Review of *The AI Revolution in Medicine: GPT-4 and Beyond*

Aiersilan Aizierjiang
alexandera@gwu.edu
The George Washington University

The release of GPT-3.5 in late 2022 marked a technological inflection point that many compare to the arrival of the internet or the personal computer, then there has been more robust and competitive versions such as GPT-4 being proposed. In *The AI Revolution in Medicine: GPT-4 and Beyond* Lee et al. [2023], authors Peter Lee, Carey Goldberg, and Isaac Kohane provide an early and optimistic exploration of how general-purpose artificial intelligence (also known as Artificial General Intelligence, AGI) might transform healthcare. Their central thesis is that Large Language Models (LLMs) have evolved beyond simple search engines or predictive text tools into entities capable of reasoning, empathy, and complex problem-solving. They argue that these models will not replace physicians but rather form a "symbiotic" partnership with them, acting as tireless medical residents capable of handling administrative burdens, suggesting diagnoses, and allowing human doctors to focus on patient connection. As a PhD student specializing in AI + Healthcare, I find their vision compelling yet requiring significant scrutiny regarding implementation and safety.

The authors distinguish their work by moving beyond theoretical pontification to provide vivid, real-time transcripts of interactions with GPT-4. These interactions serve to illustrate three core roles for AI in medicine: the *Torchbearer* for diagnostics, the *Ultimate Paperwork Shredder* for administration, and the *Empathic Consultant* for patient interaction. One of the most striking examples of the Torchbearer role involves a complex case of a newborn with ambiguous genitalia. Kohane, a pediatric endocrinologist, feeds the AI clinical details involving hypospadias and non-palpable gonads. The AI generates a differential diagnosis that rivals expert human analysis, correctly identifying a rare condition 11-beta-hydroxylase deficiency that affects fewer than one in 100,000 babies. This ability to synthesize vast medical knowledge and apply it to an N-of-1 case demonstrates the potential for AI to augment human decision-making in critical, high-stakes moments. It suggests a future where the "diagnostic odyssey" for rare disease patients could be

drastically shortened by an AI partner that "never forgets a journal article".

However, the authors argue that the most immediate impact will be in the unglamorous trenches of administration. They vividly describe the crushing burden of "pajama time", the hours doctors find themselves spending at night finishing paperwork. The book demonstrates GPT-4's ability to act as a "paperwork shredder" by processing a transcript of a patient visit such as the fictional "Dave Smith" appointment and instantly generating accurate medical encounter notes, insurance codes, and discharge summaries. Beyond text, the authors showcase the model's ability to reason across modalities, such as when asked to calculate an IV flow rate for a nurse. Not only does the AI perform the math correctly, but it also writes a functional JavaScript application on the fly to help the nurse perform future calculations. This illustrates a profound shift: the AI is not just a retriever of information but a generator of tools, fundamentally altering the workflow of clinical staff.

From an information policy perspective, the deployment of such powerful tools introduces complex challenges regarding equity, bias, and evaluation standards. The authors propose new frameworks for evaluating AI, moving from the standard "Trial" model (used for drugs) to models they term the "Trainee" (evaluating the AI like a medical resident) and the "Torchbearer" (evaluating it as a super-expert). While innovative, these frameworks must be scrutinized under the lens of Lessig's "code is law" Lessig [2009]. If GPT-4 becomes the standard "Trainee" in every clinic, its opaque algorithms effectively set the standard of care without legislative oversight. Moreover, applying Spinello's cyberethics framework Spinello [2010] reveals the moral hazard of establishing automated decision-making in life-critical scenarios. The authors' optimism about the "Doctor-Patient-AI Triad"where the AI acts as a third partner in the room must be weighed against the risk of automating inequality.

Recent literature suggests the reality might be messier than the authors' optimistic projections. For instance, McPeak et al. found that LLMs deployed in a Nigerian clinic tended to over-recommend laboratory tests that were standard in high-income settings but unavailable or unaffordable locally McPeak et al. [2024]. This highlights a critical policy gap: without careful

calibration to local contexts, AI models might exacerbate health disparities by offering advice that is technically correct but practically useless for underserved populations. Furthermore, Abaluck et al. raise questions about whether LLM assistance translates to better patient outcomes in all settings Abaluck et al. [2026]. Policies must therefore move beyond simple accuracy metrics (like the 90% USMLE pass rate cited in the book) to evaluating ecological validity and impact on actual healthcare delivery systems.

My own research intersects deeply with the mechanisms and risks described in the book. The authors recount instances where GPT-4 seemingly demonstrates "reasoning" by adapting to new scenarios without specific training. This phenomenon, known as In-Context Learning, was the focus of my work on generating traffic scenarios Aiersilan [2025] when I was at the University of Macau. Just as I use LLMs to create diverse corner cases to train robust motion planners, the book suggests medical educators could use GPT-4 to generate infinite, unique patient simulations to train medical students. This capability allows for a personalized medical education curriculum that could adapt dynamically to a student's weaknesses. However, the book's comforting description of GPT-4 as a "polite" and "avuncular" mentor also signals a danger I investigate in the Vibe-Check Protocol Aiersilan [2026]. The smooth, confident tone of AI response even when it hallucinates or fails at basic tasks like Sudoku, as admitted in Chapter 6 can lead to cognitive offloading. Users may accept the AI's output because it *feels* right, rather than confirming it is factually correct. This reliance on feeling aligns with the understanding that human verification is not merely mechanical data retrieval, but a process deeply rooted in "expert intuition" Dreyfus and Dreyfus [1986] and "tacit knowledge" Polanyi [2009]. Unlike the rigid logic of machines, humans possess an adaptive "gut feeling" Gigerenzer [2007] that allows them to grasp validity beyond explicit rules. **However, I gradually realized that this creates a profound tension here in our AGI-trend: does the ubiquity of AGI threaten to erode this distinctly human capacity, reducing the user from a comprehensive, intuitive verifier to a mere retrieval-based mechanical fact-checker — or is it actually encouraging this regression?** To prevent this regression in high-stakes domains like medicine, policy frameworks must mandate "friction" in the user interface that forces physicians

to engage their full critical faculties rather than passively accepting AI suggestions.

Looking forward to the next decade, it becomes clear that multimodal AI is poised to transform diagnostics in ways that transcend the text-heavy examples currently emphasized. I draw this conclusion not only from my own research in 3D computer vision applied to healthcare domain, but also from the rapid convergence of imaging, genomics, and clinical records in modern healthcare. The authors hint at this, but the implications are vast. In my literature review of AI-driven analysis for sarcopenia Aiersilan and Hahn [2026], I discuss how analyzing body shape from simple images could democratize access to diagnostics that currently require expensive scanners. We can project a future where the "symbiotic" doctor does not just chat with an AI but captures patient data through cameras and sensors that feed into a multimodal model, providing a holistic view of patient health that is currently impossible. This aligns with the book's vision but pushes it further towards a fully sensor-integrated clinical environment where the "pajama time" is eliminated not just by writing notes, but by automated observation.

To conclude, Lee, Goldberg, and Kohane have written a foundational text that captures the excitement of the AI dawn in medicine. Their prediction that doctors who use AI will replace those who do not is likely accurate, but it simplifies the transition. The real challenge lies in the nuance. We must navigate the tension between the efficiency of an automated resident and the risks of automation bias, ensuring that our policies protect patient safety without stifling innovation. The revolution they describe is inevitable, but its success will depend less on the code itself and more on the wisdom with which we choose to wield it.

References

Jason Abaluck, Robert Pless, Nirmal Ravi, Anja Sautmann, and Aaron Schwartz. Does llm assistance improve healthcare delivery? an evaluation using on-site physicians and laboratory tests. Technical report, National Bureau of Economic Research, 2026.

Aizierjiang Aiersilan. Generating traffic scenarios via in-context learning to learn better motion

planner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14539–14547, 2025.

Aizierjiang Aiersilan. The vibe-check protocol: Quantifying cognitive offloading in ai programming. *arXiv preprint arXiv:2601.02410*, 2026.

Aizierjiang Aiersilan and James Hahn. Literature review of ai-driven body shape analysis for sarcopenia. *Authorea Preprints*, 2026.

Hubert Dreyfus and Stuart E Dreyfus. *Mind over machine*. Simon and Schuster, 1986.

Gerd Gigerenzer. *Gut feelings: The intelligence of the unconscious*. Penguin, 2007.

Peter Lee, Carey Goldberg, and Isaac Kohane. *The AI revolution in medicine: GPT-4 and beyond*. Pearson, 2023.

Lawrence Lessig. *Code: And other laws of cyberspace*. ReadHowYouWant. com, 2009.

Grady McPeak, Anja Sautmann, Ohia George, Adham Hallal, Eduardo Arancón Simal, Aaron L Schwartz, Jason Abaluck, Nirmal Ravi, and Robert Pless. An llm's medical testing recommendations in a nigerian clinic: Potential and limits of prompt engineering for clinical decision support. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 586–591. IEEE, 2024.

Michael Polanyi. The tacit dimension. In *Knowledge in organisations*, pages 135–146. Routledge, 2009.

Richard A Spinello. *Cyberethics: Morality and Law in Cyberspace:*. Jones & Bartlett Publishers, 2010.