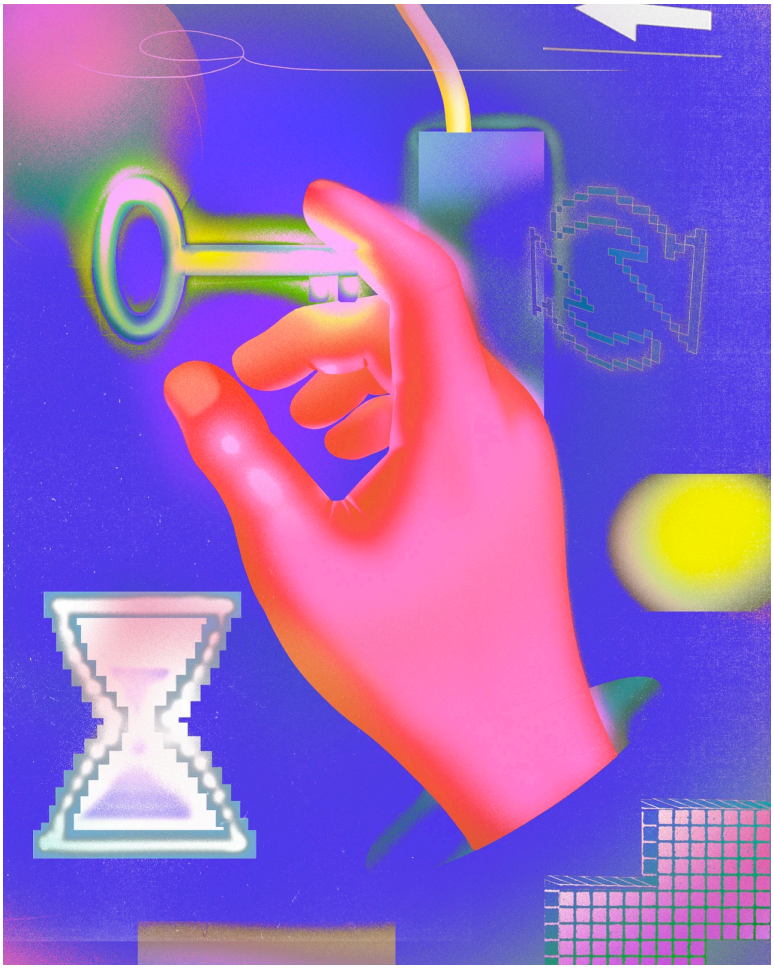

Illustration by Ohni Lisle

TECHNOLOGY

A BETTER WAY TO THINK ABOUT AI

Artificial intelligence is ready to collaborate. Why fixate on automation?

By David Autor and James Manyika

AUGUST 24, 2025, 7 AM ET

SHARE AS GIFT SAVE

No one doubts that our future will feature more automation than our past or present. The question is how we get from here to there, and how we do so in a way that is good for humanity.

Sometimes it seems the most direct route is to automate wherever possible, and to keep iterating until we get it right. Here's why that would be a mistake: imperfect automation is not a first step toward perfect automation, anymore than jumping halfway across a canyon is a first step toward jumping the full distance. Recognizing that the rim is out of reach, we may find better alternatives to leaping—for example, building a bridge, hiking the trail, or driving around the perimeter. This is exactly where we are with artificial intelligence. AI is not yet ready to jump the canyon, and it probably won't be in a meaningful sense for most of the next decade.

Rather than asking AI to hurl itself over the abyss while hoping for the best, we should instead use AI's extraordinary and improving capabilities to build bridges. What this means in practical terms: We should insist on AI that can collaborate with, say, doctors—as well as teachers, lawyers, building contractors, and many others—instead of AI that aims to automate them out of a job.

Radiology provides an illustrative example of automation overreach. In a widely discussed study published in April 2024, researchers at MIT found that when radiologists used an AI diagnostic tool called CheXpert, the accuracy of their diagnoses declined. "Even though the AI tool in our experiment performs better than two-thirds of radiologists," the researchers

wrote, “we find that giving radiologists access to AI predictions does not, on average, lead to higher performance.” Why did this good tool produce bad results?

A proximate answer is that doctors didn’t know when to defer to the AI’s judgment and when to rely on their own expertise. When AI offered confident predictions, doctors frequently overrode those predictions with their own. When AI offered uncertain predictions, doctors frequently overrode their own better predictions with those supplied by the machine. Because the tool offered little transparency, radiologists had no way to discern when they should trust it.

A deeper problem is that this tool was designed to automate the task of diagnostic radiology: to read scans like a radiologist. But automating a radiologist’s entire diagnostic job was infeasible because CheXpert was not equipped to process the ancillary medical histories, conversations, and diagnostic data that radiologists rely on for interpreting scans. Given the differing capabilities of doctors and CheXpert, there was potential for virtuous collaboration. But CheXpert wasn’t designed for this kind of collaboration.

When experts collaborate, they communicate. If two clinicians disagree on a diagnosis, they might isolate the root of the disagreement through discussion (e.g., “You’re overlooking this.”). Or they might arrive at a third diagnosis that neither had been considering. That’s the power of collaboration, but it cannot happen with systems that aren’t built to listen. Where CheXpert’s and the radiologist’s assessments differed, the doctor was left with a binary choice: go

with the software's statistical best guess or go with her own expert judgment.

It's one thing to automate tasks, quite another to automate whole jobs. This particular AI was designed as an automation tool, but radiologists' full scope of work defies automation at present. A radiological AI could be built to work collaboratively with radiologists, and it's likely that future tools will be.

Tools can be generally divided into two main buckets: In one bucket, you'll find automation tools that function as closed systems that do their work without oversight—ATMs, dishwashers, electronic toll takers, and automatic transmissions all fall into this category. These tools replace human expertise in their designated functions, often performing those functions better, cheaper, and faster than humans can. Your car, if you have one, probably shifts gears automatically. Most new drivers today will never have to master a stick shift and clutch.

In the second bucket you'll find collaboration tools, such as chain saws, word processors, and stethoscopes. Unlike automation tools, collaboration tools require human engagement. They are force multipliers for human capabilities, but only if the user supplies the relevant expertise. A stethoscope is unhelpful to a layperson. A chain saw is invaluable to some, dangerous to many.

Automation and collaboration are not opposites, and are frequently packaged together. Word processors automatically perform text layout and grammar checking even as they provide a blank canvas for writers to express ideas. Even so, we can distinguish automation from collaboration functions. The

transmissions in our cars are fully automatic, while their safety systems collaborate with their human operators to monitor blind spots, prevent skids, and avert impending collisions.

AI does not go neatly into either the automation bucket or the collaboration bucket. That's because AI does both: It automates away expertise in some tasks and fruitfully collaborates with experts in others. But it can't do both at the same time in the same task. In any given application, AI is going to automate or it's going to collaborate, depending on how we design it and how someone chooses to use it. And the distinction matters because bad automation tools—machines that attempt but fail to fully automate a task—also make bad collaboration tools. They don't merely fall short of their promise to replace human expertise at higher performance or lower cost, they interfere with human expertise, and sometimes undermine it.

The promise of automation is that the relevant expertise is no longer required from the human operator because the capability is now built in. (And to be clear, automation does not always imply superior performance—consider self-checkout lines and computerized airline phone agents.) But if the human operator's expertise must serve as a fail-safe to prevent catastrophe—guarding against edge cases or grabbing the controls if something breaks—then automation is failing to deliver on its promise. The need for a fail-safe can be intrinsic to the AI, or caused by an external failure—either way, the consequences of that failure can be grave.

The tension between automation and collaboration lies at the heart of a

notorious aviation accident that occurred in June 2009. Shortly after Air France Flight 447 left Rio De Janeiro for Paris, the plane's airspeed sensors froze over—a relatively routine, transitory instrument loss due to high-altitude icing. Unable to guide the craft without airspeed data, the autopilot automatically disengaged as it was set to do, returning control of the plane to the pilots. The MIT engineer and historian David Mindell described what happened next in his 2015 book, *Our Robots, Ourselves:*

> When the pilots of Air France 447 were struggling to control their airplane, falling ten thousand feet per minute through a black sky, pilot David Robert exclaimed in desperation, "We lost all control of the airplane, we don't understand anything, we've tried everything!" At that moment, in a tragic irony, they were actually flying a perfectly good airplane … Yet the combination of startle, confusion, at least nineteen warning and caution messages, inconsistent information, and lack of recent experience hand-flying the aircraft led the crew to enter a dangerous stall. Recovery was possible, using the old technique for unreliable airspeed—lower the pitch angle of the nose, keep the wings level, and the airplane will fly as predicted—but the crew could not make sense of the situation to see their way out of it. The accident report called it "total loss of cognitive control of the situation."

This wrenching and ultimately fatal sequence of events puts two design failures in sharp relief. One is that the autopilot was a poor collaboration tool.

It eliminated the need for human expertise during routine flying. But when expert judgment was most needed, the autopilot abruptly handed control back to the startled crew, and flooded the zone with urgent, confusing warnings. The autopilot was a great automation tool—until it wasn't, when it offered the crew no useful support. It was designed for automation, not for collaboration.

The second failure, Mindell argued, was that the pilots were out of practice. No surprise: The autopilot was beguilingly good. Human expertise has a limited shelf life. When machines provide automation, human attention wanders and capabilities decay. This poses no problem if the automation works flawlessly or if its failure (perhaps due to something as mundane as a power outage) doesn't create a real-time emergency requiring human intervention. But if human experts are the last fail-safe against catastrophic failure of an automated system—as is currently true in aviation—then we need to vigilantly ensure that humans attain and maintain expertise.

Modern airplanes have another cockpit navigation aid, one that is less well known than the autopilot: the heads-up display. The HUD is a pure collaboration tool, a transparent LCD screen that superimposes flight data in the pilot's line of sight. It does not even pretend to fly the aircraft, but it assists the pilot by visually integrating everything that the flight computer digests about the plane's direction, pitch, power, and airspeed into a single graphic called the flight-path vector. Absent a HUD, a pilot must read multiple flight instruments to intuitively stitch this picture together. The HUD is akin to the navigation app on your smartphone—if that app also had night vision, speed

sensors, and intimate knowledge of your car's engine and brakes.

The HUD is still a piece of complex software, meaning it can fail. But because it is built to collaborate and not to automate, the pilot continually maintains and gains expertise while flying with it—which, to be clear, is typically not the whole flight, but in crucial moments such as low-visibility takeoff, approach, and landing. If the HUD reboots or locks up during a landing, there is no abrupt handoff; the pilot already has hands on the control yoke for the entire time. Despite the fact that HUDs offer less automation than automatic landing systems, airlines have discovered that their planes suffer fewer costly tail strikes and tire blowouts when pilots use HUDs rather than auto-landers. Perhaps for this reason, HUDs are integrated into newer commercial aircraft.

Collaboration is not intrinsically better than automation. It would be ridiculous to collaborate with your car's transmission or to pilot your office elevator from floor to floor. But in some domains, occupations, or tasks where full automation is not currently achievable, where human expertise remains indispensable or a necessary fail-safe, tools should be designed to collaborate —to amplify human expertise, not to keep it on ice until the last possible moment.

One thing that our tools have not historically done for us is make expert decisions. Expert decisions are high-stakes, one-off choices where the single right answer is not clear—often not knowable—but the quality of the decision matters. There is no single best way, for example, to care for a cancer patient, write a legal brief, remodel a kitchen, or develop a lesson plan. But the skill, judgment, and ingenuity of human decision making determines outcomes in many of these tasks, sometimes dramatically so. Making the right call means exercising expert judgment, which means more than just following the rules. Expert judgment is needed precisely where the rules are not enough, where creativity, ingenuity, and educated guesses are essential.

But we should not be too impressed by expertise: Even the best experts are fallible, inconsistent, and expensive. Patients receiving surgery on Fridays fare worse than those treated on other days of the week, and standardized test takers are more likely to flub equally easy questions if they appear later on a test. Of course, most experts are far from the best in their fields. And experts of all skill levels may be unevenly distributed or simply unavailable—a shortage that is more acute in less affluent communities and lower-income countries.

Expertise is also slow and costly to acquire, requiring immersion, mentoring, and tons of practice. Medical doctors—radiologists included—spend at least four years apprenticing as residents; electricians spend four years as apprentices and then another couple as journeymen, before certifying as master electricians; law-school grads start as junior partners, and new Ph.D.s begin as assistant professors; pilots must log at least 1,500 hours of flight

before they can apply for an Airline Transport Pilot license.

The inescapable fact that human expertise is scarce, imperfect, and perishable makes the advent of ubiquitous AI an unprecedented opportunity. AI is the first machine humanity has devised that can make high-stakes, one-off expert decisions at scale—in diagnosing patients, developing lesson plans, redesigning kitchens. AI's capabilities in this regard, while not perfect, have consistently been improving year by year.

What makes AI such a potent collaborator is that it is not like us. A modern AI system can ingest thousands of medical journals, millions of legal filings, or decades of maintenance logs. This allows it to surface patterns and keep up with the latest developments in health care, law, or vehicle maintenance that would elude most humans. It offers breadth of experience that crosses domains and the capacity to recognize subtle patterns, interpolate among facts, and make new predictions. For example, Google DeepMind's AlphaFold AI overcame a central challenge in structural biology that has confounded scientists for decades: predicting the folding labyrinthine structure of proteins. This accomplishment is so significant that its designers, Demis Hassabis and John Jumper, colleagues of one of us, were awarded the Nobel Prize in Chemistry last year for their work.

The question is not whether AI can do things that experts cannot do on their own—it can. Yet expert humans often bring something that today's AI models cannot: situational context, tacit knowledge, ethical intuition, emotional intelligence, and the ability to weigh consequences that fall outside the data.

Putting the two together typically amplifies human expertise: Oncologists can ask a model to flag every recorded case of a rare mutation and then apply clinical judgment to design a bespoke treatment; a software architect can have t[illegible]del retrieve dozens of edge-case vulnerabilities and then decide which s[illegible] patch best fits the company's needs. The value is not in substituting e[illegible]ert for another, or in outsourcing fully to the machine, or indeed in p[illegible]ing the human expertise will always be superior, but in leveraging human and rapidly-evolving machine capabilities to achieve best results.

As AI's facility in expert judgment becomes more reliable, capable, and accessible in the years ahead, it will emerge as a near-ubiquitous presence in our lives. Using it well will require knowing when to automate versus when to collaborate. This is not necessarily a binary choice, and the boundaries between human expertise and AI's capabilities for expert judgment will continually evolve as AI's capabilities advance. AI already collaborates with human drivers today, provides autonomous taxi services in some cities, and may eventually relieve us of the burden and risk of driving altogether—so that the driver's license can go the way of the manual transmission. Although collaboration is not intrinsically better than automation, premature or excess automation—that is, automation that takes on entire jobs when it's ready for only a subset of job tasks—is generally worse than collaboration.

The temptation toward excess automation has always been with us. In 1984, General Motors opened its "factory of the future" in Saginaw, Michigan. President Ronald Reagan delivered the dedication speech. The vision, as MIT's Ben Armstrong and Julie Shaw wrote in *Harvard Business Review* in

2023, was that robots would be "so effective that people would be scarce—it wouldn't even be necessary to turn on the lights." But things did not go as planned. The robots "struggled to distinguish one car model from another: They tried to affix Buick bumpers to Cadillacs, and vice versa," Armstrong and Shaw wrote. "The robots were bad painters, too; they spray-painted one another rather than the cars coming down the line. GM shut the Saginaw plant in 1992."

There has been much progress in robotics since this time, but the advent of AI invites automation hubris to an unprecedented degree. Starting from the premise that AI has already attained superhuman capabilities, it is tempting to think that it must be able to do everything that experts do, minus the experts. Many people have therefore adopted an automation mindset, in their desire either to evangelize AI or to warn against it. To them, the future goes like this: AI replicates expert capabilities, overtakes the experts, and finally replaces them altogether. Rather than performing valuable tasks expertly, AI makes experts irrelevant.

Research on people's use of AI makes the downsides of this automation mindset ever more apparent. For example, while experts use chatbots as collaboration tools—riffing on ideas, clarifying intuitions—novices often treat them mistakenly as automation tools, oracles that speak from a bottomless well of knowledge. That becomes a problem when an AI chatbot confidently provides information that is misleading, speculative, or simply false. Because current AIs don't understand what they don't understand, those lacking the expertise to identify flawed reasoning and outright errors may be led astray.

The seduction of cognitive automation helps explain a worrying pattern: AI tools can boost the productivity of experts but may also actively mislead novices in expertise-heavy fields such as legal services. Novices struggle to spot inaccuracies and lack efficient methods for validating AI outputs. And methodically fact-checking every AI suggestion can negate any time savings.

Beyond the risk of errors, there is some early evidence that overreliance on AI can impede the development of critical thinking, or inhibit learning. Studies suggest a negative correlation between frequent AI use and critical-thinking skills, likely due to increased "cognitive offloading"—letting the AI do the thinking. In high-stakes environments, this tendency toward overreliance is particularly dangerous: Users may accept incorrect AI suggestions, especially if delivered with apparent confidence.

The rise of highly capable assistive AI tools also risks disrupting traditional pathways for expertise development when it's still clearly needed now, and will be in the foreseeable future. When AI systems can perform tasks previously assigned to research assistants, surgical residents, and pilots, the opportunities for apprenticeship and learning-by-doing disappear. This threatens the future talent pipeline, as most occupations rely on experiential learning—like those radiology residents discussed above.

Early field evidence hints at the value of getting this right. In a PNAS study published earlier this year and covering 2,133 "mystery" medical cases, researchers ran three head-to-head trials: doctors diagnosing on their own, five

leading AI models diagnosing on their own, and then doctors reviewing the AI suggestions before giving a final answer. That human-plus-AI pair proved most accurate, correct on roughly 85 percent more cases than physicians working solo and 15 to 20 percent more than an AI alone. The gain came from complementary strengths: When the model missed a clue, the clinician usually spotted it, and when the clinician slipped, the model filled the gap. The researchers engineered human-AI complementarity into the design of the trials, and saw results. As these tools evolve, we believe they will surely take on autonomous diagnostic tasks, such as triaging patients and ordering further testing—and may indeed do better over time on their own, as some early studies suggest.

Or, consider an example with which one of us is closely familiar: Google's Articulate Medical Intelligence Explorer (AMIE) is an AI system built to assist physicians. AMIE conducts multi-turn chats that mirror a real primary-care visit: It asks follow-up questions when it is unsure, explains its reasoning, and adjusts its line of inquiry as new information emerges. In a blinded study recently published in *Nature*, specialist physicians compared the performance of a primary-care doctor working alone with that of a doctor who collaborated with AMIE. The doctor who used AMIE ranked higher on 30 of 32 clinical-communication and diagnostic axes, including empathy and clarity of explanations.

By exposing its reasoning, highlighting uncertainty, and grounding advice in trusted sources, AMIE pulls the user into an active problem-solving loop instead of handing down answers from on high. Doctors can potentially

interrogate and correct it in real time, reinforcing (rather than eroding) their own diagnostic skills. These results are preliminary: AMIE is still a research prototype and not a drop-in replacement. But its design principles suggest a path toward meaningful human collaboration with AI.

Full automation is much harder than collaboration. To be useful, an automation tool must deliver near flawless performance almost all of the time. You wouldn't tolerate an automatic transmission that sporadically failed to shift gears, an elevator that regularly got stuck between floors, or an electronic tollbooth that occasionally overcharged you by $10,000.

By contrast, a collaboration tool doesn't need to be anywhere close to infallible to be useful. A doctor with a stethoscope can better understand a patient than the same doctor without one; a contractor can pitch a squarer house frame with a laser level than by line of sight. These tools don't need to work flawlessly, because they don't promise to replace the expertise of their user. They make experts better at what they do—and extend their expertise to places it couldn't go unassisted.

Designing for collaboration means designing for complementarity. AI's comparative advantages (near limitless learning capacity, rapid inference, round-the-clock availability) should slot into the gaps where human experts tend to struggle: remembering every precedent, canvassing every edge case, or drawing connections across disciplines. And at the same time, interface design must leave space for distinctly human strengths: contextual nuance, moral reasoning, creativity, and a broad grasp of how accomplishing specific tasks

achieves broader goals.

Both AI skeptics and AI evangelists agree that AI will prove a transformative technology—indeed, this transformation is already under way. The right question then is not whether but how we should use AI. Should we go all in on automation? Should we build collaborative AI that learns from our choices, informs our decisions, and partners with us to drive better results? The correct answer, of course, is both. Getting this balance right across capabilities is a formidable and ever-evolving challenge. Fortunately, the principles and techniques for using AI collaboratively are now emerging. We have a canyon to cross. We should choose our routes wisely.

ABOUT THE AUTHORS

David Autor

James Manyika