A reprint from

American Scientist

the magazine of Sigma Xi, The Scientific Research Honor Society

Self-Driving Science

End-to-end AI-automated science undercuts trust, obscures responsibility, and restricts adaptivity.

Brian Uzzi and Julio M. Ottino

The deliberative quality of science exists in tension with a mandate to produce more findings, faster, and that pressure has grown in recent decades. Now the push to produce has acquired fresh force amid rising calls to employ AI across all aspects of science, from the conception of ideas and questions to the final publication of papers. Proponents of this "end-to-end science" tout the claimed benefits of these systems and their alleged efficiencies but seem loath to consider what science stands to lose in the trade-off: the very features that make it productive and trustworthy.

We argue that closed loops that would automate most or all stages of scientific work will disastrously undermine the qualities necessary for rigorous investigation and creative discovery. If the sole goal of AI is to boil down what researchers do and then do it much faster with far less oversight and no diversity of opinion, then the likely output will be the rapid output of a mountain of junk.

Yes, the AI moment is upon us, and few would deny its strengths and uses. Indeed, scientists are at the forefront of AI usage, whether as a means of analyzing vast datasets or as a proposed solution to an overstressed peer review system. The former application, which uses machine learning via artificial neural networks to produce results that are validated and confirmed by human scientists, has scored significant successes in areas such as predicting cancer outcomes, using retinal scans to gauge cardiovascular risk, and forecasting weather with an accuracy that rivals or exceeds conventional physics-based models. Also promising are AI reasoning systems, which solve problems by connecting data, drawing inferences, and determining steps needed to reach a conclusion rather than simply recalling facts from memory or matching patterns. Other so-called AIs, such as large language models (LLMs), have a far less stellar track record, especially when scientists rely on them for research or writing.

What happens when we mistake the machinery of scientific production for science itself?

AI reasoning and neural network systems have already enabled Nobel-caliber breakthroughs. Indeed, one of them helped win that very prize in 2024. AlphaFold, which uses deep learning techniques to predict protein structures, largely solved one of biology's most essential and challenging puzzles that had stymied humans for decades: predicting the final 3D shape that proteins fold into and that determines their critical biological functions. That breakthrough, and the fact that most of the component parts of end-to-end systems already exist and are being used separately, is now helping to drive the push to build such systems.

But end-to-end systems go far beyond using machine learning to crunch large datasets; proponents seek to put scientific work almost totally in the virtual hands of AIs—from vision to hypothesis generation and on through experimentation, interpretation, writing, and publishing. Progress toward this goal is happening fastest in research stages that are digital, repetitive, measurable, and easy to evaluate. But we're already seeing "shakedown cruises" of systems meant to perform literature reviews, gin up hypotheses, design experiments, write manuscripts, and run simulated peer reviews. The news has widely covered so-called "vibe coding."

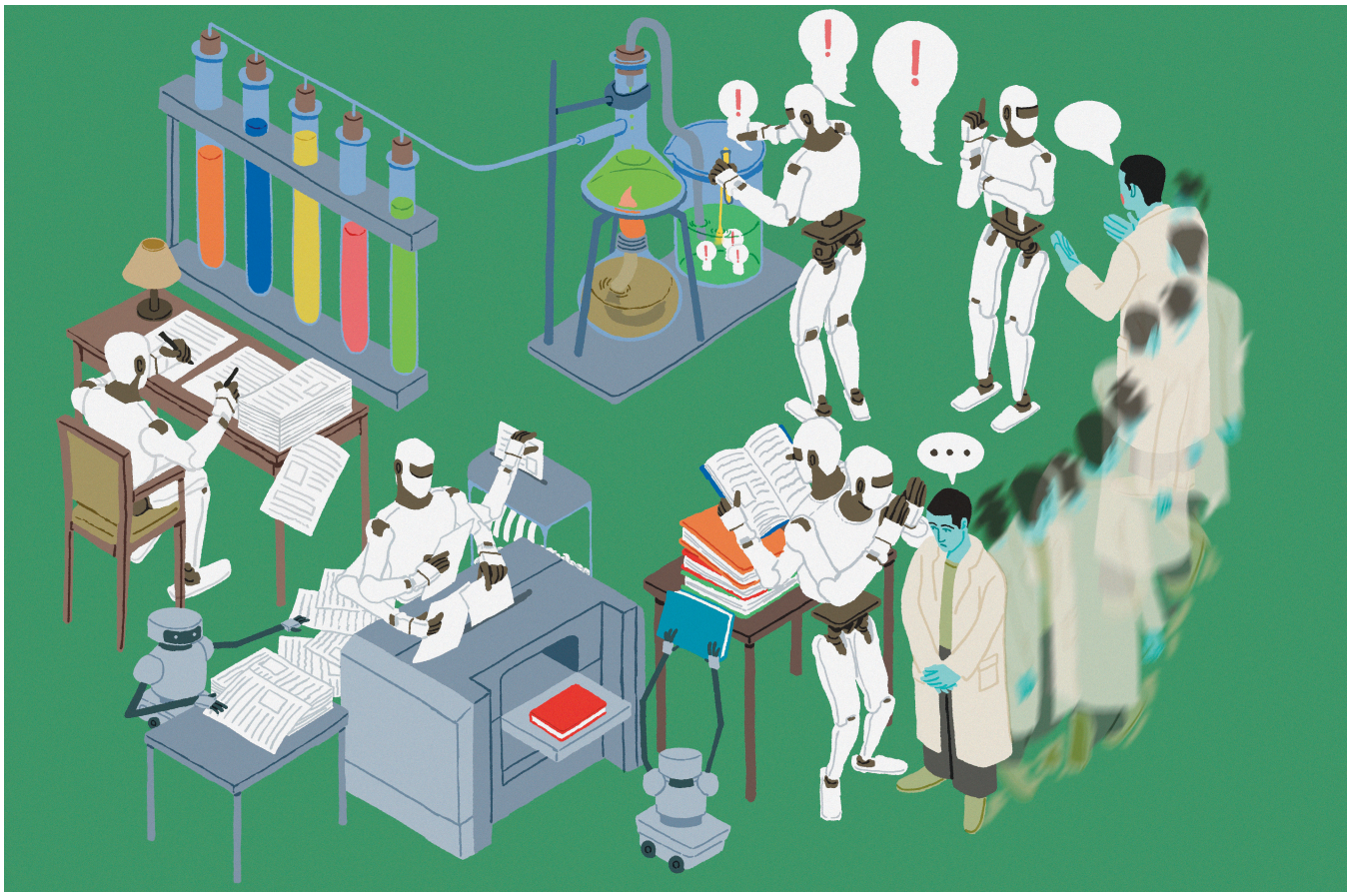
While we wait for AIs that can tackle the more difficult tasks of laboratory science, AI assistants, copilots, and agents are already being connected to instruments and robotic platforms. One such system resides at the supercomputer-powered Lawrence Berkeley National Laboratory, which is managed by the University of California. The facility's A-lab uses a "lab in the loop" system to reportedly churn out results about 100 times as fast as a human scientist can. Human participation may be constitutive, merely supervisory, or, in fully autonomous systems, superfluous.

QUICK TAKE

The accelerating AI age is driving a push to increase the involvement of AI in every stage of research, creating so-called end-to-end science systems.

Emphasizing speed and productivity at the expense of more deliberative human processes threatens to undermine rigorous investigation, creative discovery, and trust.

Gaffes, disasters, or malfeasance may result unless we design governance into such systems before the stampede to be first and most widely adopted make reform impossible.



Yuki Murayama

End-to-end science would put AI in charge of every stage of research: generating hypotheses, gathering data, running experiments, performing analyses, writing papers—even deciding what should be published, revised, or rejected. This human scientist is a mere facilitator between AIs. The authors argue that such systems will undermine rigor, creative discovery, and trust in science.

To see how these constitutive and supervisory approaches differ, consider an x-ray: A system that includes humans in a constitutive role cannot work without expert approval—an AI system can make a recommendation, but only a doctor can make a diagnosis. A system with supervisory roles for humans would make its own AI diagnoses, but a human radiologist or physician would oversee and review them. A fully automated system, however, removes humans from the equation entirely. For example, the U.S. Department of Energy's multimillion-dollar Frontiers in Artificial Intelligence for Science, Security, and Technology (FASST) initiative develops next-generation AI models and automated labs in the domains of science and energy using supercomputers. The U.K. government has partnered with Google DeepMind to establish an automated, AI-driven, robotic research lab focused on materials science research in areas such as superconductors, semiconductors, clean energy, medical imaging, computer chips, and transport. In the corporate sector, major technology companies in the United States are heavily investing in end-to-end systems, including Microsoft Discovery and Anthropic's Claude Science. These workbenches—integrated software environments that combine everything needed to carry out a certain type of work—are being developed with full automation in mind.

Our concern is not whether we should use AI to augment specific research steps; AlphaFold and two of Google DeepMind's projects—the weather-forecasting GraphCast and geometry-solving AlphaGeometry—demonstrate how powerful a partnership between humans and AI reasoning models can be. Rather, we must weigh the wisdom of replacing an established process, including its productive tension and critical debate that have historically generated trustworthy scientific insights, with an autonomous pipeline optimized for throughput that produces more science faster by smoothing out the very frictions that have shaped science.

End-to-End Ethos

It is difficult to carefully weigh such questions when the current momentum toward using end-to-end systems feels like a gold rush; it is a race, driven partly by financial and career incentives, that values automating faster science at scale above achieving the greatest possible meaning and understanding. Research paper mills already provide a cautionary precedent, revealing how emphasizing productivity and publication over caution and rigor can lead to industrial-scale fraud, undermine peer review, and contaminate the scientific record.

End-to-end systems are capable of the same faults and failures, only faster and with even less oversight. This is not the "fail fast" philosophy expounded by startups and used by scientists to quickly separate the theoretical wheat from the chaff; it is more akin to failing invisibly or failing to achieve fully. Amid this headlong rush, what is rarely addressed is how such systems will preserve scientific integrity, which is a consequence not of good algorithms but of a broader social system of norms, incentives, and accountability mechanisms that make science adaptive, trustworthy, and, in a broad sense, wise.

Yuki Murayama

Marie Curie's work shows how science benefits from creative thinkers, debate, replication, and validation that end-to-end systems may lack. 1. Curie presented her work to peers and discovered polonium and radium. 2. Eugène-Anatole Demarçay's spectroscopy confirmed radium as an element, while Ernest Rutherford's work added to radioactivity theory. 3. Gerhard Carl Schmidt added thorium to the list of radioactive elements, while other researchers worked to validate such hypotheses. 4. Curie's Nobel Prize-winning research led to fields such as radiation medicine.

The difference between end-to-end systems and traditional science is not only speed but also architecture. Conceptually, an end-to-end system is a tightly integrated, automated assembly, whereas traditional science is an open and fragmented system consisting of stages led by separate humans. Traditionally, a scientist views a problem, has an idea, reads the literature, designs a study, collects data, analyzes it, writes a paper, submits it to a journal, and makes revisions, while peer experts confirm, extend, repudiate, or ignore the final product. This fragmentation is not a weakness; it is a strength. The distribution of intellectual puzzle pieces invites debates, varied points of view, and dissent. Traditional science leverages tacit knowledge, preserves the norms of openness that contribute to trust, maintains the incentives that make individual scientists accountable for their work, and incorporates the diversity of approaches that make science adaptable.

Beyond the lab, it is critical that science be embedded in human society, because science doesn't just produce answers; it chooses the questions worth asking and weighs the value of evidence. If end-to-end systems are to strengthen science rather than hollow it out, then social embeddedness cannot be an afterthought. As with any tool or discovery, we must ask difficult, essential questions: Who chooses the system's goals? What authority should the AI have? How might we reconstruct and analyze the AI's decision processes and winnowing of options? Who is responsible for the results? Who gets the credit and the funding for the work? Who can contest the findings? What skills must we maintain to ensure that we can understand or overrule such systems?

Figures from the history of science have previously revealed the wisdom of such questions. They show us what is at stake and why the end-to-end approach threatens to fail on its own terms.

Guidance From History

Scientists are neither perfect nor incorruptible. Part of the beauty and trustworthiness of science lies in its self-correcting nature, which helps to uncover such misfeasance, and in its ethos, which helps give pause to those who might stray. AIs currently lack the latter and, depending on their implementation, have only the vague potential for the former. Where is the capacity for self-correction? As theoretical physicist Richard Feynman of the California Institute of Technology put it in his 1974 "Cargo Cult Science" address: "The first principle is that you must not fool yourself—and you are the easiest person to fool."

Scientific ethos was central to the work of Robert Merton of Columbia University, one of the founders of the sociology of science. Merton argued that science's integrity rests on shared norms embedded in communities where reputations and peer scrutiny hold work accountable. Those norms included communalism, universalism, disinterestedness, and organized skepticism. When a machine system generates hypotheses, designs experiments, analyzes data, and drafts papers, accountability becomes opaque. Decisions once contestable in a community under Merton's norms become invisible in engineered systems, while machine production rates overwhelm human capacity to vet output.

The value of a scientific fellowship of skeptical peers is well illustrated by Polish physicist and chemist Marie Curie. She transformed physics and chemistry not through genius alone, but because a scientific community scrutinized, challenged, and ultimately confronted the reality of her work. Curie's 1898 claim was not simply finding something radioactive; she asserted that radioactivity appeared to be an intrinsic property of atoms. That idea ran counter to the then-dominant view that atoms were stable, indivisible units. Changing that view required that Curie make her evidence available for open debate, along with evidence from separate communities and other researchers. Only then could her finding be accepted not as a physical anomaly, chemical edge case, or medical curiosity, but as a foundational discovery.

End-to-end systems trade such productive friction for efficiency; without a normative structure that assigns community and personal account-

ability prior to their deployment, such systems could weaken or collapse the conditions necessary for trust. History is replete with examples of resistance to adding precautionary measures to an efficient or profitable system after implementation. The World Wide Web was deployed without security as a first-order design constraint, leading to decades of patches, overlays, and protocol retrofits that continue to struggle to integrate with encoded architectural choices.

Returning to Feynman's point, one aspect of not fooling ourselves lies in accepting what science tells us, even if it flies in the face of what we reckon to be true. Clinging to entrenched beliefs is a human foible, one that science's methods and norms are meant to correct; whether AIs and end-to-end systems will share it will depend on how we train and monitor them—vital issues that, in our opinion, are receiving too little attention. German theoretical physicist Max Planck illustrated the capacity to transcend one's own biases through his solution to the problem of radiation by blackbodies, ideal objects that absorb all incoming radiation and emit light solely according to their temperatures. Although it pained him to break with classical physics, predictions of certain blackbody emission spectra made by established theories were incorrect, leading him reluctantly to introduce energy quantization. The effort, the resistance, and the reluctant acceptance of a radical new framework were not incidental to the discovery. They were the essence of it.

Pressure Without Probity

AI can be fast, but it can also crank out a lot of low-quality work. That capacity makes it harder to weed out flawed or fraudulent results and to validate if advances are genuine. Those factors, combined with the blurring of ownership and ethical responsibility, ultimately reduce trust and add impediments, potentially nullifying the speed that is a key selling point of end-to-end systems.

Consider how such systems stress the principles illustrated by these historic examples. Lower barriers to production will flood science with lower-quality work and generate incentive spirals: If competitors publish 10 times more findings and papers using end-to-end systems, then everyone—especially those with the vast resources needed for the deployment of such systems—will be pressured to do likewise. The once-sterling currency of scientific findings will debase and, as effortless science grows more plentiful, trust in science's collective value will decrease. That trust has been granted not because science is fast, but because scientists are willing to be corrected by reality and to own their mistakes. End-to-end systems can error-correct to a point but would automate around any issues that lie outside of coding and beyond experimental and data issues, untrammeled by outside influences and unaware of errors caused by faulty problem formulation, self-confirming validation, or blind-spot-riddled training. These systems can catch errors in the pipeline but not in the paradigm.

A monoculture is efficient right up until it isn't—and then it collapses without backup strategies.

An end-to-end approach also skirts the question of who ultimately is ethically responsible for scientific work and its products. Theoretical physicist J. Robert Oppenheimer's invocation of the *Bhagavad Gita*—"Now I am become Death, the destroyer of worlds"—showed there was a human "I" who could be held accountable and who bore the moral weight of the consequences of his work. End-to-end system developers risk building systems in which potentially world-changing research arrives without any durable and culpable "I" attached. Instead, responsibility will fragment across workflows, platforms, algorithms, and institutions that are focused on productivity, not probity. When no one can say, "I should have known better," no one will. Self-driving cars exemplify the point. In most systems there is a symmetry between credit and culpability, but it is not guaranteed. When self-driving cars do well in traffic and on long drives, the company takes credit. When self-driving systems fail, producing injuries and accidents, assigning or accepting culpability remains a chronic problem. The law is evolving quickly to address these issues, but how culpability will be handled in intellectual circles is far less clear.

Proponents of end-to-end science often invoke evolution to legitimate the enterprise, which is framed as the next adaptive step in science's development, a natural extension of the tools that science has always adopted. But this invocation falls apart under scrutiny, and the collapse is instructive. Biological evolution does not succeed by optimizing a single pipeline but via variation—multiple independent lineages testing different solutions simultaneously, under selection pressure that preserves diversity even as it eliminates unfit traits. Efficiency tends to narrow the variation that makes a system adaptive. A monoculture is efficient right up until it isn't—and then it collapses without backup strategies.

The argument that end-to-end science is evolutionary is not only mistaken but also self-undermining. Research shows that when AIs select ideas in blind "taste tests," they rank AI ideas over human ideas most of the time. End-to-end science does not extend evolutionary logic; it erodes it. Optimize away variation, and you get more science on paper but less capacity to generate the unexpected results that science depends on to advance. Maybe end-to-end systems can correct this problem, but current evidence suggests that the incentives are not there. A study of 41.3 million research papers showed that scientists who adopt AI tools significantly increase their productivity but collectively narrow their scientific focus.

Science Is Not a Clock

Science's need for adaptability and adjustable viewpoints stems partly from the range of problems with which it must contest. In a 1965 lecture, Austrian–British philosopher Karl Popper defined two ends of this spectrum, called "clocks" and "clouds." Clock problems can be decomposed and rebuilt piece by piece; their behavior is predictable from their parts. Examples include building a bridge or calculating a satellite's orbit. Cloud problems are complex, with so many interacting parts that they must be analyzed as wholes; they describe behaviors that emerge from interactions, making it practically impossible to draw conclusions from decomposition and pointless to analyze parts in isolation. Ecosystems and financial markets are examples of cloud problems.

Most scientific questions fall between clocks and clouds. But end-to-end science embraces clock thinking—the assumption that by automating each component of the scientific workflow we can optimize the whole. The history of science makes clear why this approach fails. Merton shows that institutional trust is cultivated through the shared norms and practices of a scientific community over time. Curie makes clear that socially embedded frictions in the traditional system that embrace diversity of thought are how science course-corrects. Feynman argues that self-correction depends on the willingness to face facts, expose errors, and turn away from self-deception. Oppenheimer demonstrates that responsibility requires the continual involvement of a morally responsible being, the sort of clear ethical links that distributed pipelines dissolve. Darwinian evolution shows that adaptive capacity requires variation, which end-to-end efficiency destroys.

These perspectives converge on a single necessity: To head off the worst potential outcomes of end-to-end systems, be they embarrassing gaffes or full-blown disasters, we must design governance into them before gold-rush competitive pressure makes reform impossible. Even basic safeguards such as promoting adversarial review, instituting publication limits, and preserving time for productive ideation would help. We can cultivate a culture of responsibility, expanding scientific norms to encompass the silicon world and ensuring that such an ethos retains provenance and accountability rules. Institutions could incentivize competing perspectives, rewarding a diversity of ideas and views as grist for the mill instead of penalizing them as grit in the gears.

The question is not whether to build end-to-end systems; it is whether we are wise enough to build them in ways that preserve the conditions for wisdom. Without the social embeddedness of science, end-to-end systems share the same failure modes as traditional science, but operating faster, around the clock, and with even less oversight—unless we tread carefully now. The gold rush may be worth it, but we should embark upon it knowing what we stand to lose and who stands to gain. If we fail to build in the values that scientific culture has acquired through hard experience, then we are not accelerating science; we are replacing it with something that resembles science—and hoping the resemblance holds.

Bibliography

Candia, C., and B. Uzzi. 2021. Quantifying the selective forgetting and integration of ideas in science and technology. *American Psychologist* 76:1067–1087.

Hao, Q., F. Xu, Y. Li, and J. Evans. 2026. Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature* 649:1237–1243.

Lu, C., et al. 2026. Towards end-to-end automation of AI research. *Nature* 651:914–919.

Pei, G., and H. Huang. 2025. Open science falling behind in the era of artificial intelligence. *Frontiers in Research Metrics and Analytics* 10:1595824.

Tang, G., and H. Cai. 2025. Citation contamination by paper mill articles in systematic reviews of the life sciences. *JAMA Network Open* 8:e2515160.

Brian Uzzi is the Richard L. Thomas Professor of Leadership at the Kellogg School of Management at Northwestern University. A member of the American Academy of Arts and Sciences, Uzzi studies the links between collaboration, AI, and human achievement. Julio M. Ottino is the McCormick Institute Professor of Engineering and a professor at the Kellogg School of Management at Northwestern University in Evanston, Illinois. He is also a member of the National Academy of Sciences and the National Academy of Engineering. Email for Uzzi: uzzi@kellogg.northwestern.edu