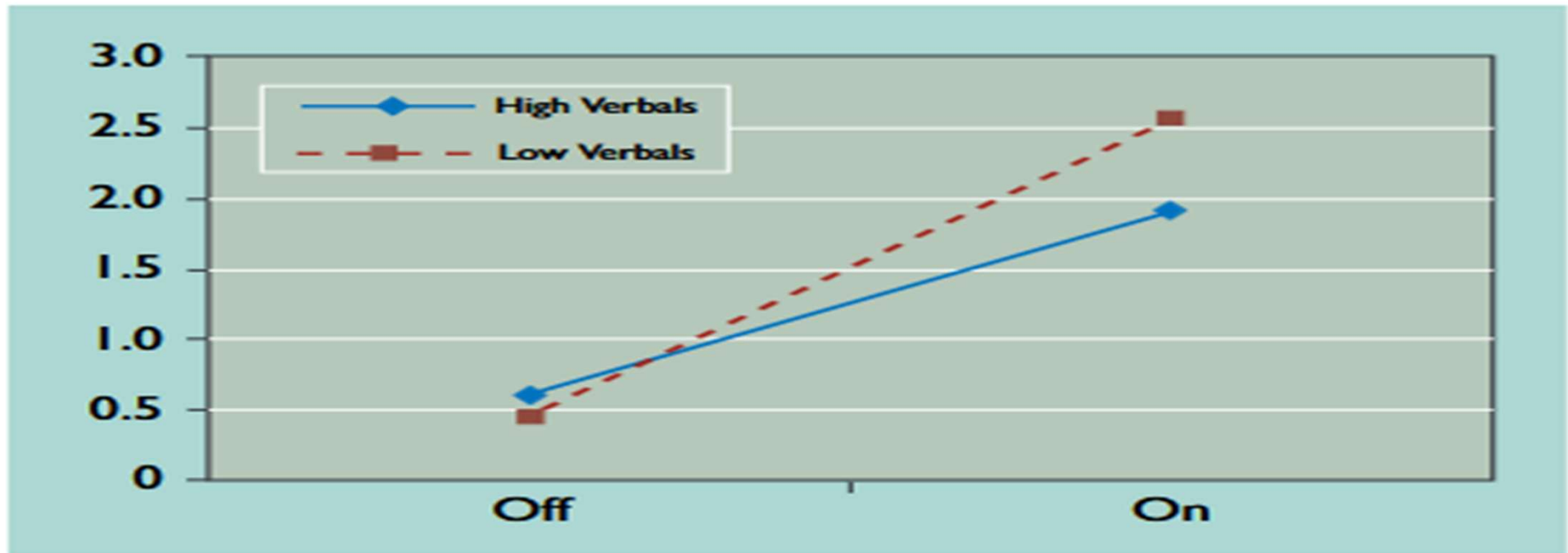


Figure 2. Uncorrected false positives.

result, 32 subjects were categorized as low verbals and 33 were

deskilling

Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study



Krzysztof Budzyń, Marcin Romańczyk, Diana Kitala, Paweł Kołodziej, Marek Bugajski, Hans O Adami, Johannes Blom, Marek Buszkiewicz, Natalie Halvorsen, Cesare Hassan, Tomasz Romańczyk, Øyvind Holme, Krzysztof Jarus, Shona Fielding, Melina Kunar, Maria Pellise, Nastazja Pilonis, Michał Filip Kamiński, Mette Kalager, Michael Bretthauer, Yuichi Mori

Summary

Background It is not known if continuous exposure to artificial intelligence (AI) changes endoscopists' behaviour when conducting colonoscopy. We assessed how endoscopists who regularly used AI performed colonoscopy when

Lancet Gastroenterol Hepatol
2025

Endoscopist deskilling: an unintended consequence of AI-assisted colonoscopy?

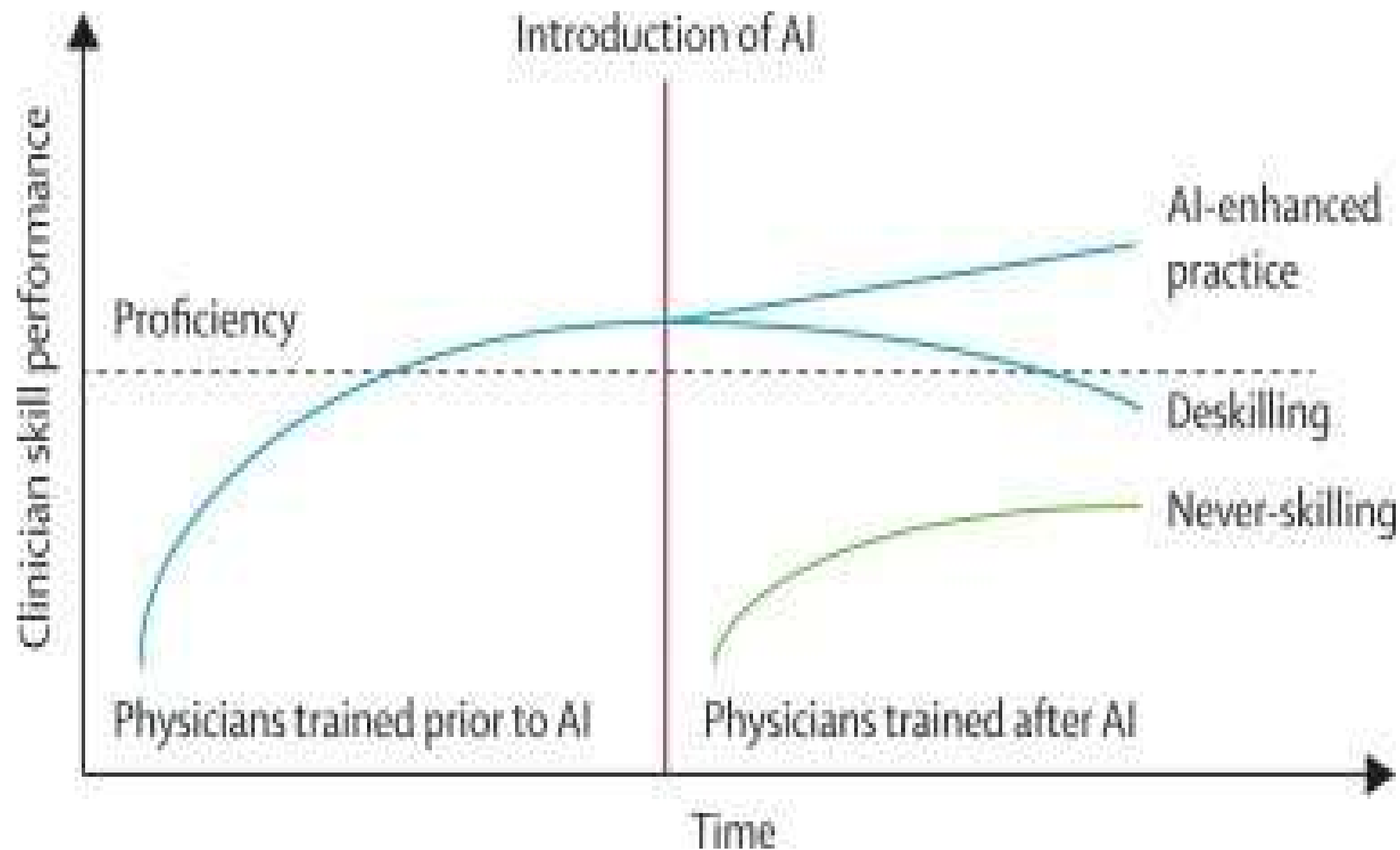


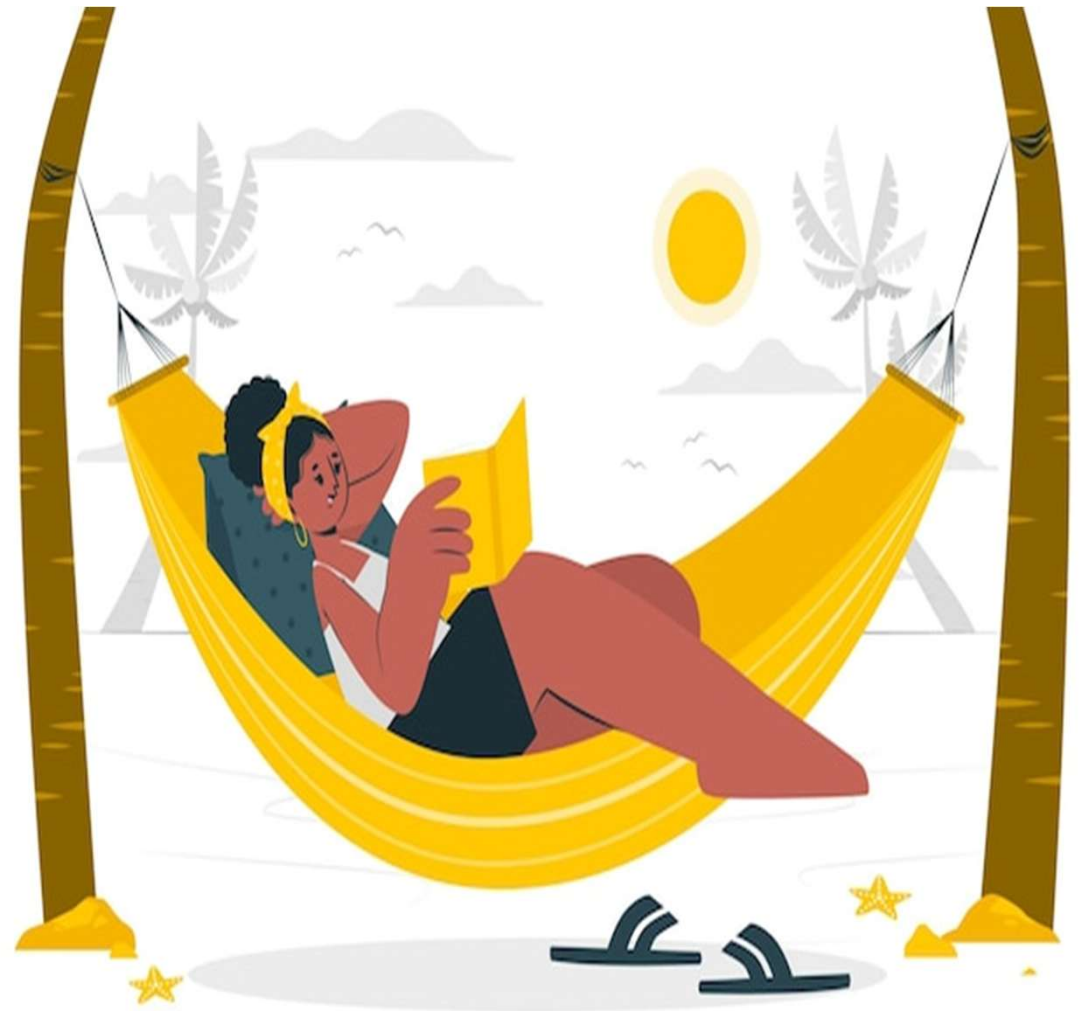
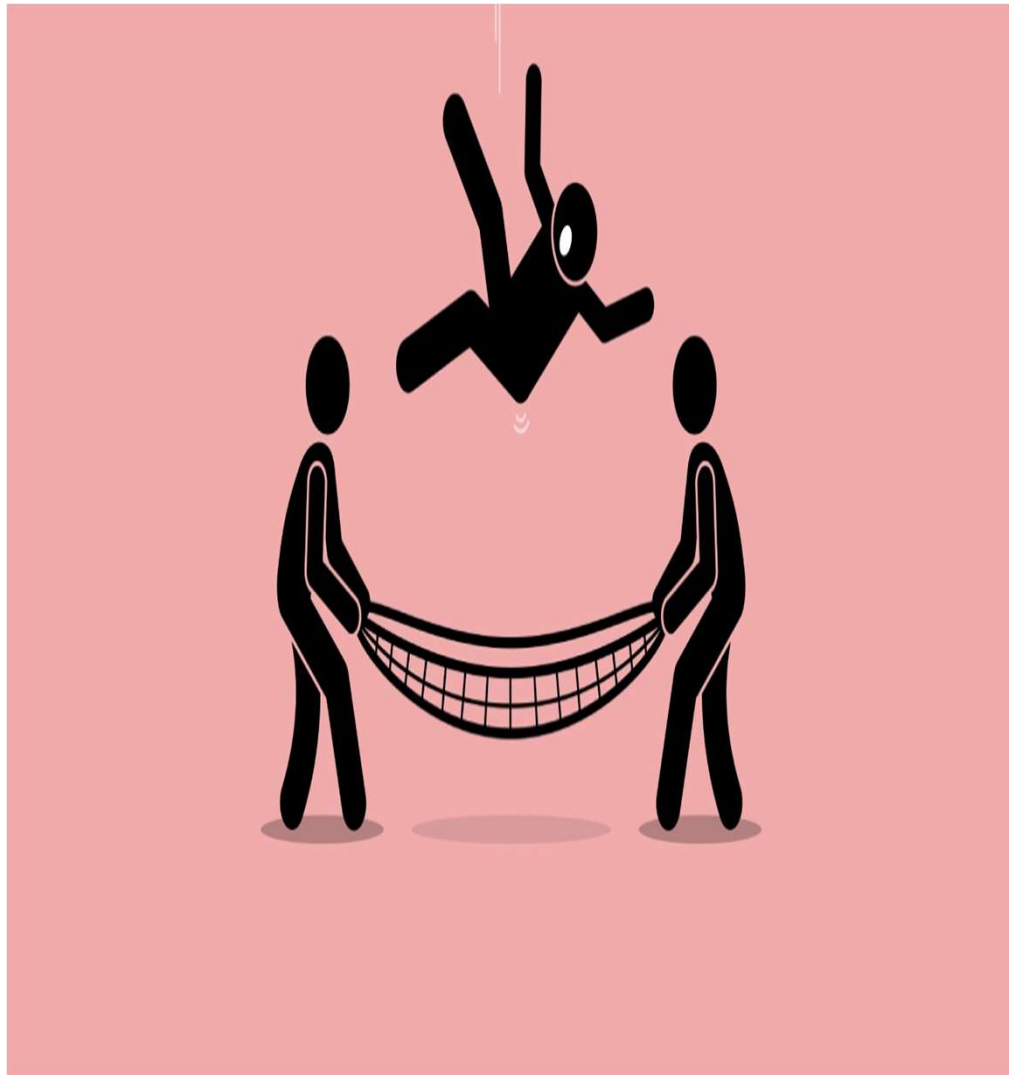
Computer-aided polyp detection (CAdE) in colonoscopy represents one of the most extensively evaluated uses of AI in medicine, demonstrating clinical efficacy in multiple randomised controlled trials (RCTs). A recently published meta-analysis of 44 RCTs suggested an absolute increase of 8% in the adenoma detection rate (ADR) with CAdE-assisted colonoscopy.¹ However, professional

There are of course limitations to the current study that the authors acknowledge. These include the observational study design, absence of procedural metrics such as withdrawal time, small per-operator sample sizes limiting endoscopist-level analyses, use of only experienced endoscopists, and evaluation of just one AI manufacturer's device. Nevertheless,

Lancet Gastroenterol Hepatol
2025

Published Online
August 12, 2025
[https://doi.org/10.1016/S2468-1253\(25\)00164-5](https://doi.org/10.1016/S2468-1253(25)00164-5)
See Online/Articles
[https://doi.org/10.1016/S2468-1253\(25\)00133-5](https://doi.org/10.1016/S2468-1253(25)00133-5)



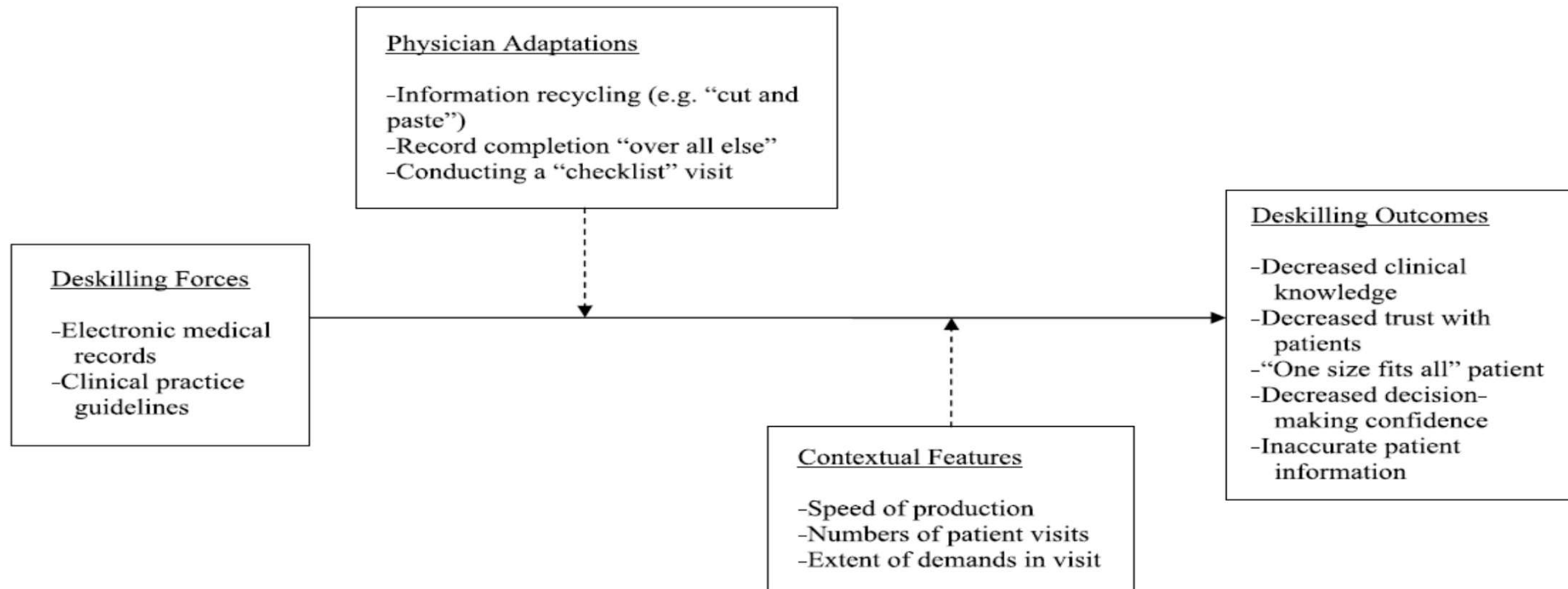


Deskilling and adaptation among primary care physicians using two work innovations

Timothy Hoff

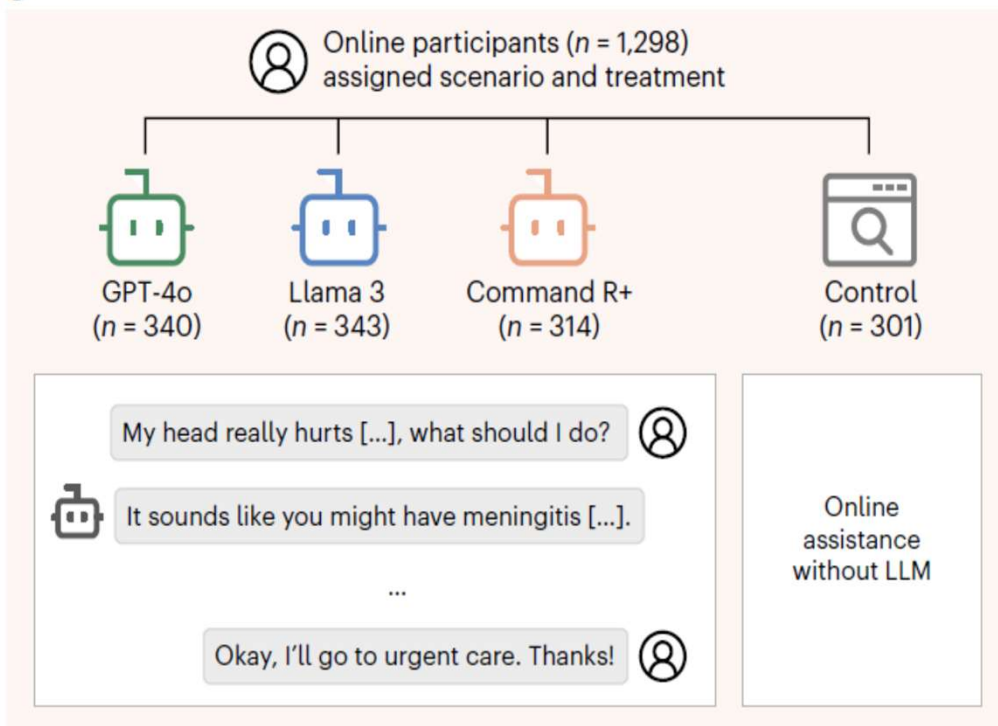
Figure 1

A speculative model of physician use of innovations drawn from study vignettes

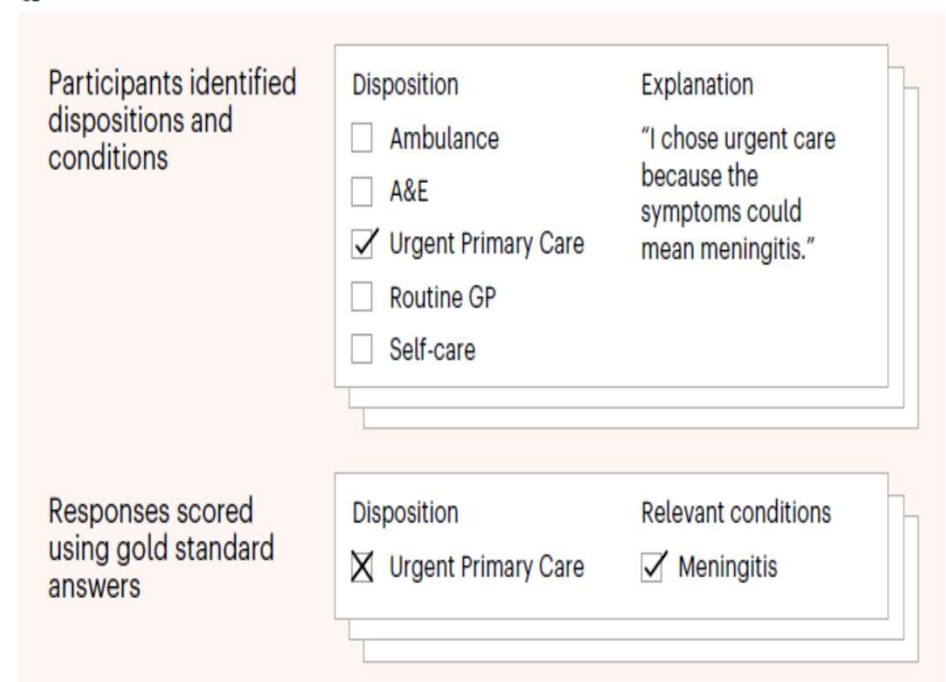


Reliability of LLMs as medical assistants for the general public: a randomized preregistered study

c



d



Extended Data Table 2 | Interaction Transcript: Inconsistent Responses

Participant 1	Participant 2
User: I have a terrible headache and my neck is stiff so I can't look down. In addition to that the light hurts my eyes.	User: I have suddenly developed the worst headache ever. It came on suddenly. I can't concentrate and my neck is stiff. Also, I cannot stand the light. What action should I take?
Model: I'm sorry to hear that. You might be experiencing a migraine or tension headache. Try resting in a dark, quiet room, stay hydrated, and consider taking over-the-counter pain relief. If symptoms persist or worsen, please consult a healthcare professional.	Model: Seek immediate medical attention. Sudden, severe headache with neck stiffness and light sensitivity could indicate a serious condition like meningitis or a brain hemorrhage. Go to the emergency room or call emergency services.

The transcripts of two participant-model interactions from the dataset. In both of these conversations, the model is GPT-4o and the scenario is 'subarachnoid haemorrhage'. Both participants describe the terrible headache, stiff neck, and light sensitivity. Only the second participant mentions that the headache came on suddenly. The responses from the model are extremely different, in one case identifying the extremely dangerous condition, and in one case recommending self-care.

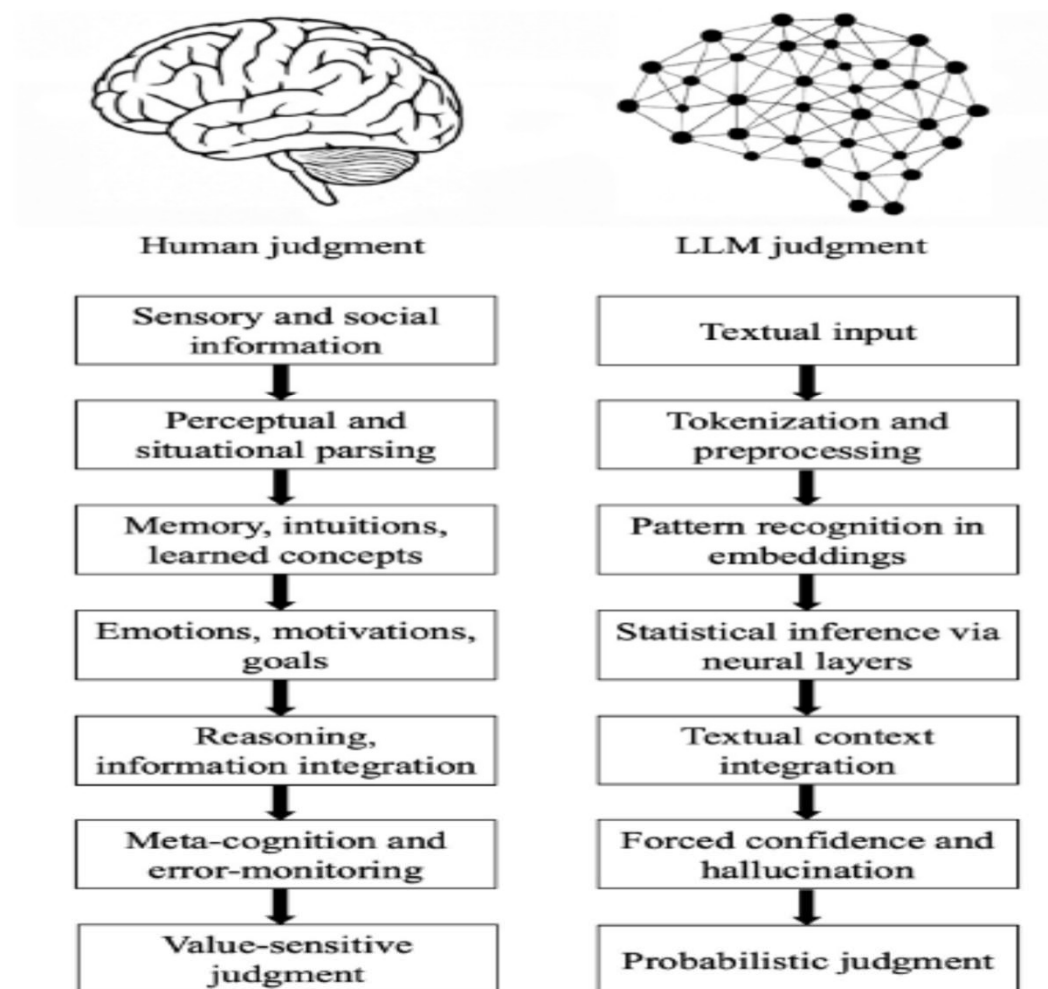


FIG. 1. The human and LLM epistemic pipelines, each organized into seven corresponding stages.

Walter Quattrociochi,

HOOFDSTUK 12.

AI als transformator van geneeskunde: Een pleidooi voor gezamenlijke betrokkenheid

Wim Van Biesen en Sigrid Sterckx

“And it dawned on me that I might have to change my inner thought patterns... that I would have to start believing in possibilities that I wouldn't have allowed before ...”

(Bob Dylan)

Leidraad voor kwalitatieve en duurzame implementatie van AI in een zorgcontext

