



Generative AI as a Revolutionary Knowledge Technology

Tech for Humanity Case Study

Developed by Brian Lutgens

Introduction

While the social and material conditions for producing reliable knowledge are varied and complex, six technologies have arguably had the greatest impact: writing, the alphabet, arithmetic, the printing press, binary and the algorithm, and the transistor. In order to be invented, each of these required an array of pre-existing technologies and institutions, and in order to be widely adopted, complex social circumstances had to prevail. Given an alignment of those circumstances, these six technologies—both individually and in their cumulative effect of leading to others—have probably done more in the pursuit of knowledge than anything else. Without them, what we know today—or rather, what we *think* we know, and *how* we know it—would not be possible.

This case study is about generative AI as a new technology with the potential to become as important for the production of reliable knowledge as these six predecessors. While still in its early development, scientists and physicians already recognize generative AI as revolutionary for their fields; it creates new possibilities for producing or applying reliable knowledge, as opposed to simply adding capacity to old ones. With that in mind, this case study will examine why scientists and physicians view AI in this way, and it will do so by examining how scientists, engineers, and physicians *currently use* it, and how they foresee using it in the future. In this way, this case study sets aside the hype, hyperbole, and speculation about its potential, limits, and dangers and examines the facts on what generative AI *is doing now* in science and medicine. Hopefully this focus offers an antidote to the pessimistic view prevalent in the humanities, which focuses on bias, hallucinations, risks, and inequities.

In effect, by letting scientists and physicians speak for themselves through their actions, one will get a better sense of how generative AI is being used now and how it will likely be used in the future, in the area where it probably has the most potential—the production and application of *reliable knowledge*.

Current uses in science and medicine best illustrate that potential, so this case study will examine it in order to pose discussion questions about its use in human applications.

Background

The Converging Nexus of Knowledge Technologies

The cumulative convergence of the six knowledge technologies can be seen in a technology we take for granted—the Global Positioning System. GPS is a satellite-based location system that uses atomic clocks to determine position on Earth. Writing, the alphabet, arithmetic, the printing press, binary and the algorithm, and the transistor lie at the heart of its development and use; without these six core achievements, the collateral technologies that make GPS possible would not themselves be possible, making those six—arguably—“first among equals” in a nexus of knowledge-producing innovations. This can be seen in the way in which *theory* governs how GPS works.

For instance, GPS relies on relativistic corrections in order to determine accurate position. The speed at which the atomic clocks in the satellites travel relative to the “stationary” atomic clocks on earth require corrections for time-dilation, and the distance the signals travel and the differing strength of the gravitational field through which they travel both require further corrections. GPS function requires rocketry, electricity, factories, computers, metallurgy, data farms—just about everything one can imagine in a vast, interconnected social and technological nexus—but it stands out in the way it relies on *scientific* knowledge. Without the advancement of scientific knowledge (as opposed to practical know-how), a technology like GPS would not be possible. Even after factoring in the ways in which all its component technologies and collateral dependencies function, corrections from a *theoretical understanding* of space, time, and radiation are “first among equals” in that function. Even if everything else worked, without those theoretical corrections, GPS positioning would not.

It is in this *knowledge* element of technology where knowledge technologies all come into practice. The theoretical understanding of the nature of the universe that makes GPS work would not be possible without—minimally—writing, the alphabet, arithmetic, the printing press, binary and the algorithm, and the transistor. Without writing and the literacy that alphabetized writing affords, without printing and the dissemination of knowledge that printing produces, without the mathematics developed from arithmetic, without the capability to digitize information and think about it, and then to combine the knowledge technologies in a manipulable device (the transistor), the science supporting GPS would not be possible. We simply could not make atomic clocks and launch satellites using the same practical know-how sailors utilized for navigation and time-keeping on ships in the modern European Age of Sail.

Simply put, *scientific knowledge* drives most technology now. Technologies and capacities derived from writing, the alphabet, arithmetic, the printing press, binary and the algorithm, and the transistor now drive that knowledge. This makes scientific knowledge dependent on technology in a unique way, even as that knowledge drives technology in turn.

Generative AI as a New Knowledge Technology

Now we begin our discussion of generative AI. Since the advent of computing and networked connectivity, knowledge has been produced at an unprecedented and exponentially accelerating rate. In the late 19th and early 20th centuries, it was possible for mathematicians like Henri Poincaré to be familiar with almost the whole of known mathematics. Today, a mathematician at Poincaré's level of brilliance would—at most—be familiar with less than 1% of it.

The scale of knowledge production and analysis taking place today is vast. In fields like physics and engineering, still relying on knowledge created by human effort unassisted by computational technologies, approximately 3.3 million articles were published in 2022 alone.¹ As such, knowledge is being produced at such a rate, on such a scale, with so much variety that it is utterly impossible for any human being to be aware of, much less consume and assimilate, anything approaching even a significant portion of it. However, generative AI can!

As an opening anecdote, consider this account from the National Academies of Science, Engineering, and Medicine. Around 2023, physicist Mario Krenn and his research team were

¹ Google Search AI result, March 21, 2025.

struggling to come up with an experiment that would observe a unique kind of quantum entanglement. Krenn developed an AI engine that could “design” quantum experiments. It was a very specialized form of generative AI, developed for and trained exclusively on quantum mechanics. After only a few hours, the AI suggested a design which actually gave the scientists the answer that had eluded them for weeks. In a subsequent problem, the AI was able to solve it by “reviving a long-forgotten technique and applying it to a new context.” From there, the human scientists were able to advance yet another discovery.²

This case is only one anecdote, but it happens to be typical of how generative AI is changing the nature and accelerating the pace of scientific discovery. This stands to make it the most significant knowledge technology since the transistor, adding to the noble line that began with writing and the alphabet.

How Generative AI is *Not* Used in Science

In order to understand this progress, first and foremost it is crucial to differentiate how generative AI is being used in science, from the now-ubiquitous public portals to large language models (LLM) available online.

Simply put, scientists are *not* accessing the “Ask me a question” prompt and having an AI program like Claude, Llama, or ChatGPT help design experiments or solve problems. Instead, when they use these engines, they are re-training them on specific information, then writing procedural code that instructs it how to “learn and “think” in a way that exploits built-in functions (APIs) and modifiable parameters (hyperparameters)—in other words, the scientists write (as best they can) what they themselves do into instructions and algorithms. Subsequently, they use the AI to replicate knowledge processing on a scale and scope that far surpasses any human capacity in terms of the amount of information one can learn and use.

This is essentially what Krenn and his team did with their quantum mechanics AI, which was able to make a connection between a forgotten technique and a new need because it could “read, retain, recollect, and reorganize” vast amounts of information, including all of the 3.3 million articles published in 2022, for instance. In this case it was the digitized quantum

² National Academies of Science, Engineering and Medicine. 2023. “How AI is Shaping Scientific Discovery.” <https://www.nationalacademies.org/news/2023/11/how-ai-is-shaping-scientific-discovery>

mechanics articles up to that time, but the principle is the same: scientists are essentially retraining, modifying, and customizing generative AI models to make them suitable for specific, knowledge-producing tasks. They are *not* using AI “out of the box” in the typical way students and professors might do via the online portals we are more familiar with.

An important point follows from this different kind of use. Among scientific applications, bias, inequity, and risk are typically non-issues, and the “hallucination problem” is largely mitigated. Where it is not completely so, however, it presents no more of a problem than plain old human error—which is much harder to mitigate, in any case. For this reason, the concerns about AI prevalent in the humanities simply do not arise in the scientific community. Scientific use of generative AI to produce reliable knowledge is so different from ordinary social uses and everyday human applications that concerns about bias, risk, equity, and hallucination don’t apply. They either do not arise in the kind of work being done, or they are relatively easy to overcome.

Generative AI in Science: A Sketch of Current Uses

Generative AI has been used by researchers to make significant new discoveries and solve previously unmanageable problems, *in just the last five years*—or about the time when Open AI released the early versions of ChatGPT, and large language models became available to scientists. It has also been used to model biological and physical processes or environments either more efficiently or in entirely new ways—ways beyond human capacity, as well as beyond the capacities of existing methods.

For instance, customized generative AI has been used to reveal the structure of crystalline materials where X-ray diffraction—the usual technique—fails. This happens for materials that can only exist in powdered form. After training a model on the Material Project database, which contains data on more than 150,000 materials, then “teaching” it to recognize structures which X-ray diffraction outputs from intact crystalline minerals, “Crystalyze” (the engine) was able to create a structure from the X-ray diffraction patterns of *powdered* crystals, where the orientations are randomized, resulting—eventually—in three new materials that were created in the lab. In this case these were new materials that could be used in permanent magnets. As a result of this advance, AI like Crystalyze can be used to generate new materials with different physical properties, even though their chemical composition is the same. Aside from industrial

applications, this allows experimentation with materials in ways previous limits on X-ray diffraction could never afford.³

In the biological sciences, similar advances have been made using generative AI. For instance, PlantRNA-FM is an AI trained on 54 billion pieces of RNA information comprising the genetic information across 1,124 plant species. After tailoring and training, PlantRNA-FM can now make precise predictions of RNA functions and identify functional RNA structural patterns across a wide variety of plants. These predictions and identifications can then be—and *have been*—validated experimentally: RNA structures identified by PlantRNA-FM affect the translation of genetic information into proteins. While PlantRNA-FM currently only applies to plants, in principle the same technique can be used in invertebrates and bacteria as well, given current information on these organisms.⁴

Simply put, it is unlikely advances in genomics will occur *without* generative AI similar to PlantRNA-FM, used across all species, as the number of possibilities, interactions, and elements are simply far too vast for any human to detect patterns in, and current statistical techniques have proven inadequate. Hence generative AI's revolutionary power to break the stalemate in between *information*, of which we have much, and testable *predictions*, of which we can make few. Once these predictions are made, they can be tested in the lab, rendering any “hallucination” problem insignificant. In fact, “hallucination” in AI used in this particular way will be indistinguishable from plain old-fashioned human error. In a field like genomics, one is grateful to have this new advantage, given the current unmanageable scale of predicting and identifying developmental and signaling pathways. Generative AI will improve the situation beyond previous possibilities.

In addition to advancing the capacity to predict structures and functions in materials and biological or chemical processes, generative AI has also been used to generate research hypotheses, based on the prevailing state of the art in a field. While still in the early stages, engineers at MIT have developed “SciAgents,” which take inputs in the form of scientific papers, then develop relations between concepts in those papers, and subsequently devise areas of

³ Eric Riesel et al. 2024. “Crystal structure determination from powder diffraction patterns with generative machine learning.” *Journal of the American Chemical Society*, vol 146, 44: 30340-30348.

⁴ Haopeng Yu et al. 2024. “An interpretable RNA foundation model for exploring functional RNA motifs in plants.” *Nature Machine Intelligence*, 6: 1616-1625.

interest for investigation and research, including specific targets for experimental work. In test trials so far, the results have been “robust and comparable” to ideas generated by actual scientists familiar with the input articles. As dehumanizing as this might sound to human creativity, consider again that when *millions* of articles are published in a given field in a year, tools like SciAgents will be instrumental in processing—at least as a first pass—this unfathomable amount of knowledge, in ways no human beings possibly could.⁵

In other words, given the rate at which new knowledge is produced, without the help of generative AI, the possibilities for new knowledge derived from existing knowledge will be *massively underachieved*.

These three accounts are only the tip of an iceberg that is travelling through science, which is changing how scientific work is done and expanding what it can do. Other examples include AI that can 1) reduce climate modelling that would take weeks on a supercomputer to hours on a workstation, in a lab; 2) model infants’ microbiomes to predict neurodevelopmental disorders linked to changes in them; 3) quickly calculate 3D genomic structures from DNA sequences; and, 4) determine the functional structure of proteins up to ten times more accurately than traditional methods. A search on a clearing house like *EurekAlert* (published by the *American Academy for the Advancement of Science*) shows *dozens* of similar breakthroughs in a matter of just a few years—far too many to cover in a case study like this.

Suffice it to say, generative AI is becoming a mainstay in the scientific community. The question is not “*Will* it be useful?”; rather it is: “*How many ways* will it be of use, and *how many areas* of research will it transform?”

Generative AI at the Bridge Between Basic Science and Medicine

Relatively few social or ethical questions hinge on these scientific uses of generative AI. It’s of little social concern whether Crystalyze or PlantRNA-FM are used in scientific discovery, as opposed to other techniques. It’s also hard to see an ethical dimension to their use, either, unless it involves downstream effects in industry or medicine, where impact on society and individuals can be anticipated. In any case, questions of bias, risk, and equity rarely hinge on

⁵ Alireza Ghafarollahi and Markus J. Buehler. 2024. “SciAgents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning.” *Advanced Materials*. doi: 10.1002/adma.20241523.

esoteric scientific choices regarding which tools aid in which discovery; and for this reason, application of generative AI in science need not be burdened with such questions, which otherwise arise in social or medical situations.

The situation changes in applied medicine, however, where direct consequences result in benefits or harms for individuals and society. Here, the questions of equity, bias, risk, and hallucination become front and center. Prior to these applications, though, basic science is the precursor to medical practice. In this science, knowledge for use in medicine is prepared, and in this kind of knowledge production, generative AI is taking a leading role, similar to that in non-medical scientific applications. Before addressing how generative AI can—and therefore *should* or *should not*—be used in medicine, it is worth taking a look at what it can do *now*, and how it *could* be used, in the event that the ethical and social implications are worked out. As with basic science, generative AI is revolutionizing the production and application of medical knowledge, and it is doing so at an astonishing pace.

To give an idea of the capabilities and limitations of the new technology, in the early 2020s, an AI model called PaLM (both Flan-PaLM and Med-PaLM) performed fairly well in answering one-off medical questions from data sets like MedMCQA and PubMedQA—obtaining 58% and 79% accuracy, respectively. The model performed almost as well on the medical licensing exam (USMLE, 67% accuracy). However, when it came to diagnosis and treatment, it underperformed human clinicians by a wide margin, nor could it be integrated into physician workflow (i.e., charting, handoff notes, messaging, etc.).⁶

Eighteen months after that performance test was attempted, a custom AI called MedFound was developed. MedFound was trained on a large database of medical texts, articles, and real-world clinical records.

It turned out that MedFound outperformed baseline LLMs like ChatGPT 4o, Llama 3-70B, and Clinical Camel-70B, achieving between 80-90% accuracy on diagnostic scenarios across eight specialties (compared to 50-60% for these baselines). This puts Medfound well within the estimated range of medical error among human clinicians. In fact, MedFound DX *outperformed* both junior and intermediate physicians in diagnosis in all eight specialties it was tested in, and it

⁶ Karan Singhai et al. 2023. “Large language models encode clinical knowledge.” *Nature*, 620: 172-179.

performed *nearly as well* as senior clinicians in most areas, and *slightly better* in some others. In addition to being accurate, MedFound was also trained to align with and integrate into physician workflow (perform charting, generate clinical notes, and answer patient queries, via MedFound DX-PA); it could also provide clinical summaries of and rationales for its decisions. In these respects as well, it beat baseline LLMs and performed as well as or better than human practitioners.⁷

In a year and a half, generative AI in medicine went from underperforming compared to human clinicians in diagnosis and treatment to performing better than most, and as well—or almost as well—as all but the best people in their fields. This is one example of what generative AI *can do* in medicine in the present day. Whether it *should* be used in this capacity will be left for discussion.

While not currently used for diagnosis, generative AI *is* being used in drug development, to identify biological targets for new drugs and even to design molecules to take advantage of these newly identified targets. At the forefront of this effort is Insilico Medicine, a Boston company determined to replace animal testing in drug development with generative AI models. Insilico is using generative AI at the bridge between basic science and practical medicine, and some of its accomplishments are worth noting.

Since 2021, Insilico has found twenty-two preclinical candidates for drug trials: several for cancer (including one for immunotherapy, one for a previously undruggable transcription factor, and one for drug-resistant solid tumors); one for idiopathic pulmonary fibrosis; one for inflammation in central nervous system diseases; and one for inflammatory bowel disease. It has also determined novel targets for existing drugs to treat endometriosis.⁸ Ten of those preclinical candidates have reached clinical trials, which means Insilico's use of generative AI has enabled it to identify therapeutic targets and develop potential drugs faster, more cheaply, and more efficiently than traditional drug companies are able to (those companies can typically bring only five compounds to clinical trials for every 5000 preclinical candidates, at a cost of

⁷ Xiaohong Liu et al. 2024. "A generalist medical language model for disease diagnosis assistance." *Nature Medicine*: <https://doi.org/10.1038/s41591-024-03416-6>.

⁸ EurekAlert News Releases, Oct 30, 2024, Nov 12, 2024, Nov 21, 2024, Dec 10, 2024, Dec 17, 2024, Jan 2, 2025, Jan 7, 2025. <https://www.eurekalert.org/>

\$15-100 million per compound, spanning one to six years per trial).⁹ If this success continues and can be expanded to include a broader class of disorders and targets, generative AI has the potential to *revolutionize* drug development, implementing direct benefits from basic science into practical medicine at an unprecedented pace.

Additionally, this efficiency will enable rare diseases to be targeted, where the market for a particular drug is too small to be profitable (or in cases where the drug has to be exorbitantly expensive in order to generate profit). This unprecedented technological advance would rectify one of the worst inequities in the pharmaceutical industry.

Generative AI models are currently capable of doing many of the tasks physicians do, as well as human physicians can usually do them, and AI models stand to revolutionize drug development. This raises the question: *what* should AI be used for in medicine, and *how* should it be used? It's unlikely that many people are worried about its use in drug development, for the same reasons they are relatively unbothered by its use in basic science. The methods scientists use currently have few ethical implications; and current methods mitigate the errors that might occur with AI in drug development. However, in practical medicine—*where patients' well-being is at stake*—the risks and benefits of using generative AI are critical and must be discussed.

Before that, however, we will examine three ways in which AI inventions are *currently* being applied in medicine. One of these three has already gone from concept to implementation.

Case Study

Generative AI in Medicine: Three Applications

As noted, at least one generative AI model is almost as good—and in some cases, as good or better than—human physicians at diagnosis, prescription, charting, and clinical notes (MedFound). As this model is emulated and replicated, diagnostic accuracy is likely to improve and become more common. However, as impressive—and potentially frightening—as this achievement might be, there is little discussion among physicians whether AI should take over this primary role. Instead, AI is being examined for use in other areas, where physicians are either overworked, bottlenecked, or otherwise not available: patient messaging and emergency

⁹ <https://greenfieldchemical.com/2023/08/10/the-staggering-cost-of-drug-development-a-look-at-the-numbers/>

medicine handoff notes (two main areas where they are overworked or bottlenecked), and in device usage cases when physicians are scarce and/or equipment is unavailable. Each area is worth looking at more closely.

Answering patient messages in electronic medical records is one of the main causes of nurse and physician burnout, which in some specialties is as high as 60%. A typical family practice doctor might spend three or four hours after seeing patients and finishing charts answering patient messages. In many cases, this is unpaid work. As a result, the potential for generative AI to assist in this task has been widely embraced, and most medical records providers are working to integrate it into their products and practices. One of the leading providers, Epic Systems, has already done so on an experimental basis, in collaboration with OpenAI. If this integration can be completed, generative AI stands to significantly reduce physicians' after-hours work load, without compromising patient safety.

So, how is this done, how effective has it been, and how might it be improved?

In a recent trial, to test AI assistance in physician-patient messaging, investigators created PAM Chat, a generative AI based on ChatGPT 4o, which was integrated into Epic's electronic records system. PAM Chat gives the provider the option for a template response to a patient's message: the provider has the option to use the draft itself, edit it and then use it, or extract parts from it and use it in another message which they create overall. At this time, however, according to Dr. Eden English, the lead on the trials (at nine clinics, with 166 users), this integration has met with limited success; only 12% of physician replies used elements from these AI-generated drafts. The doctors reported that their chief concern was *accuracy*—they were worried that PAM Chat would give patients inaccurate information—which in many cases it did. While most users (90%) agreed PAM Chat's replies were empathetic and patient-friendly, most were more concerned about accuracy, and 92% found that significant edits were needed. At the end of the trial, few physicians and medical assistants recommended the tool to others.¹⁰

This trial was a failure, but like many scientific failures, it may have been an instructive one. One shortcoming of the technology was that it relied on an “out of the box” large language model *not*

¹⁰ Eden English, MD, et al. 2024. “Utility of artificial Intelligence-generated draft replies to patient messages.” *JAMA Open Network*, 7(10): e2438573 and Roy Perlis, MD and Rita Rubin, MA. 2025. “Researcher tested an AI tool that drafts responses to patient messages—here's what they found.” *JAMA*, 33(8): 647-650.

specifically trained in medicine. As Dr. English notes: they had “no ability to fine-tune the model. We’re beholden to OpenAI’s fine-tuning, and it’s not linked to UpToDate [a medical reference library]. It can’t search websites in real time. It only has the information that GTP-4o has when it finished its most recent fine-tuning. All we have the ability to control is what we tell it in the prompt.”

Given these limits, one can’t help but wonder how the trial would have gone if PAM Chat relied on a model like MedFound (discussed above), or some other model specifically trained in medicine and not contaminated with *misinformation* in its original training data. Tailored models can be told *not* to use information from their original training and to rely *only* on the information in subsequent customization. As noted above, MedFound’s performance compares very well with that of physicians at diagnosis, treatment, and prescription. This success once again opens the question: would users be less likely to worry about its accuracy—and therefore be more likely to use it—in their replies to patient messages, if that were an option now?

What the Epic/OpenAI trial shows is that off-the-shelf, online large language models are still likely *inadequate* for use in medicine, particularly with patient messaging, just as they are for basic science and drug research. Given the newness of MedFound (January, 2025) relative to this particular trial (October, 2024), it remains to be seen how medically-specific large language models can be integrated into electronic medical records systems to assist with answering patient messages and other cumbersome tasks.

A second area where generative AI stands to reshape how physicians work is in emergency medicine handoff notes, which cause bottlenecks in busy ERs. In this case, generative AI has been more successful, most likely because this AI integration into electronic medical records relied on a large language model *specifically trained* for the task. The result: handoff notes written by the LLM were rated almost as useful and safe as physician-written notes, but they could be generated in a fraction of the time.

The specifics of the study bear mentioning.

In 2023, investigators introduced a customized Large Language Model into the electronic records system at New York-Presbyterian/Well Cornell Medical Center to generate customized

handoff notes from the ER physician to the next provider. Once a patient was seen in the ER, the LLM could access that patient's records and the ER summary, then generate a handoff note to be used by the specialist who took over the case. When the handoff notes for 1600 patients were examined, the LLM-generated notes scored higher than physician-generated notes on automatic measures evaluating inclusion of information relative to the original records—however, they did score slightly lower in terms of physician-rated evaluations on usefulness (4.04 out of 5, vs 4.36), completeness (4.00 vs 4.16), curation (4.24 vs 4.76), readability (4.00 vs 4.64), correctness (4.52 vs 4.90), and patient safety (4.06 vs 4.5). Yet, none of the notes generated by the LLM were rated “unsafe.” As the study noted: “the majority of identified quality limitations and incorrectness would have minimal impact on patient safety, even when extrapolated to the worst-case scenario of the LLM-generated summary content not being reviewed by a clinician before completion.”¹¹

While not rated as highly as physician-generated handoff notes, notes generated by a customized LLM were both safe and accurate and could be generated in a fraction of the time. The principal differences between physician vs. LLM-generated notes were: 1) physician notes could and sometimes did include information not found in the source notes; and, 2) LLM-generated notes were generated in a fraction of the time. These differences may surely factor into any decision on whether to rely on AI assistance in this case.

A third area where generative AI stands to revolutionize medical care is in “smart devices” that can do the work physicians or experts do, without the physician's or the expert's experience. One device in particular has already been developed and tested: a portable AI ultrasound that can determine gestational age of a fetus.

Gestational age is one of the principal factors determining prenatal care, in terms of steps the mother should take depending on the age of the fetus. In developed countries, gestational age is determined by ultrasounds once pregnancy is determined, but in underserved areas where physicians and equipment are unavailable, gestational age is usually determined anecdotally from the patient's history. This uncertainty affects both maternal and fetal health. Since

¹¹ Vince Hartman, MS et al. 2024. “Developing and evaluating large language model-generated emergency medicine handoff notes.” *JAMA Open Network*, 7(12): e2448723.

ultrasound is impractical in these areas—i.e., the machines themselves are prohibitively expensive, and they require a trained specialist to operate—an alternative is needed.

This kind of problem can be solved by introducing generative AI integration into existing, low-cost, portable devices. In a field trial in Lusaka, Zambia and Chapel Hill, North Carolina, a generative AI was integrated into the software of a low-cost, portable, battery powered, “blind” ultrasound device, which was then used to estimate gestational age during pregnancy, rather than a traditional ultrasound performed by a trained technician. Across a sample of 400 women in their first trimester, the AI-powered ultrasound was equally accurate to a trained technician using a non-portable device. As the study notes: “between 14- and 27-weeks’ gestation, novice users with no prior training in ultrasonography estimated GA [gestational age] as accurately with the low-cost, point-of-care AI tool as a credential sonographer performing standard biometry on the high-specification machine.” As it further notes: “These findings have immediate implications for obstetric care in low-resource settings, advancing the World Health Organization goal of ultrasonography estimate of GA for all pregnant people.”¹²

Unlike the two previous cases, it is hard to see this application of AI in medicine as anything but a resounding success that could be replicated in other medical devices.

So, generative AI has already been introduced on an experimental basis in three important areas in medicine, two of which are overburdens or bottlenecks in a physicians’ work, and one of which addresses medical care for underserved areas. Each has met with different levels of success. First, the attempted application of generative AI to doctor-patient messaging via electronic medical records was unsuccessful, but this failure relied on an “off-the-shelf” model, not a customized version. It remains to be seen how a customized, medical-specific AI might work. For ER handoff notes, a customized AI was much more successful than patient messaging: it produced notes comparable to physician-generated notes, but in a fraction of the time. Lastly, the use of a generative AI-enabled device in obstetrics appears to be a godsend: virtually any user can determine gestation age using a low-cost, AI-powered portable device, and do so just as well as any trained technician using an expensive, fixed device in a clinic or hospital. This is a benefit *now* to women in underserved areas.

¹² Jeffrey Stringer, MD, et al. 2024. “Diagnostic accuracy of an integrated AI tool to estimate gestational age from blind ultrasound sweeps.” *JAMA*, 332(8): 649-657

This case study has only focused on three areas where AI stands to reshape medicine. Many more are in the works, and overall, the reception of and openness to AI among physicians appears to be warm, if not without reservation, if we judge by the editorials in the *Journal of the American Medical Association*. However, medical AI adoption is only in the earliest stages; it remains to be seen just how effective it will be, where it will be used, and whether it should be used. Unlike its use in basic science and pharmaceutical research, generative AI in medicine and social applications raises critical and complex ethical issues that must be addressed first. This case study has focused on what AI in medicine *can currently do*, and what, experimentally, is being *tried*. Whether these advances are ready for mass implementation is left for discussion.

Conclusion

Generative AI is arguably the next phase in the evolution of knowledge technologies. Just like the alphabet, arithmetic, the printing press, binary and algorithms, and the transistor, AI stands to revolutionize how reliable knowledge is acquired and applied. In basic science, it is already being used to solve previously intractable problems, and it is in the early stages of being able to assimilate existing work to generate new areas of interest, with specific targets for experimentation. In drug research, it is already being used to speed up the pipeline from drug targets to compounds to clinical trials, and as the tools are fine-tuned, this efficiency will improve. In medicine, AI is making tentative strides in alleviating physician work-loads, it has as-of-yet unutilized diagnostic capacities, and it has a proven ability to make portable devices “smarter,” thereby serving underserved areas. Through its ability to incorporate the impact and effects of the previous innovations cumulatively in the pursuit of reliable knowledge, generative AI stands to become *the most important revolution yet*, first and foremost in the practices of science and medicine.

Nevertheless, its use in both medicine and scientific research still raises critical social and ethical questions, which are now left for discussion.

Discussion

1. Is the use of generative AI in “basic science” as ethically neutral as this case study indicates? Are there important ethical questions at the level of basic research that can and should be raised to assess its downstream effects on individuals and society?

We have seen an analogous dilemma with the advances in physics that led to the invention and use of nuclear weapons. If the research into the atomic structure of matter had not been undertaken or successful, nuclear weapons would not have been possible. We see the ethical dilemma even more clearly in biomedical research. For example, we have the capacity to engineer biological weapons, but there is a universal prohibition against doing so. Similar issues arise in genetic engineering.

Along these lines, then, if generative AI can advance research investigating a potential genetic basis for differences in capabilities and behaviors between population groups (i.e., by “race” or ancestry), is it humane and ethical to do that? In other words, are there roads in research opened by generative AI that should *not* be taken?

2. According to a recent estimate, medical error is the leading cause of death and permanent disability in the United States. Misdiagnosis accounts for a third of all medical error, and as many as 20% of fatal illnesses are misdiagnosed the first time.¹³ How might these statistics affect the decision to use generative AI in diagnosis and treatment? As noted in this case study, in less than two years, AI models have gone from significantly underperforming human doctors in diagnosis to performing better than junior and intermediate physicians, and almost as well as senior physicians. With this in mind, *should* AI be used in diagnosis and treatment, and if so, *how* should it be used?

3. Study after study has shown that algorithmic or actuarial methods either perform as well as or more frequently outperform human judgement. This has been demonstrated in areas as diverse as predicting 1) psychotherapeutic outcomes, 2) criminal recidivism, 3) race horse success, 4) the quality of wine vintages, 5) movie box office success, 6) a novel’s success, 7) supply chain procurement, 8) employee retention, and 9) teacher performance.¹⁴

¹³ David-Newman-Toker et al. 2023. “Burden of serious harms from diagnostic error in the USA.” *BMJ Qual Saf*, 0: 1-12. doi: 10.1136/bmjqs-2021-014130.

¹⁴ These are covered in Ian Ayres. 2007. *Super Crunchers*. New York: Bantam Dell.

Additionally, fifty years of empirical work has established that the vast majority of human judgements are *biased*. They commit cognitive distortions rooted in the unquestioning acceptance of prior information, the apparent resemblance of one piece of information to another, and the availability of information to recall.¹⁵ Due to these human tendencies, most people engage in motivated reasoning rooted in confirmation bias, and where statistical estimation is required, they are even more likely to display prejudicial bias.¹⁶

Given these two aspects of “machine versus intuitive” judgment, how should the use of generative AI in medicine and other social applications be assessed? Do these differences between human and machine matter in medical and social assessment? Would they matter if AI could be proven more accurate and reliable, as well as *without* bias? Are other factors besides accuracy and bias more important to consider in the use of generative AI, and if so, what are they, and why?

4. Assume that generative AI becomes as reliable at prediction and diagnosis as human judges in the legal system, and physicians. Should it be integrated into social uses and medicine? If so, how and why? If not, why?

In this integration, which assessment—human or machine—gets the final say, and why?

5. As discussed above, MedFound, a custom AI trained on medical knowledge, is more reliable in diagnosis, treatment, and prescription than junior and intermediate physicians. If you had a choice, would you rather have your medical care determined by 1) the AI, 2) the physician, 3) the physician using the AI (with the physician having the final say), or 4) the physician using the AI (with the AI having the final say)? Why?

¹⁵ See Daniel Kahneman. 2011. *Thinking Fast and Slow*. New York: Farrar, Straus, and Giroux.

¹⁶ See Dan Sperber and Hugo Mercier. 2017. *The Enigma of Reason*. Cambridge, MA: Harvard University Press.