Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.263928

ARTIFICIAL INTELLIGENCE AND INFORMATICS

COMMENTARY



# Beyond autonomy: why medicine needs artificial intelligence teammates, not artificial intelligence doctors

Gorkem Durak
Alpay Medetalibeyoglu
Vedat Cicek
Elif Keles
Ulas Bagci

Northwestern University, Machine and Hybrid Intelligence Lab, Department of Radiology, Chicago, United States of America

The medical artificial intelligence (AI) field is racing toward fully autonomous systems that promise to diagnose, treat, and manage patients independently. Yet, real-world deployments tell a different story. Although systems such as CE-verified ChestLink[1] (by Oxipit)—which achieves 99.9% sensitivity in chest X-ray screening—and PathFinder[2]—which outperforms average pathologists in melanoma diagnosis—demonstrate impressive performance, the most sophisticated attempts at clinical autonomy reveal fundamental limitations. Despite outperforming primary care physicians in diagnostic dialogue, Google's Articulate Medical Intelligence Explorer (AMIE)[3] cannot practice medicine. Stanford's Optimization Paradox study[4] demonstrates that combining the best individual AI agents paradoxically creates worse clinical systems, with diagnostic accuracy dropping from 77.4% to 67.7%. Evidence from 2023–2025 suggests the real paradigm shift is not replacing clinicians with autonomous AI agents but developing sophisticated AI teammates that augment human expertise. Our perspective postulates that success requires abandoning fantasy of fully autonomous medical AI and embracing collaborative intelligence that preserves clinical judgment while alleviating cognitive burden and physician burnout.

In January 2024, Google DeepMind introduced AMIE,[3] a system that surpassed primary care physicians on 30 of 32 diagnostic conversation criteria. The headlines were sensational: "AI Beats Doctors at Diagnosis." However, buried within the paper was an important caveat: AMIE operated in a controlled, text-only setting with unlimited time, no real patients, and no actual clinical responsibility. Despite its impressive results, AMIE cannot prescribe medications, order tests, or take responsibility for its decisions. This technically advanced system is, at its core, just an elaborate chatbot.

This gap between algorithmic success and real-world clinical practice highlights a broader challenge in medical AI. Although technology companies promote autonomous agents that will independently diagnose and treat patients, the gap between computational capability and clinical utility remains substantial. The narrative of inevitable progress—from rule-based systems to machine learning, then to generative AI, and ultimately to fully autonomous agents—ignores growing evidence that autonomy itself may not be the appropriate goal. The true paradigm shift in medicine is not just moving from automation to autonomy but evolving from isolated tools to collaborative partners. This difference is not just a matter of words—it is essential for creating AI systems that clinicians will trust, and that will truly benefit patients.

Agentic AI marks a real breakthrough in computing power. Unlike traditional AI, which responds to specific inputs with fixed outputs, agentic systems have three main traits: goal-driven independence (pursuing objectives independently), complex reasoning and planning (breaking down ambiguous tasks into clear steps), and flexible adaptation (learning from environmental feedback). These systems use large language models (LLMs) as reasoning engines, coordinating specialized tools to perform complex, multi-step clinical tasks.

The promise appears compelling. ChestLink,[1] the world's first CE-certified autonomous medical imaging AI, analyzes chest X-rays without radiologist involvement, achieving 99.9% sensitivity while autonomously reporting 36.4% of normal cases. PathFinder,[2] a multi-agent

**Corresponding author:** Gorkem Durak

**E-mail:** gorkem.durak@northwestern.edu

Received 09 February 2026; revision requested 13 February 2026; accepted 16 February 2026.

Epub: 13.04.2026

DOI: 10.4274/dir.2026.263928

You may cite this article as: Durak G, Medetalibeyoglu A, Cicek V, Keles E, Bagci U. Beyond autonomy: why medicine needs artificial intelligence teammates, not artificial intelligence doctors. *Diagn Interv Radiol*. XXX April 2026 DOI: 10.4274/dir.2026.263928 [Epub Ahead of Print].

system for melanoma diagnosis published in 2025, achieved 74% accuracy—outperforming average pathologists by 9%. These successes suggest that autonomous AI could address healthcare's most pressing challenges: diagnostic variability, workflow inefficiencies, and access disparities.

Yet, the most revealing finding comes from a 2025 study on multi-agent diagnostic systems. Stanford researchers developed a Best of Breed system,[4] (also known as the Optimization Paradox), by selecting the top-performing AI agents for each diagnostic subtask. Logic suggested this dream team of top AI agents would excel. Instead, it catastrophically failed. Despite superior individual components (85.5% accuracy in lab interpretation), the combined system achieved only 67.7% diagnostic accuracy—compared with 77.4% for a system built with less capable but better-integrated agents, exposing a fundamental truth: medical AI system performance is not determined by the excellence of individual components but by the quality of connections between them. The interfaces—how agents communicate, share data, and coordinate actions—matter more than the agents themselves. This finding challenges the premise of autonomous medical AI, revealing that isolated optimization can introduce dangerous brittleness in complex clinical workflows.

Three landmark studies from 2023–2025 illuminate the chasm between autonomous AI's promise and clinical reality:

**1.** Google's AMIE's Brilliance Without Practice Rights:[3] This system achieved superior diagnostic performance in standardized patient scenarios, demonstrating remarkable clinical reasoning capabilities. However, the system exists in a regulatory and practical vacuum. It cannot access real patient data, integrate with electronic health records, or assume clinical responsibility. The authors acknowledge these limitations, noting that "deployment in real clinical settings would require addressing numerous practical, ethical, and regulatory challenges." This system exemplifies the paradox of modern medical AI: systems sophisticated enough to outperform doctors in tests yet unable to practice medicine in real life.

**2.** The False Conflict Error/Automation Bias: A pivotal 2024 Nature Communications study[5] revealed that high-performing clinicians' diagnostic accuracy decreased by almost 50% when using AI support. When AI and experts disagreed, experts deferred to the machine—even when their initial judgment was correct. This "false conflict error" demonstrates that AI does not just fail in isolation; it can actively degrade human performance by undermining clinical confidence. The study's authors warned that "AI assistance can paradoxically harm the performance of highly skilled decision-makers."

**3.** Microsoft Azure's Tumor Board Orchestrator:[6] In partnership with Johns Hopkins, Stanford, and other university hospitals, this multi-agent system was designed to streamline tumor board preparation, reducing it from multiple hours to minutes by coordinating specialized agents for radiology, pathology, and clinical data synthesis. Despite successes in workflow acceleration, multimodal data fusion, and conflict detection, the findings cannot eliminate the need for extensive human oversight and fall short of true autonomy, with clinicians reporting that "the system augments but cannot replace multidisciplinary discussion."

All these failures follow a pattern: autonomous systems excel at narrow tasks but struggle with integration, context, and the messy realities of clinical practice. They cannot navigate the implicit knowledge, edge cases, and value judgments that define real medical care.

The evidence points toward a different paradigm: LLM-based human–agent systems, in which AI functions as an intelligent teammate rather than an autonomous practitioner. This structure, described in a 2025 paper,[7] is gaining traction because it leverages each party's strengths: AI handles rapid data synthesis, pattern recognition across vast datasets, and repetitive documentation tasks, while the human clinician provides oversight, contextual understanding, ethical judgment, and accountability.

A real-world randomized controlled trial published in NEJM AI demonstrated that ambient AI scribes effectively decreased documentation time, reduced clinicians' work exhaustion, and improved interpersonal engagement.[8] Abridge is one example of a leading ambient scribe that has undergone considerable clinical validation, providing compelling quantitative data: a 2025 study published in JAMIA Open[9] found that clinicians using Abridge experienced a 43% reduction in documentation time, and they were seven times more likely to describe their workflow as easy and five times more likely to believe they could complete their notes before seeing the next patient. This demonstrates the power of a collaborative approach where AI eliminates administrative burden rather than completely replacing physicians.[10] Microsoft Azure's agent orchestrator for tumor boards, as another example, does not autonomously decide treatment but reduces data preparation from hours to minutes in Stanford experiments,[11] freeing oncologists for higher-level strategic discussion. Nuance's Dragon Ambient eXperience (DAX) Copilot is one of the most prominent tools functioning as a background documentation assistant, trained on over 15 million clinical encounters to produce high-quality notes with minimal clinician input.[12] The platform has recently evolved into the unified Microsoft Dragon Copilot, which integrates DAX's ambient listening capabilities with the established voice dictation technology of Dragon Medical One, enabling hands-free note creation and editing. It succeeds not only because it performs medical tasks up to a certain level but also because it liberates physicians from the keyboard, allowing them to focus on patients. In all these studies, the key insight is to optimize for human–AI collaboration rather than AI independence.

These collaborative models address the fundamental limitations of agentic AI. Humans provide what machines cannot: empathy in difficult conversations, judgment in unprecedented cases, and accountability when decisions affect lives. AI provides capabilities that humans struggle with: perfect recall across thousands of guidelines, pattern detection in massive datasets, and sustained attention to routine tasks. Collaborative medical AI requires fundamental shifts in development, regulation, and implementation. Technically, we must prioritize interface design over algorithm optimization. The Optimization Paradox teaches that brilliant components mean nothing without seamless integration. Success demands end-to-end system validation, not isolated benchmarking.

Implementation strategy matters: autonomy is not a single endpoint but a spectrum, and its appropriateness is task dependent. For example, similar to ChestLink's success in autonomously reporting normal X-rays, low-risk, high-volume classification tasks are suitable for autonomy, whereas high-stakes, multifaceted diagnostic and treatment decisions are not. Regulatory frameworks need updating: the Food and Drug Administration's predetermined change control plans address adaptive algorithms but not autonomous clinical actions. We need new pathways that evaluate human–AI teams as integrated systems, measuring not just algorithmic accuracy but workflow integration, clinical utility,

and safety in collaborative contexts. The path forward is to abandon ambitious AI development but to redirect it.

Stop chasing the phantom of full autonomy; instead, start building AI teammates that augment human capabilities while preserving the judgment, empathy, and accountability that define good medicine. The paradigm shift is not from human to machine but from isolation to collaboration. Medicine needs AI teammates, not AI doctors—and recognizing this distinction will determine whether medical AI fulfills its promise or joins the graveyard of technologies that were brilliant in theory but useless in practice.

### Footnotes

### Conflict of interest disclosure

Ulas Bagci, PhD, serves as Section Editor for Diagnostic and Interventional Radiology. He had no involvement in the peer-review of this article and had no access to information regarding its peer review.

## References


1. Plesner LL, Müller FC, Nybing JD, et al. Autonomous chest radiograph reporting using AI: estimation of clinical impact. Radiology. 2023;307(3):e222268. [Crossref]
2. Ghezloo F, Seyfioğlu Saygın M, Soraki R, et al. PathFinder: a multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. arXiv preprint arXiv:2502.08916, 2025. [Crossref]
3. Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. Nature. 2025;642(8067):442-450. [Crossref]
4. Bedi S, Mlauzi I, Shin D, Koyejo S, Shah NH. The optimization paradox in clinical AI multi-agent systems. arXiv preprint arXiv:2506.06574, 2025. [Crossref]
5. Rosenbacke R, Melhus A, Stuckler D. False conflict and false confirmation errors are crucial components of AI accuracy in medical decision making. Nat Commun. 2024;15(1):6896. [Crossref]
6. Microsoft. Multi-agent orchestration platform for complex workflows like tumor boards. 2025. [Crossref]
7. Zou HP, Huang WC, Wu Y, et al. A call for collaborative intelligence: why human-agent systems should precede AI autonomy. arXiv preprint arXiv:2506.09420, 2025. [Crossref]
8. Afshar M, Baumann MR, Resnik F, et al. A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. NEJM AI. 2025;2(12). [Crossref].
9. Albrecht M, Shanks D, Shah T, et al. Enhancing clinical documentation with ambient artificial intelligence: a quality improvement survey assessing clinician perspectives on work burden, burnout, and job satisfaction. JAMIA Open. 2025;8(1):ooaf013. [Crossref]
10. Hudson TJ, Albrecht M, Smith TR, et al. Impact of ambient artificial intelligence documentation on cognitive load. Mayo Clin Proc Digit Health. 2025;3(1):100193. [Crossref]
11. Alvarez IG. The agentic AI assist Stanford University cancer care staR needed. 2025. [Crossref]
12. Kakaday R, Herrera EZ, Coskey O, Hertel AW, Kaiser P. The STREAMLINE pilot study on time reduction and efficiency in AI-mediated logging for improved note-taking experience. Appl Clin Inform. 2025;16(3):614-621. [Crossref]