

PERSPECTIVE

AI IN MEDICINE

Will Autonomous AI Exceed AI-Aided Physicians as the Best Medical Care?

Ezekiel J. Emanuel, MD, PhD; Abe Baker-Butler, BA; Neal Khosla, MS; Vinod Khosla, MS, MBA

The prevailing view of artificial intelligence (AI) in medicine is that it will support physician-led care. The American Medical Association regularly calls AI augmented intelligence to focus on **AI's assistive role**. Similarly, the **American College of Physicians** argues that AI "should be limited

to a supportive role in clinical decision-making" and "should not replace physician decision-making." In *A Giant Leap: How AI Is Transforming Healthcare and What That Means for Our Future*, Wachter¹ argues that the highest tier of care will be AI-aided physicians, whereas AI-only care will be medicine's "economy class."

We disagree. In cognitive medical functions, AI-alone medical care is likely to be better than physician-only or physician-AI hybrid care. Large language models (LLMs) were only publicly introduced in November 2022, and already generative AI rivals or outperforms licensed physicians at 5 fundamental cognitive medical tasks: (1) eliciting medically relevant information; (2) establishing a differential diagnosis; (3) specifying diagnostic testing; (4) prescribing guideline-concordant treatment; and (5) managing chronic diseases. The gap between physicians' and LLMs' performance will likely widen because AI is rapidly improving, whereas physicians' skills are threatened by AI-induced deskilling.^{2,3}

Data from medicine and other fields suggest that when AI-alone performance is consistently superior to human-alone performance, AI alone surpasses human-AI hybrids. Paradoxically, hybrid care in which humans are in (or on) the loop to correct AI errors is likely to worsen rather than improve AI performance. Review of all published articles on AI in medicine since January 1, 2024, shows that medicine is rapidly approaching the transition point at which AI alone will exceed physicians and physician-AI hybrids in providing the best care at 5 fundamental cognitive medical tasks.

When Does AI Alone Exceed Physicians?

First, AI gathers patient information as effectively as, if not more effectively than, physicians (eTable 1 in the [Supplement](#)). Using 159 objective structured clinical examination simulated case scenarios covering multiple specialties, physicians judged Google's Articulate Medical Intelligence Explorer as significantly better than physicians at eliciting patient actors' complaints (97% vs 50% favorable), systems review (88% vs 35%), medical history (85% vs 50%), family history (50% vs 21%), and medication history (68% vs 45%) ($P < .001$ for all comparisons).

Second, many LLMs, including ChatGPT o3, Microsoft AI Diagnostic Orchestrator, and Google's Articulate Medical Intelligence Explorer, produce more accurate diagnoses than physicians (eTable 1 in the [Supplement](#)). ChatGPT o3 ranked the final diagnosis first in

60% of 377 real-world complex cases, whereas 20 internal medicine physicians did so in only 15.9% of a 302-case subset.

Third, AI can more precisely select tests to establish a definitive diagnosis than physicians can (eTable 1 in the [Supplement](#)). When instructed to stay within an \$8000 budget on 56 real-world complex cases, Microsoft AI Diagnostic Orchestrator reached the correct final diagnosis 4.02 times more frequently at 19.1% lower cost (\$2396 vs \$2963) than physicians without access to colleagues, textbooks, or the internet.

Fourth, compared with physicians, AI alone excels at recommending guideline-concordant treatments (eTable 1 in the [Supplement](#)). In the Google Articulate Medical Intelligence Explorer study, AI prescribed more appropriate treatment than licensed primary care physicians (90% vs 37% favorable, respectively; $P < .001$). Similarly, Cedars-Sinai's AI provided optimal treatment recommendations in 77.1% (95% CI, 72.7%-80.9%) of 461 real patient cases, whereas physicians did so in only 67.1% of cases (95% CI, 62.9%-71.1%).

Fifth, at least for hyperlipidemia, osteoarthritis, diabetes, and breast cancer, AI is better at chronic disease management, with the ability to monitor patients and adjust medications more frequently and accurately than physicians (eTable 1 in the [Supplement](#)). For instance, a 2023 randomized clinical trial found that 16 patients with type 2 diabetes who discussed their clinical data with a voice-based AI that provided them with updated insulin dosing instructions after each conversation had shorter time to optimal insulin dose (median, 15 vs >56 days; $P = .006$), greater insulin adherence (83% vs 50%; $P = .01$), and lower diabetes-related emotional distress (difference, -3.6 points; $P = .03$) than 16 patients who had standard physician-titrated insulin augmented by automated daily voice reminders to log insulin.

Overall, the preponderance of studies published since 2024 (eTable 1 in the [Supplement](#)) suggests AI alone is currently equivalent or superior to humans alone at 5 core cognitive medical tasks. More important, LLMs are rapidly improving. Most studies do not use **reasoning models**, a newer class of LLMs introduced in late 2024 that use optimized problem-solving methods. Reasoning models exceed nonreasoning models by greater amounts at cognitively demanding tasks, such as clinical care.

When Do Humans Alone Equal or Exceed AI Alone?

Excluding radiology, a domain in which AI alone does not reliably exceed physicians, only 9 studies published since January 1, 2024, concluded that physicians alone are equivalent or superior to AI alone at 5 fundamental cognitive medical tasks (eTable 1 in the [Supplement](#)).⁴⁻⁶ Most exclude the best AI models, are outdated, or have methodological flaws.

One study showed humans alone exceeding AI alone at differential diagnosis, test ordering, and prescribing guideline-concordant treatment in 80 emergency and intensive care cases (Hager et al in eTable 1 in the [Supplement](#)). However, the study excluded the best AI models—OpenAI's and Google's LLMs—underestimating AI's abilities. Similarly, an 83-study meta-analysis claimed no significant difference between AI and all physicians ($P = .10$) and worse performance by AI than expert physicians ($P = .007$). But 80 of the 83 studies used outdated versions of ChatGPT, likely failing to accurately represent AI's current capabilities.

Overall, although many of the studies indicating that humans alone exceed AI alone are flawed, some evidence suggests humans alone currently match or exceed the best AI-alone performance at particular cognitive medical tasks, such as diagnosing paroxysmal atrial fibrillation and cases with atypical features (eTable 1 in the [Supplement](#)).

Studies comparing AI and physicians often use LLMs out of the box without customized prompts and agent architectures (eTable 1 in the [Supplement](#)). Such customization includes task-specific safety guardrails and structured logic that instructs LLMs on sequences of clinical steps for particular scenarios and the LLM's role (eg, "You are a primary care physician seeking to stabilize chronic disease"). Large language models with customized architectures markedly outperform naive architectures, generating more accurate diagnoses, test ordering, and guideline-concordant treatment and reducing safety incidents.⁷⁻¹⁰ Thus, existing studies may systematically underestimate AI performance (eTable 1 in the [Supplement](#)).

How Does AI Alone Compare With Physician-AI Hybrids?

Counterintuitively, when AI alone is better at a task than humans alone, human-AI hybrids actually degrade performance compared with AI alone (eTable 2 in the [Supplement](#)). Thus, as AI improves, having humans in the loop will likely worsen patient care.

A 2024 meta-analysis of 106 experiments with human-AI hybrids—at least 20 were in medicine—found that when humans alone were better at a task than AI alone, human-AI hybrids improved human performance ($g = 0.46$; 95% CI, 0.28-0.66; $t_{104} = 5.06$; 2-tailed $P < .001$) (eTable 2 in the [Supplement](#)). Conversely, when AI alone was better than humans alone, deploying human-AI hybrids significantly worsened performance compared with AI alone ($g = -0.54$; 95% CI, -0.71 to -0.37; $t_{104} = -6.20$; 2-tailed $P < .001$). The meta-analysis found that for content-creation tasks specifically, hybrids exceed AI alone. However, this finding included all content-creation tasks regardless of whether humans alone or AI alone was better. Consequently, this finding has limited relevance to whether AI alone exceeds hybrids when AI alone exceeds humans alone. A 2025 review of 52 clinical studies found that physician-AI hybrids "had a lower reliability level than the best performer [of either AI alone or human alone]" ($\beta = -0.019$; SE = 0.007; $t = -2.872$; $P = .006$) (eTable 2 in the [Supplement](#)). The authors concluded that physician-AI hybrids "neither outperformed medical AI alone nor surpassed the best of clinicians or medical AI alone."

Similarly, using real patient cases, OpenAI's ChatGPT-4 alone had a median diagnostic reasoning score of 92% (95% CI, 76%-100%; IQR, 82%-97%), whereas physicians using ChatGPT-4 scored 76% (IQR, 66%-87%). A study found that on the same cases, o1 preview alone, a more advanced OpenAI reasoning model, scored even

higher, at 97% (IQR, 95%-100%) (Brodeur et al in eTable 2 in the [Supplement](#)). However, they did not conduct a direct comparison using o1-preview hybrids.

A few studies concluded that physician-AI hybrids exceed AI alone (eTable 2 in the [Supplement](#)). One study found that a "human-AI hybrid team outperform[ed] both [human and AI] agents" at identifying colon lesions (Reverberi et al in eTable 2 in the [Supplement](#)). Another study found that physician-AI hybrids exceed AI alone at diagnosing paroxysmal atrial fibrillation (Zhang et al in eTable 2 in the [Supplement](#)). In both studies, AI alone did not clearly exceed physician-alone performance.

A 2025 study reported that physician-AI hybrids exceeded AI alone at differential diagnosis (Zoller et al in eTable 2 in the [Supplement](#)). However, in this study, AI—not humans—controlled the hybrid's final diagnostic decisions, which is not the decision-making model of physician in the loop advocated by the [American Medical Association](#), [American College of Physicians](#), and hybrid proponents.¹ Another study tested an AI-controlled hybrid and a physician-controlled hybrid at tumor identification (Ruffe et al in eTable 2 in the [Supplement](#)). The AI-controlled hybrid exceeded AI alone, whereas the physician-controlled hybrid underperformed AI alone, suggesting that when AI alone exceeds hybrids, it is not merely a training issue. In AI-alone and AI-controlled hybrids, physicians' ability to introduce error is greatly diminished compared with their ability to do so with physician-controlled hybrids.

Artificial intelligence alone exceeds human-AI hybrids beyond medicine. In chess, in 1997 [Deep Blue](#) beat Garry Kasparov, the world champion. The last time a human alone beat AI alone at chess was 2005.¹¹ By 2017, [AI alone surpassed even the most skilled human-AI hybrids](#) in chess. Despite a 12-year period of hybrid dominance, as AI improved AI alone eventually consistently dominated. Similar findings that AI alone exceeds human-AI hybrids exist in other fields (eTable 2 in the [Supplement](#)). Medicine is more complex than chess. Physicians' choices are less clearly defined and less certainly linked to outcomes. However, the evolution seems to be that when AI alone completely exceeds humans alone, then AI alone eventually exceeds human-AI hybrids because humans in the loop degrade AI performance.

Why Does AI Alone Exceed Physician-AI Hybrids?

Although integrating humans in the loop catches some AI errors and adds insights, it also introduces errors when humans contribute errors and incorrectly overrule AI.¹² Humans perform poorly at deciding when to trust AI, especially when AI exceeds them.^{13,14} This poor human decision-making is partly driven by algorithm aversion (distrust of AI). Algorithm aversion is evident in 75% of studies on algorithmic advice and persists even when humans know they are less accurate than AI.^{15,16} In high-stakes tasks with high-expertise decision-makers, such as expert physicians, algorithm aversion seems consistently greatest.^{14,17} In medicine, algorithm aversion causes significant inaccuracy because humans have lower medical accuracy than AI and struggle at integrating cross-specialty medical knowledge, which AI does easily (eTable 1 in the [Supplement](#)).¹⁸ As AI improves and exceeds humans, physicians' error-catching and insight-adding benefits dwindle, whereas physician-introduced errors remain constant or increase. This problem may worsen because physicians are likely to become increasingly deskilled. Artificial intelligence generates some upskilling benefit for new physicians

but induces an overall physician [deskilling](#).^{3,19} Such deskilling seems inevitable without intentional upskilling programs or AI nonuse.

Caveats and Additional Research

Most studies comparing AI alone with physicians alone and physician-AI hybrids are simulations of discrete cognitive medical tasks, not analyses of real clinical interactions. More assessments of AI in real-life clinical encounters are essential and seem to be expanding.^{20,21} Liability, regulatory, and reimbursement barriers limit studies examining AI alone in real clinical interactions, which impedes accurate assessments of AI-alone real-life clinical performance.

Adversarial stress testing suggests that LLMs are still brittle in some cases.²² Bean et al²³ found that “transmission of information between the LLM and the user...[is] a particular point of failure.” This failure is not detected when text-based clinical vignettes are fed directly to LLMs, as they are in many simulated studies, which may overestimate AI clinical performance. Johri et al²⁴ proposed a framework to address this problem in future studies.

Simultaneously, more studies should compare AI alone with physician-AI hybrids. Physician-AI hybrids reliably exceed physicians alone, which has induced the erroneous assumption that physician-AI hybrids also exceed AI alone,¹³ causing studies to treat physician-AI hybrids as the criterion standard and to minimize comparisons with AI alone.²⁵ This method may obscure the extent and frequency to which AI alone exceeds physician-AI hybrids.

Artificial intelligence alone exceeding hybrids at cognitive medical tasks is necessary but insufficient for clinical implementation. Many, though not all, cognitive medical tasks are associated with physical procedures and examinations that robotics does not yet enable AI to perform autonomously. Surgeries, deliveries, interventional radiology, and colonoscopies are a few examples. Accordingly, in many workflows autonomous AI for cognitive tasks must be integrated with physicians performing physical tasks. There are also

engineering challenges and high start-up costs associated with AI clinical implementation, which could slow AI deployment.

Also, AI fails differently than physicians. Tail risks from sources, such as internet loss, cyberattacks, and hallucinations, will be greater with autonomous AI than hybrids and must be weighed against autonomous AI's higher diagnostic and treatment accuracy (eTables 1 and 2 in the [Supplement](#)).

One study showed that the design of hybrid workflows can affect hybrid performance, and there has been insufficient testing on which AI implementations are optimal (Agarwal et al in eTable 2 in the [Supplement](#)). However, as long as a hybrid is human controlled, workflow design improvements seem unlikely to eliminate the human-introduced error present in hybrids and absent in AI alone.

Finally, publication bias exists. A 2025 study (Liu et al in eTable 2 in the [Supplement](#)) noted that “failures or negative outcomes of medical AI are often not formally reported,” although this is changing as AI failures become more rare and notable. Additionally, the best AI-alone algorithms in other areas, such as finance, are rarely reported because firms seek to preserve advantages of superior AI. These biases are likely to distort our understanding.

Conclusions

The prevailing claim that physician-AI hybrids, such as AI decision support for physicians, will be the highest standard of care is unproven. That in the near future AI-alone may provide better patient care than physicians or physician-controlled hybrids at 5 fundamental cognitive medical tasks is unsettling but seems probable.

Even when AI alone is superior, significant barriers to implementation remain. Nonetheless, superior autonomous AI will likely be ready to be deployed for real-world cognitive medical tasks in some, maybe many, workflows by 2030. Consequently, physicians, policymakers, and others need to urgently devise approaches to workflow, liability, regulation, reimbursement, and medical education.

ARTICLE INFORMATION

Author Affiliations: Department of Medical Ethics and Health Policy, University of Pennsylvania, Philadelphia (Emanuel, Baker-Butler); Curai Health, San Francisco, California (N. Khosla); Khosla Ventures, Menlo Park, California (V. Khosla).

Corresponding Author: Ezekiel J. Emanuel, MD, PhD, Department of Medical Ethics and Health Policy, University of Pennsylvania, 3600 Civic Center Blvd, 8th Floor, Philadelphia, PA 19104 (zemanuel@upenn.edu).

Published Online: August 17, 2026.
doi:[10.1001/jama.2026.15380](https://doi.org/10.1001/jama.2026.15380)

Author Contributions: Dr Emanuel and Mr Baker-Butler served as co-first authors.

Conflict of Interest Disclosures: Dr Emanuel reported honoraria from the 10th Annual CVS Accountable Care Symposium, Building Together, California Orthopaedic Association, HealthCare Royalty Partners, RAISE Symposium, SpringTide, Avalon Healthcare Solutions, Massachusetts Association of Health Plans, UCSF Department of Medicine Medical Grand Rounds, and Munk Debates; grants from the Bergen Center for Ethics and Priority Setting in Health, University of Bergen, and Bill & Melinda Gates Foundation outside the submitted work; serving in an advisory capacity for

Daymark Health, University of Pennsylvania Parity Center, Peterson Center on Healthcare, JSL Health Capital, Notable Health, Smirk Health, Dendro Technologies (CalmiGo), FeelBetter, Clarify Health Solutions, and Cellares; and consulting for Korro. Mr N. Khosla reported serving as chief executive officer for Curai Health; being a stockholder in Alphabet outside the submitted work; holding a patent for ensemble machine learning systems and methods; having a patent pending for heterogeneous multiagent safety verification architecture using parallel deterministic and neural validators for cross-checking autonomous AI reasoning pipelines; and holding a patent for systems and methods for responding to health care inquiries. Mr V. Khosla reported investing in OpenAI, Curai Health, and Limbic. No other disclosures were reported.

REFERENCES

1. Wachter R. *A Giant Leap: How AI Is Transforming Healthcare and What That Means for Our Future*. Portfolio; 2026:261.
2. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi:[10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2)

3. Budzyń K, Romańczyk M, Kitala D, et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *Lancet Gastroenterol Hepatol*. 2025;10(10):896-903. doi:[10.1016/S2468-1253\(25\)00133-5](https://doi.org/10.1016/S2468-1253(25)00133-5)
4. Horiuchi D, Tatekawa H, Oura T, et al. ChatGPT's diagnostic performance based on textual vs visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. *Eur Radiol*. 2025;35(1):506-516. doi:[10.1007/s00330-024-10902-5](https://doi.org/10.1007/s00330-024-10902-5)
5. Resch D, Lo Gullo R, Teuwen J, et al. AI-enhanced mammography with digital breast tomosynthesis for breast cancer detection: clinical value and comparison with human performance. *Radiol Imaging Cancer*. 2024;6(4):e230149. doi:[10.1148/rycan.230149](https://doi.org/10.1148/rycan.230149)
6. Le Guellec B, Bruge C, Chalhoub N, et al; ARIANES Investigators. Comparison between multimodal foundation models and radiologists for the diagnosis of challenging neuroradiology cases with text and images. *Diagn Interv Imaging*. 2025; 106(10):345-352. doi:[10.1016/j.diii.2025.04.006](https://doi.org/10.1016/j.diii.2025.04.006)
7. Hassanein FEA, Ahmed Y, Maher S, Barbary AE, Abou-Bakr A. Prompt-dependent performance of multimodal AI model in oral diagnosis:

a comprehensive analysis of accuracy, narrative quality, calibration, and latency versus human experts. *Sci Rep*. 2025;15(1):37932. doi:10.1038/s41598-025-22979-z

8. Chai M, Zomorodi AR. Prompt engineering does not universally improve large language model performance across clinical decision-making tasks. *arXiv*. Preprint posted online December 28, 2025. doi:10.48550/arXiv.2512.22966
9. Wang L, Chen X, Deng X, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *NPJ Digit Med*. 2024;7(1):41. doi:10.1038/s41746-024-01029-4
10. Esmailzadeh P. Ethical implications of using general-purpose LLMs in clinical settings: a comparative analysis of prompt engineering strategies and their impact on patient safety. *BMC Med Inform Decis Mak*. 2025;25(1):342. doi:10.1186/s12911-025-03182-6
11. History of chess matches between human and computer. Deep Chess. Published September 24, 2018. Accessed June 25, 2026. <https://deepchess.org/blog/f/human-against-computer-chess-matches>
12. Ong KT, Seo J, Kim H, et al. Success and failure of human-AI collaboration in clinical reasoning: an experimental study on challenging real-world cases. *Int J Med Inform*. 2026;211:106342. doi:10.1016/j.ijmedinf.2026.106342
13. Vaccaro M, Almaatouq A, Malone T. When combinations of humans and AI are useful: a systematic review and meta-analysis. *Nat Hum Behav*. 2024;8(12):2293-2303. doi:10.1038/s41562-024-02024-1
14. Filiz I, Judek JR, Lorenz M, Spiwoks M. The extent of algorithm aversion in decision-making situations with varying gravity. *PLoS One*. 2023;18(2):e0278751. doi:10.1371/journal.pone.0278751
15. Kaufmann E, Chacon A, Kausel EE, Herrera N, Reyes T. Task-specific algorithm advice acceptance: a review and directions for future research. *Data Inf Manag*. 2023;7(4):100040. doi:10.1016/j.dim.2023.100040
16. Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J Exp Psychol Gen*. 2015;144(1):114-126. doi:10.1037/xge0000033
17. Logg JM, Minson JA, Moore DA. Algorithm appreciation: people prefer algorithmic to human judgment. *Organ Behav Hum Decis Process*. 2019;151:90-103. doi:10.1016/j.obhdp.2018.12.005
18. Hashem A, Chi MTH, Friedman CP. Medical errors as a result of specialization. *J Biomed Inform*. 2003;36(1-2):61-69. doi:10.1016/S1532-0464(03)00057-1
19. Gokkurt Yilmaz BN, Ozbey F, Yilmaz BE. Effect of artificial intelligence-assisted personalized feedback on radiographic diagnostic performance of dental students: a controlled study. *BMC Med Educ*. 2025;25(1):1403. doi:10.1186/s12909-025-07875-4
20. Mello MM. Utah's experiment with AI-driven prescription renewals. *JAMA Health Forum*. 2026;7(3):e261001. doi:10.1001/jamahealthforum.2026.1001
21. Schaeckermann M, Chen C. Collaborating on a nationwide randomized study of AI in real-world virtual care. Google Research. Published February 3, 2026. Accessed June 24, 2026. <https://research.google/blog/collaborating-on-a-nationwide-randomized-study-of-ai-in-real-world-virtual-care/>
22. Gu Y, Fu J, Liu X, et al. Evaluating the robustness and readiness of large frontier models in health AI applications. *Nat Med*. Published online June 26, 2026. doi:10.1038/s41591-026-04501-8
23. Bean AM, Payne RE, Parsons G, et al. Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nat Med*. 2026;32(2):609-615. doi:10.1038/s41591-025-04074-y
24. Johri S, Jeong J, Tran BA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat Med*. 2025;31(1):77-86. doi:10.1038/s41591-024-03328-5
25. Lång K, Josefsson V, Larsson AM, et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol*. 2023;24(8):936-944. doi:10.1016/S1470-2045(23)00298-X