Cell
Leading Edge

CellPress

Perspective

Toward autonomous science with agentic artificial intelligence

Lucie Y. Guo,[1] Darren S.J. Ting,[2,3,4,5] Xinyi Su,[4,6,7] Alex Aliper,[8] Alex Zhavoronkov,[8] and Daniel S.W. Ting[4,5,9,*]
[1]F. M. Kirby Center, Scheie Eye Institute, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
[2]Academic Ophthalmology, Department of Inflammation and Ageing, School of Infection, Inflammation, and Immunology, College of Medicine and Health, University of Birmingham, Birmingham, UK
[3]Birmingham and Midland Eye Centre, Sandwell and West Birmingham NHS Trust, Birmingham, UK
[4]Singapore National Eye Center, Singapore Eye Research Institute, Singapore, Singapore
[5]Duke-NUS Medical School, National University Singapore, Singapore, Singapore
[6]Institute of Molecular Cell Biology, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore
[7]Department of Ophthalmology and Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University Hospital, Singapore, Singapore
[8]Insilico Medicine, International Renewable Energy Agency HQ, Masdar City, Abu Dhabi, UAE
[9]Byers Eye Institute, Stanford University, Palo Alto, CA, USA
*Correspondence: daniel.ting45@gmail.com
https://doi.org/10.1016/j.cell.2026.08.052

SUMMARY

Artificial intelligence (AI) in biomedicine has evolved from pattern-recognition for medical image analysis to generative models like AlphaFold that predict protein structures. Tools such as Elicit, GPT-Rosalind, and Claude Science now bridge prediction and reasoning through literature synthesis, domain-tuned models, and agentic workbenches. Three recent systems (Co-Scientist, Robin, and Biomni) demonstrated agentic AI capable of generating hypotheses, designing experiments, computational analysis, and refining reasoning through experimental feedback. This perspective traces the emergence of agentic scientific reasoning, discusses scaling laws and autonomous laboratories, and argues that AI's transformative potential is inseparable from challenges in hallucination, bias, irreproducibility, and dual use. Scientific progress will depend on responsible human-AI collaboration.

INTRODUCTION

Artificial intelligence (AI) in biomedical sciences has rapidly evolved from deep learning in medical imaging analysis to generative AI that includes large language models (LLMs), vision language models, ambient AI, and multimodal AI.[1,2] Convolutional neural networks (CNNs), which can process grid-structured data like images and video and extract spatial features like edges and textures, classified clinical data such as retinal images[3] and histopathology slides[4] with near-expert level accuracy by examining spatial structure in image data. Statistical models predicted protein-protein interactions from amino acid sequence and gene expression data.[5–7] Variant callers distinguished true genetic signal from sequencing noise,[8] and further models learned to predict pathogenicity of uncharacterized genetic variants.[9–11] Recently, foundation models and LLMs have demonstrated potential in supporting clinical reasoning. These advancements shared a common structure: a well-defined input mapped to a well-defined output, specified in advance by a human scientist who knew what question to ask.

In 2026, papers appeared simultaneously that mark an inflection point in this trajectory (Figure 1). Gottweis and colleagues introduced Co-Scientist,[12] a multi-agent system built on iterative hypothesis generation, tournament-style self-critique, and test-time compute scaling. Rather than answering a fixed question posed by a human investigator, Co-Scientist orchestrates multiple AI agents in "tournament"-style competition in which they assess two hypotheses at a time, scan the literature, and rank the hypotheses. In one validation, the system proposed drug repurposing candidates for acute myeloid leukemia (AML), including KIRA6, an inhibitor of the stress-responsive enzyme IRE1α that had not been previously proposed as a prepurposing candidate for AML; KIRA6 showed selective tumor inhibition in patient-derived cell lines, killing leukemia cells at concentrations up to 18-fold lower than those required to kill control cells. In another, it identified epigenetic targets for liver fibrosis with anti-fibrotic activity and evidence of regeneration in human hepatic organoids; in a third, it independently recapitulated a then-unpublished experimental finding on capsid-forming phage-inducible chromosomal island gene transfer in bacteria.

A contemporaneous multi-agent AI system, Robin, couples hypothesis generation directly to autonomous analysis of experimental data.[13] Robin is composed of three collaborating agents with avian-inspired names: Crow performs shallow literature scans, Falcon performs deep dives in literature, and Finch analyzes experimental data obtained from benchside testing. Robin was asked to focus on a disease of interest: dry age-related macular degeneration (dAMD), a leading cause of severe vision

Please cite this article in press as: Guo et al., Toward autonomous science with agentic artificial intelligence, Cell (2026), https://doi.org/10.1016/j.cell.2026.08.052

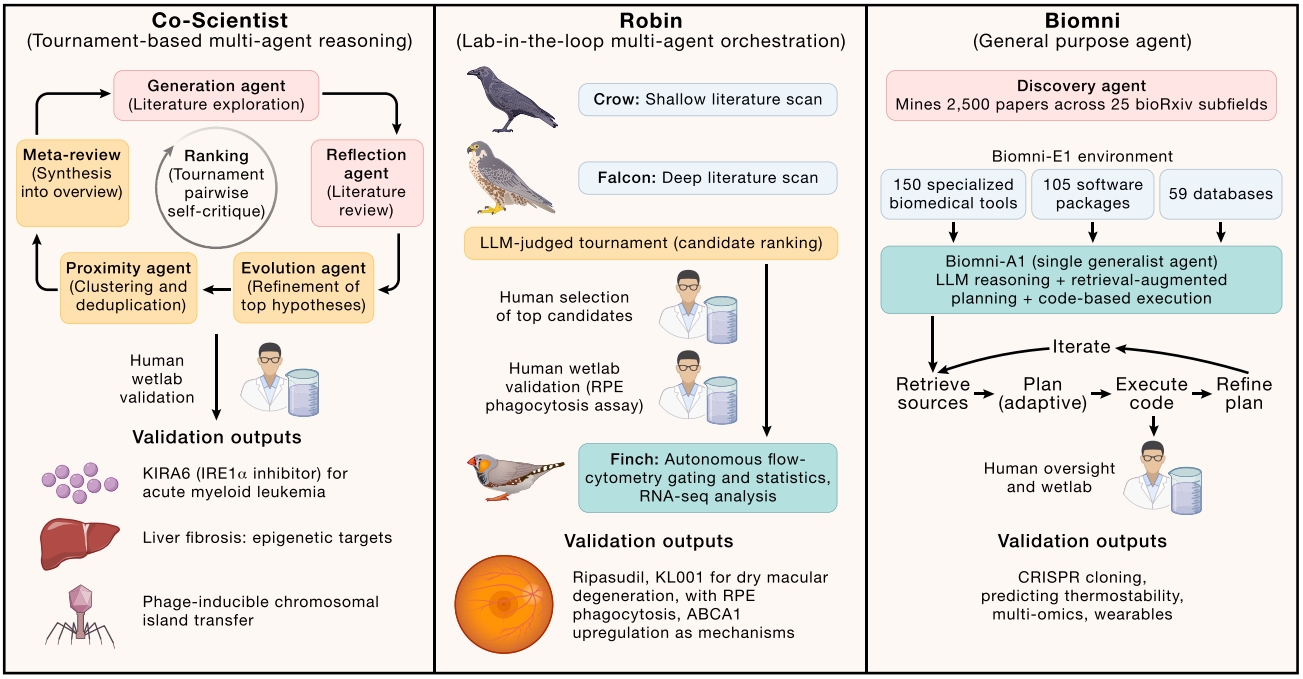


Figure 1. Anatomy of agentic scientific systems

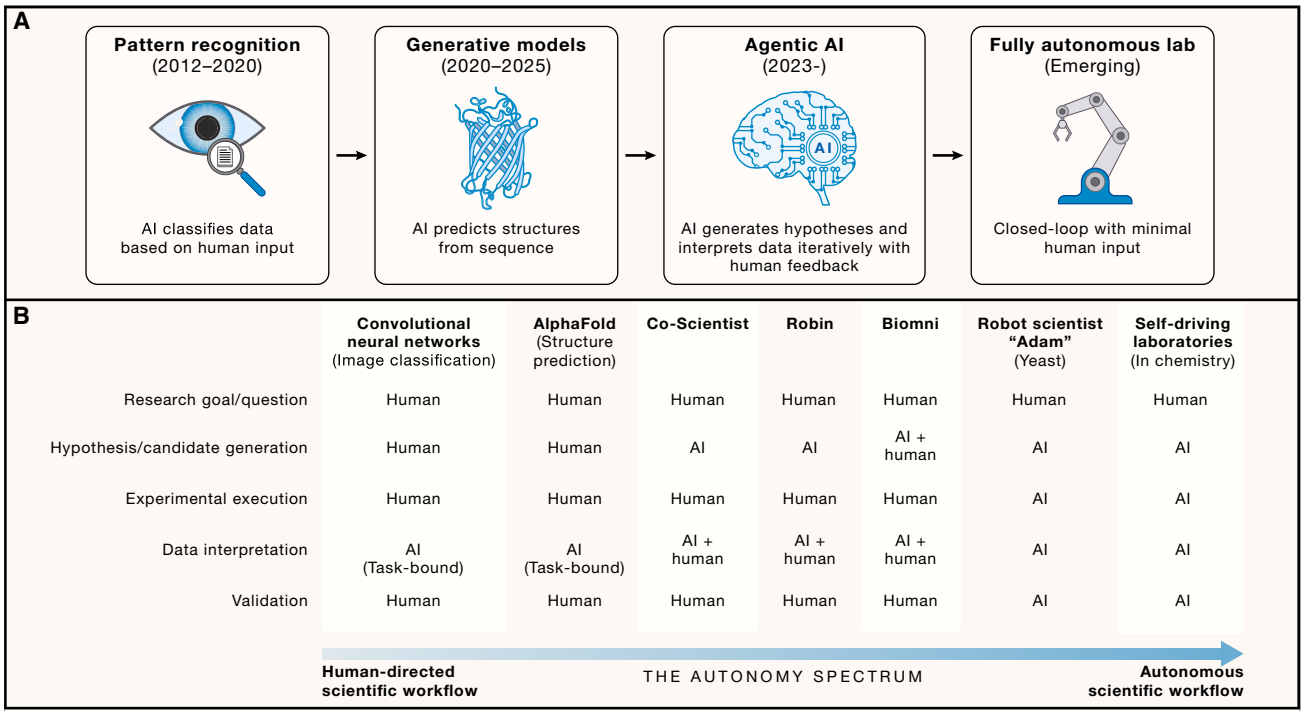
loss. Robin identified and reviewed 151 papers and proposed 10 biologically relevant dAMD mechanisms, ranked them and the corresponding experimental strategies, and proposed enhancing phagocytosis by retinal pigment epithelium (RPE) as a therapeutic strategy, as well as testing drugs to enhance phagocytic capacity with patient-derived or stem cell-derived primary RPE cells. In this pipeline, (1) Crow conducted a literature review of about 400 papers and proposed 30 existing drug candidates for experimental testing; (2) Falcon then comprehensively evaluated each of these molecules, and they were ranked in an LLM-judged tournament. A human selected the top 5 candidates from this ranking and tested them experimentally, and (3) Finch then gated the raw flow cytometry data and performed statistics. Through an iterative lab-in-the-loop discovery cycle, Robin identified ripasudil and KL001 as enhancers of RPE phagocytosis. After human experimentalists validated predictions in primary human RPE cells, Robin then autonomously analyzed the resulting RNA-sequencing data and proposed the upregulation of the lipid transporter ABCA1 as a mechanistic pathway.

Concurrently, Huang and colleagues introduced Biomni,[14] a general-purpose AI agent (a system that can interpret and execute a wide variety of open-ended tasks across multiple domains instead of one) (Figure 1). Rather than using multiple competing agents in a tournament, Biomni is a single agent that can reason over a curated environment (i.e., Biomni-E1, which had 150 specialized biomedical tools, 105 software packages, and 59 databases from mining 2,500 publications across 25 subfields on bioRxiv). When a user query is given, Biomni uses retrieval-augmented planning (which recalls past experiences or references external data, instead of relying on pre-trained knowledge) to select relevant resources, generate a step-by-step plan, and execute each step through code. This code-as-action interface lets the agent compose workflows dynamically, without predefined templates.

These three systems now go beyond retrieving or summarizing existing knowledge; they drive iterative cycles of hypothesis generation, experimental design, and data interpretation that begin to approximate aspects of human scientific reasoning itself. The authors of Robin argue that its ability to effectively generate novel, testable hypotheses by combinatorially synthesizing non-obvious connections between disparate fields is useful for finding "low-hanging fruits" that may elude human researchers with compartmentalized knowledge. This architectural diversity among these systems (multi-agent tournament versus single-agent code execution) indicates that the field is exploring multiple viable paths in building an artificial scientist.

This perspective traces the arc from supervised pattern recognition through generative models to agentic scientific reasoning, examines whether scaling laws transfer from language modeling to discovery, considers the prospect of closed-loop autonomous laboratories, and highlights the scientific and ethical risks that accompany this transition. We acknowledge that this arc from pattern recognition toward autonomy represents one framing of the field's trajectory, shaped by the recent emergence of agentic systems, but alternative narratives also exist that emphasize increasingly capable human-AI collaboration instead of autonomy or data infrastructure instead of reasoning capability. Here, we prioritize the narrative of agentic scientific reasoning because the systems that we examine (Co-Scientist, Robin, and Biomni) make the question of autonomy especially timely.

Please cite this article in press as: Guo et al., Toward autonomous science with agentic artificial intelligence, Cell (2026), https://doi.org/10.1016/j.cell.2026.08.052



	Convolutional neural networks (Image classification)	AlphaFold (Structure prediction)	Co-Scientist	Robin	Biomni	Robot scientist "Adam" (Yeast)	Self-driving laboratories (In chemistry)
Research goal/question	Human	Human	Human	Human	Human	Human	Human
Hypothesis/candidate generation	Human	Human	AI	AI	AI + human	AI	AI
Experimental execution	Human	Human	Human	Human	Human	AI	AI
Data interpretation	AI (Task-bound)	AI (Task-bound)	AI + human	AI + human	AI + human	AI	AI
Validation	Human	Human	Human	Human	Human	AI	AI

Figure 2. Evolution of AI-assisted science

(A) Evolution of AI-assisted science from pattern recognition toward autonomous laboratories.

(B) A comparison of various systems on the autonomy spectrum, illustrating the respective contributions of AI and human investigators at each stage of the discovery pipeline.

THE ERA OF PATTERN RECOGNITION

The first wave of biomedical AI, spanning roughly 2012 to 2020, was defined by supervised learning applied to single, well-specified tasks (Figure 2). Deep CNNs became useful in skin lesion classification and diabetic retinopathy screening,[3,4,15] at least in identifying high-risk pathology that warranted referral to a specialist. In genomics, machine learning models learned to call genomic variants from short-read data with high accuracy[8] and predict the pathogenicity of uncharacterized variants. In drug discovery, quantitative structure-activity relationship models predicted whether a candidate molecule would bind a given target,[16] but this task still required a human medicinal chemist to have already proposed the molecule and identified the target worth pursuing. In this first era, the human investigator provided the scientific question, the relevant data, and the criteria to differentiate success from failure, and the model performed the pre-determined tasks.

In the field of biological structure prediction, AlphaFold,[17] which first competed in the Critical Assessment of Protein Structure Prediction (CASP) in 2018 and achieved near-experimental accuracy[18] with AlphaFold2 at CASP14 in 2020, marked a transition from classification to generative modeling. For decades, predicting the three-dimensional fold of a protein from its amino acid sequence had been one of the central unsolved problems in computational biology. RoseTTAFold[19] extended the transformer paradigm with a three-track neural network integrating one-dimensional sequence, two-dimensional distance, and three-dimensional coordinate information to produce predictions approaching AlphaFold2 accuracy while running on a single workstation. ESMFold used the ESM-2 protein language model to predict structures directly from single sequences without multiple sequence alignments.[20] AlphaFold's near-experimental accuracy across a substantial fraction of the proteome, combined with the decision to release predicted structures as a public resource,[21] demonstrated that AI could function as a generator of biologically actionable knowledge instead of just a faster classifier of it. AlphaFold 3's extension to protein-ligand, protein-nucleic acid, and antibody-antigen complexes broadened its impact beyond protein structural prediction.[22]

AlphaFold was the most visible example of a broader shift toward generative and foundation models in biology. Protein language models such as ESM-2 learned rich representations of sequence space from billions of proteins,[20] which enable not only structure prediction, but also function annotation and variant effect prediction without task-specific training. Foundation models extended this paradigm to other modalities including single-cell multi-omics,[23,24] while generative models for protein design (e.g., RFdiffusion[25]) and small-molecule generation moved from predicting biological entities to designing novel ones. These developments expanded the range of questions that AI can address, but each still operated within a fixed boundary defined in advance by a human. The model generated a structure, a sequence, or a molecule when asked, but did not decide what to generate, why it mattered, or what to do

Please cite this article in press as: Guo et al., Toward autonomous science with agentic artificial intelligence, Cell (2026), https://doi.org/10.1016/j.cell.2026.08.052

with the result. The agentic systems of the next era attempt to cross this boundary.

THE AGENTIC ERA

A second wave of tools emerged that began to assist with the reasoning process itself. Literature and evidence-synthesis assistants such as Elicit[26] applied semantic search and language models to screen, extract, and synthesize findings across tens of millions of papers, accelerating systematic review. Domain-tuned reasoning models such as GPT-Rosalind, a frontier model specialized for life sciences, are designed to reason across molecules, proteins, genes, and pathways to support hypothesis generation and experiment planning. Agentic research workbenches such as Claude Science orchestrate numerous scientific databases and tools within a single environment.

Co-Scientist[12] and Robin[13] are multi-agent architectures that can formulate questions themselves. They are collections of specialized model instances that take on distinct functional roles: generating candidate hypotheses, critiquing and ranking them in tournament-style competitions, retrieving and synthesizing relevant literature, and proposing concrete experimental protocols. These communicate iteratively rather than producing a single forward pass from input to output. This architecture allows the system to allocate additional computation not only to making a single answer more accurate, but to exploring a wider hypothesis space and refining candidates through self-critique before committing to a final proposal. Co-Scientist runs pairwise comparisons between hypotheses using simulated scientific debates and scoring them with an Elo-based system, which is a method for ranking competitors from pairwise comparisons originally designed for chess. Here, the ''competitors'' are the hypotheses, and the ''matches'' or ''tournaments'' are pairwise contests in which a Ranking agent judges which of the two hypotheses is the most promising. This tournament structure provides the selection pressure that converts generative quantity into quality, and the win/loss patterns of the different hypotheses are fed back to drive iterative improvement. This is the mechanism through which test-compute scaling can translate into better hypotheses, not just more hypotheses. This allows the system's output to take a form that more closely resembles scientific reasoning: not a single prediction, but a ranked set of hypotheses, each accompanied by supporting evidence, an account of its novelty relative to the existing literature, and a proposed experimental path for resolving the remaining uncertainty.

With Robin, the system enabled both hypothesis generation as well as data interpretation. After proposing ripasudil as a repurposing candidate for dAMD, the system did not stop at the prediction. It designed the follow-up transcriptomic experiment, and after human collaborators executed the wet-lab assay, Robin autonomously analyzed the resulting RNA-seq data and identified ABCA1 upregulation as a mechanistic pathway.[13] The human collaborator still plays a substantial role in the iterative process. Several of Robin's experimental suggestions were overridden by human scientists who judged them impractical, and the analytical agent performed less well on questions that required statistical and bioinformatic reasoning, relying heavily on human-supplied prompts. In Co-Scientist as well, a human scientist still executes the physical wet-lab experiments and retains authority to evaluate, accept, or discard the system's proposals throughout. Neither system has yet closed the loop fully, from hypothesis through physical experimentation back to model update, without human intervention at each stage.

Biomni represents a third architectural paradigm.[14] Instead of having multiple agents in competition, Biomni is a single general-purpose agent that composes workflows over a single biomedical environment. Its action-discovery agent mined 2,500 publications across 25 bioRxiv subfields and yielded Biomni-E1, with specialized tools, 105 software packages, and 59 databases. Biomni uses code as its primary way to interact with scientific tools and databases; each step of a plan is expressed as executable code, and it can perform complex, multi-step procedures such as running a process repeatedly, running multiple operations simultaneously, or making decisions based on data. The agent employs adaptive planning where it formulates an initial plan grounded in biomedical knowledge, then iteratively refines it through execution to enable context-aware behavior. Where Co-Scientist's tournament provides selection pressure that converts generative quantity to quality through pairwise competition, Biomni's code-based system provides the ability to compose arbitrary workflows across diverse tools.

A remaining gap is the connective execution layer that translates computational reasoning into physical experimentation and returns structured results to the agent. This would be the difference between an agentic co-scientist and an autonomous laboratory. General-purpose agent runtimes like OpenClaw can provide persistent agent loops capable of planning, calling tools, or executing code and could in principle drive laboratory hardware, laboratory information management systems (LIMSs), and other infrastructure under human oversight. Biomni's integration with PyLabRobot,[27] which generates executable liquid-handling protocols for Hamilton STAR robotic systems from natural-language experimental specifications, is a complementary approach to this gap of generating executable automation code rather than providing a runtime orchestration layer. These approaches start to bridge computational scientific reasoning and physical experimentation.

In all systems, human curation is still important. A small fraction of candidates, chosen by human scientists from the system's ranked output, were carried forward to bench testing. These were ranking-and-selective validation paradigms, not a comprehensive assessment of the system's predictive accuracy. Co-Scientist explored 2,300 approved drugs across 34 cancer types for repurposing, but only a few candidates selected by humans were tested *in vitro*; 3 showed some activity, and 1 showed particular promise.[12] Robin generated 30 therapeutic candidates for dAMD, tested 15 across two iterations, and confirmed 2 hits.[13] Biomni's benchmark tasks used ground-truth labels, but its open-ended case studies still involved human selection of what to test.[14] These all support the value of the system-plus-human pipeline, not necessarily the autonomous system itself. When expert reviewers select which hypotheses to advance from a generated pool, the success also reflects human judgment to spot a needle in a haystack, rather than the system's ability to prioritize correctly. Without knowing

Please cite this article in press as: Guo et al., Toward autonomous science with agentic artificial intelligence, Cell (2026), https://doi.org/10.1016/j.cell.2026.08.052

whether the experimentally validated candidates were the system's top-ranked proposals or human overrides, and without evaluating how these systems behave when pointed at problems with no viable answer, we cannot yet assess whether these successes reflect a generalizable capability versus broad generation paired with selective human curation. This is not a criticism of the breakthroughs themselves, which appropriately focused on proving that agentic hypotheses can survive experimental validation, but it represents a methodological gap the field will likely fill as systems advance toward higher automaticity. Although Biomni-Eval1 includes direct comparisons to human expert performance on benchmark tasks with known answers, it is less known how agentic AI will perform against human experts asked to generate and rank candidates under the same constraints. Studies in natural language processing (NLP) ideation had shown that LLM-generated research ideas were judged to be more novel at the ideation stage,[28] but then judged to be less exciting after execution.[29] Without such rigorous comparisons, it is unclear whether newer agentic systems will outperform human scientists at hypothesis prioritization, or whether they primarily accelerate the process by generating a larger, more well-informed, or more intensely curated candidate pool from which human expertise can select.

Another gap is the absence of comparison to domain-specific computational baselines. Drug repurposing is not a new problem; decades of computational biology research have produced targeted methods, such as network-based approaches that predict drug-target interactions and drug-disease proximity on the human interactome,[30,31] signature matching that connects disease- and drug-induced gene expression profiles,[32,33] and graph neural networks trained on drug-target interaction data.[34] Head-to-head evaluation against domain-specific baselines would help to evaluate claims that agentic AI is a new paradigm for biomedical discovery. Moreover, the biological significance and novelty of the successful validations remain an open question. Tumor inhibition in patient-derived cell lines is a low bar: many compounds have shown anti-tumor activity *in vitro* that fails to translate to clinical efficacy, and many promising preclinical candidates fail in phase II or beyond.[35] Anti-fibrotic activity in hepatic organoids is similarly suggestive, but far from conclusive. Along similar lines, Rho-associated protein kinase (ROCK) inhibition has long been known to be effective for enhancing RPE phagocytosis,[36] and ripasudil as nominated by Robin had already been actively repurposed within ophthalmology in diabetic macular edema[37] and retinal ganglion cell neuroprotection.[38] Agentic systems will continue to mature; it will soon be seen whether AI systems can robustly make novel biomedical discoveries that survive the transition through *in vivo* models and human clinical trials.

IS MORE AUTONOMY BETTER?

The aforementioned methodological gaps raise a broader question of whether more autonomy is always desirable. In many contexts, AI may not be most helpful as autonomous investigators but as bounded reasoning components, constrained to specific decisions and interpretable in their outputs. A recent study[39] supports this alternate architecture: rather than using multi-agent architecture, the authors used three frontier LLMs (GPT-4o, Claude-3.7, and Gemini-2.5-Pro) as bounded reasoning components within a fixed, human-designed pipeline for chimeric antigen receptor (CAR)-T cell target discovery. The models optimized feature weights for a multi-criteria scoring system and nominated the most promising targets from a shortlist, each across 1,000 independent simulations. The pipeline itself (single-cell atlas construction, feature extraction from six public databases, and all experimental validation) was entirely human, and the LLMs did not generate hypotheses *de novo*, retrieve literature, or propose experiments. After glycoprotein non-metastatic melanoma protein B (GPNMB) emerged as the most frequently nominated target across all three models, the authors engineered a CAR construct and demonstrated anti-tumor activity *in vivo* across three xenograft models.

Two reports from independent groups appeared contemporaneously, both validating GPNMB as a CAR-T target through non-AI approaches: a proteogenomic surfaceome screen in microphthalmia/transcription factor E (MiT/TFE)-fusion-driven sarcomas[40] and a multi-omic profiling platform in glioblastoma.[41] The sarcoma study[40] reported a first-in-human clinical case (ClinicalTrials.gov: NCT07104682) in which GCAR1, the same GPNMB-directed CAR construct, produced stable disease for up to 3 months, a 52% decrease in lung lesion counts, and no cytokine release syndrome or neurotoxicity. The glioblastoma study[41] showed durable disease control beyond 160 days in orthotopic patient-derived xenograft models. The convergence of an LLM-driven pipeline and two conventional computational approaches on the same target is strong evidence for the validity of GPNMB, and this was achieved by using AI as a narrow, bounded component within an otherwise fully human-designed pipeline or by not using AI for target discovery at all. The appropriate level of autonomy is itself an empirical question, and a narrowly scoped model embedded in a human-designed pipeline can yield robust, experimentally validated discoveries without requiring end-to-end AI control.

More autonomy may not necessarily drive more impactful discoveries. Recent work[42] distinguishes three modes of inference in scientific discovery: induction (finding patterns in data), deduction (deriving consequences from established premises), and abduction (the creative leap to a new explanatory hypothesis). Current AI systems are mastering induction through statistical learning and are advancing in deduction, but they arguably lack the capacity for abduction, which is the generation of genuinely novel thought rather than recombinations of existing knowledge. LLMs operate over text without sensory or physical grounding, and LLMs may be structurally incapable of creating new foundational axioms.[42] However, biomedical discovery can be largely inductive, driven by large datasets rather than scarce observations, and worthwhile hypotheses can often be derived from connecting existing knowledge across domains rather than inventing entirely new axioms. Nonetheless, the most transformative biomedical discoveries usually require conceptual leaps that go beyond pattern matching in existing data. Whether current architectures can perform such leaps, or whether they require physical grounding that emerging multimodal models may provide, remains an open question.

Please cite this article in press as: Guo et al., Toward autonomous science with agentic artificial intelligence, Cell (2026), https://doi.org/10.1016/j.cell.2026.08.052

SCALING LAWS

To evaluate increasingly capable AI co-scientists, it is helpful to consider scaling laws: the observation that AI systems tend to be more capable in predictable ways as they are trained on larger datasets, use more computing power, and contain more parameters.[43,44] Early work focused on "training-time compute," the computational resources used to build and train a model before it is deployed. Recent work has identified sub-scaling regimes in which returns diminish once training corpora saturate or become redundant.[45] More recently, researchers have recognized the importance of "test-time compute," which is the computational effort a model expends while solving a specific problem after training.[46] Rather than producing an immediate answer, a model can devote additional computation to exploring alternatives, evaluating evidence, checking reasoning, and refining potential solutions. In parallel, a growing body of work demonstrates that agents can also improve over time through online adaptation, refining their own prompts, memory, and workflows across successive tasks without retraining the underlying model.[47] This ability to allocate computational resources during problem-solving has emerged as a major driver of recent advances in AI reasoning and agentic systems.

In a reasoning system equipped with test-time compute scaling, additional computation translates into additional deliberation: the system can explore a broader region of the hypothesis space, generate and discard more candidate solutions, and iteratively refine a proposal before committing to a final answer. This was the logic that distinguished AlphaGo's advance over earlier game-playing systems[48] and allowed it to defeat world champion Go players. The board game Go is an enormous search space, with many possible move sequences to evaluate exhaustively. AlphaGo combined neural networks with Monte Carlo Tree Search, which involves building a decision tree through random sampling and simulations. Rather than relying on a single prediction, it searched through competing possibilities before deciding on a move. Similarly, Co-Scientist's tournament-based, self-critique method spends computation at the moment of reasoning, instead of producing a single response in one forward pass. In both cases, improved performance arises from additional computation to evaluate alternative possibilities before committing to a decision. But AlphaGo is not an analogy for scientific discovery. Go possesses a perfect objective function of winning the game, as well as a fully specified and bounded state space. Scientific discovery has neither: there is no unambiguous signal that distinguishes a good hypothesis from a bad one prior to experimental testing, and the space of possible hypotheses is open-ended and undefined. More search in Go produces better moves because the evaluation function is ground truth, and while more search in science may produce more candidate hypotheses, it is not a guarantee that more hypotheses will lead to better science.

In the pattern recognition era, a classifier trained to recognize diabetic retinopathy does not become a better scientist if given more compute at inference time; it simply applies the same fixed function more quickly. A reasoning system built around test-time search, by contrast, becomes capable of qualitatively different behavior—deliberation, self-critique, and iterative refinement—as more computation becomes available to it.[46] Biomni's adaptive planning is an alternative mechanism for the same principle: test-time compute is allocated to iterative plan refinement and code debugging, rather than tournament-style hypothesis competition. By optimizing Biomni-R0 through trial-and-error training, expanding the underlying model from 8 billion to 32 billion parameters yielded a 10% absolute improvement on the Biomni-Eval1 benchmark tasks. This shows that model scaling improves performance on bounded biomedical tasks with defined correct answers, but it does not directly establish that comparable gains will translate to open-ended hypothesis generation, where no ground-truth signal is available to guide optimization.

Scaling laws in language modeling operate over a relatively smooth, well-defined space of token sequences, where more compute reliably yields more fluent and more capable output.[43,44] The hypothesis space in biology is neither smooth nor well-defined: it contains decades of accumulated knowledge that is incomplete, sometimes contradictory, and unevenly distributed across domains. A system that explores more of this space is not guaranteed to find better hypotheses in it, because the relationship between search effort and hypothesis quality depends on the quality of the prior knowledge that constrains the search. In domains where the literature is sparse or unreliable (such as rare diseases, understudied organisms, non-coding regulatory biology), more searching may simply amplify noise rather than converge on truth. The tournament-based self-critique that drives refinement in systems like Co-Scientist assumes that the critic agents can reliably distinguish good hypotheses from bad ones, but if the training data for those critics is itself biased or incomplete, the tournament may converge confidently on the wrong answer rather than slowly on the right one. Biomni's action-discovery approach introduces a related concern of recency bias: by mining only the most recent papers per bioRxiv category to refine its environment, perhaps the system would overrepresent trending methods and underrepresent foundational techniques that have faded from current discourse.[14]

Many agentic systems are currently reasoning primarily from the scientific literature— the vast body of text through which humans describe and interpret the natural world. This is a powerful substrate and allows the models to synthesize findings across disciplines, identify gaps in existing knowledge, and propose connections that individual investigators may overlook. But language is an imperfect proxy for biological reality, and the scientific literature is ultimately an interpretation of experimental observations, rather than the observations themselves. Out of necessity, LLMs compress, summarize, and abstract quantitative measurements, whereas structured biological data (such as genomic sequences, protein structures, transcriptomic profiles, cellular microscopy, or functional assays) preserve information that may never be fully captured in text. A system that reasons primarily over words about proteins or pathways can connect published claims, but it has limited ability to evaluate those claims against the underlying experimental evidence from which they were derived. It remains an open question whether the scale laws that govern text generation will apply to higher-level scientific reasoning.

Please cite this article in press as: Guo et al., Toward autonomous science with agentic artificial intelligence, Cell (2026), https://doi.org/10.1016/j.cell.2026.08.052

The frontier for AI-driven scientific reasoning is in multimodal systems that ground linguistic reasoning in quantitative biological data, linking the concepts and relationships expressed in text to the sequences, structures, and measurements that constrain their interpretation. Such multimodal grounding would enable these systems to navigate not only what has been written about biology but also biology itself. Much of the quantitative content of the scientific literature is in tables and figures instead of text, and vision language models can interpret published figures and answer questions about them[49,50]; there are efforts in semantic annotation of figures and their underlying source data,[51] all for turning published results into analyzable and trainable datasets. Fully autonomous robotic laboratories would offer another route to this grounding, because they generate structured, instrument-level measurements that are not mediated through narrative publication. Images, plate-reader traces, liquid-handling logs, reagent metadata, passage number, incubation history, and failed runs can all become part of the model's memory when captured in machine-readable form. In this sense, the relevant scaling axis for scientific AI may increasingly include not only training- and test-time compute but also the rate at which grounded experimental observations can be produced, standardized, and returned to the reasoning system.

In benchmarking AI-assisted science, it is helpful to consider Goodhart's law, which states that when a measure becomes a target, it ceases to be a good measure. The scaling laws governing language models assume a simple training task: given a passage of text, the model predicts the next word, and its "loss" is a numerical score of how wrong that prediction is. The model is trained to minimize this loss, and researchers evaluate whether a model has improved by measuring this same loss. There is no gap between the training signal and true objective; therefore, scaling works cleanly because more compute reliably produces better models, as "better" is defined unambiguously. However, in scientific discovery, the true objective of generating hypotheses that lead to real discoveries can only be confirmed through experiments that often take weeks or months to execute, and even longer to replicate. In the interim, any system must optimize a proxy, such as internal consistency, literature support, and novelty scores. These proxies will be imperfect, but what will be even more problematic is that the optimization pressure will specifically favor hypotheses that score well on whatever is measurable. Therefore, the relationship between compute and meaningful discovery in science will likely be fundamentally less predictable than in domains where the objective function is self-validating.

TOWARD THE AUTONOMOUS LABORATORY

Perhaps the most consequential near-term development suggested by this convergence is the autonomous, or self-driving, laboratory: a closed-loop system in which AI generates hypotheses, robotic platforms execute experiments, data are analyzed automatically, and the results are used to refine the next iteration of hypotheses generation with minimal human intervention.[52,53] Self-driving laboratories are already operational in chemistry and materials science, where robotic platforms synthesize and characterize hundreds of candidate compounds under the direction of machine learning systems that continuously select the next experiment based on results from the previous iterations.

The same transition is now beginning in biology, where the experimental substrate is noisier but the value of closed-loop iteration is high. A closed-loop system autonomously executed cell-based senescence assays and identified pharmacological inhibition of TRAF2- and NCK-interacting kinase (TNIK) as a senomorphic intervention, with the AI-driven robotics platform running the hypothesis-to-experiment cycle rather than a human bench scientist.[54] TNIK was independently surfaced as a target by a separate target-discovery approach: TNIK is the target of rentosertib (ISM001-055), an AI-discovered and -designed candidate that has reported phase 2a results in idiopathic pulmonary fibrosis.[55] Two independent pipelines, one an autonomous robotics laboratory operating on senescence phenotypes and the other a target-discovery engine operating on disease biology, converged on the same protein, which is an example of the closed hypothesis-to-experiment loop producing a hypothesis that can be tested in a clinical-stage program.[54]

Co-Scientist and Robin represent partial realizations of this broader vision. The systems generated hypotheses, proposed experimental designs, and interpreted the resulting data, while human investigators performed laboratory work.[12,13] Biomni extends this trajectory by generating executable wet-lab protocols: its PyLabRobot integration[14,27] produced liquid-handling code for Hamilton STAR systems (including deck configuration, error handling, and volume validation) from natural language descriptions. But the protocols are generated and executed separately, without the continuous feedback loop that defines a true autonomous laboratory.

These systems represent a reorganization of the traditional division of labor in scientific research. Historically, laboratory automation has focused on repetitive physical tasks (pipetting, imaging, sequencing, liquid handling) while cognitive tasks such as hypothesis generation, experimental design and interpretation were human endeavors. These recent developments suggest that the traditional boundaries between cognitive and physical labor in science are being redrawn. Agentic AI now inverts this pattern by automating elements of scientific reasoning, even before the corresponding physical work has been fully automated. The autonomous laboratory would not simply be another form of automation, but the convergence of these two trajectories into a closed-loop system, coupling agentic reasoning systems directly to robotic experimental platforms,[56] moves the human handoff between hypothesis and experimental execution. The practical mechanism for this coupling is now taking shape in the form of open agentic operating systems such as OpenClaw and OpenClaw-like runtimes. These systems provide the orchestration layer that connects language-model reasoning to liquid handlers, imagers, incubators, LIMS, and analysis pipelines, so that a proposed experiment can be scheduled, executed, and interpreted without a manual handoff at each step. The TNIK senomorphic study[54] is an early demonstration of what this coupling yields when a reasoning system and a robotic platform operate as one loop rather than as two stages separated by human transcription.

The limiting factor in biological discovery has rarely been the speed at which scientists generate hypotheses; it has been the

Please cite this article in press as: Guo et al., Toward autonomous science with agentic artificial intelligence, Cell (2026), https://doi.org/10.1016/j.cell.2026.08.052

speed at which the hypotheses can be tested. *In silico*, iteration is usually rapid and inexpensive. Biology, by contrast, imposes substantial experimental cost: cells must be cultured, reagents purchased, protocols tested and optimized. Reasoning systems without rapid experimentation will remain constrained by wet-laboratory throughput. The most transformative AI systems for biology may not be ones that reason most cleverly, but the ones that iterate the fastest between each hypothesis-to-experiment cycle. On this measure, a fully autonomous robotics laboratory is the decisive variable, because it removes the human scheduling latency that separates a proposed assay from its execution. In the senescence study,[54] the advantage came not from a more sophisticated hypothesis but from an agentic layer that could run, read, and re-plan cell-based assays continuously. As these orchestration layers mature and connect to standardized instrument application programming interfaces (APIs), the binding constraint shifts from reasoning quality toward the physical cycle time of the underlying biology.

There are technical barriers to realizing this vision. Early autonomous biology platforms have demonstrated closed-loop experimentation in constrained settings, from Ross King's robot scientists[57–59] to recent LLM-driven cell culture systems.[60–62] Robotic platforms in materials and synthetic chemistry operate within comparatively well-defined reaction spaces and relatively predictable, binary failure modes.[52,53,63] Autonomous biology would operate in a fundamentally more complex environment, often relying on living, evolving systems. Biological systems are vulnerable to experimental noise, microenvironmental fluctuations, and genetic drift. Batch effects can introduce artifacts that are difficult to distinguish from genuine biological signals,[64,65] and biological systems (particularly *in vivo* animal models) exhibit intrinsic variability. These challenges will become important in autonomous closed-loop systems. A hallucinated hypothesis may waste a single experiment, but a hallucinated or erroneous interpretation may be silently incorporated into the next cycle of hypothesis generation, and errors may be propagated through the discovery process and corrupt entire trajectories of experiments. Until we build interoperable, unified data fabric capable of contextualizing biological noise, the throughput of autonomous laboratories will remain bottlenecked by the chaotic realities of live biology, not by the limits of AI reasoning.

Another challenge is experimental data infrastructure. The autonomous laboratory assumes that experimental data flow cleanly from instruments to computational models based on standardized experimental protocols; in practice, biological data are often fragmented across incompatible software platforms, produced through heterogeneous or non-standardized protocols, stored in proprietary formats, and accompanied by tacit knowledge that is difficult for machines to read. Valuable experimental evidence, especially unsuccessful studies and negative results, frequently remain inaccessible to future investigators and learning systems. The compounding loop that the autonomous laboratory promises requires structured data and persistent memory that the industry has not yet fully built. Fully autonomous discovery will require the convergence of increasingly capable AI reasoning, laboratory robotics, interoperable data infrastructure, and human oversight into a unified scientific ecosystem. An open agentic operating layer could bind these components together, provided its orchestration logic is inspectable rather than proprietary. If the sequence of hypotheses, instrument calls, and data-ingestion steps is logged in a form that other groups can read and replay, then the persistent memory the autonomous laboratory requires can accumulate as a reproducible record rather than as tacit institutional knowledge.

Whether the field will converge on a common orchestration standard remains to be seen. OpenClaw represents one emerging approach that provides an open agent runtime for coordinating tools, tasks, and multi-step workflows. Biomni generates and executes code that interfaces with laboratory automation through PyLabRobot.[14] These approaches may prove complementary, but standardization may face practical barriers. Laboratory instruments from different manufacturers may have proprietary or incompatible control interfaces, while biological experimentation adds context-dependent variability and failure modes beyond those encountered in more constrained automated domains. Historically in genomics, convergence may coordinate incentives through community standards, such as consortia and funder policies. In the interim, the case for open orchestration frameworks may rest less on the maturity of any particular implementation than on a broader principle: as autonomous experimentation becomes more capable, inspectable and reproducible orchestration will become increasingly important for scientific transparency.

RISKS AND CHALLENGES

Hallucinations

LLMs, including those underlying current agentic scientific systems, are susceptible to hallucinations (Figure 3): they can generate plausible claims or citations that have no basis in reality.[66–68] When AI-generated hypotheses are used to design wet-lab experiments, such errors could result in wasted reagents, unnecessary animal use, and lost research time and expenses. Many multi-agent architectures, including tournament-style self-critique frameworks such as Co-Scientist, may create an illusion of consensus. If multiple ''critic'' agents are instances of the same underlying model operating under similar prompts, apparent agreement may simply reflect shared biases or blind spots, rather than truly independent evaluation. Agreement among AI agents should not be interpreted as evidence of correctness. Grounding hypotheses in retrieved, verifiable sources and requiring explicit citation of primary evidence can reduce, though not eliminate, the rate at which unsupported claims enter the experimental pipeline. Experimental validation remains essential and cannot be substituted by model self-assessment. Biomni's use of executable code may reduce some types of hallucination, by allowing code errors to be identified during execution and by deriving results directly from database queries and computational analyses, rather than from text generation alone. However, this potentially introduces a different class of error: a workflow may execute successfully while remaining scientifically inappropriate, or an agent may select an incorrect statistical test, apply a method outside its intended context, or misinterpret the output of an

Please cite this article in press as: Guo et al., Toward autonomous science with agentic artificial intelligence, Cell (2026), https://doi.org/10.1016/j.cell.2026.08.052

CellPress

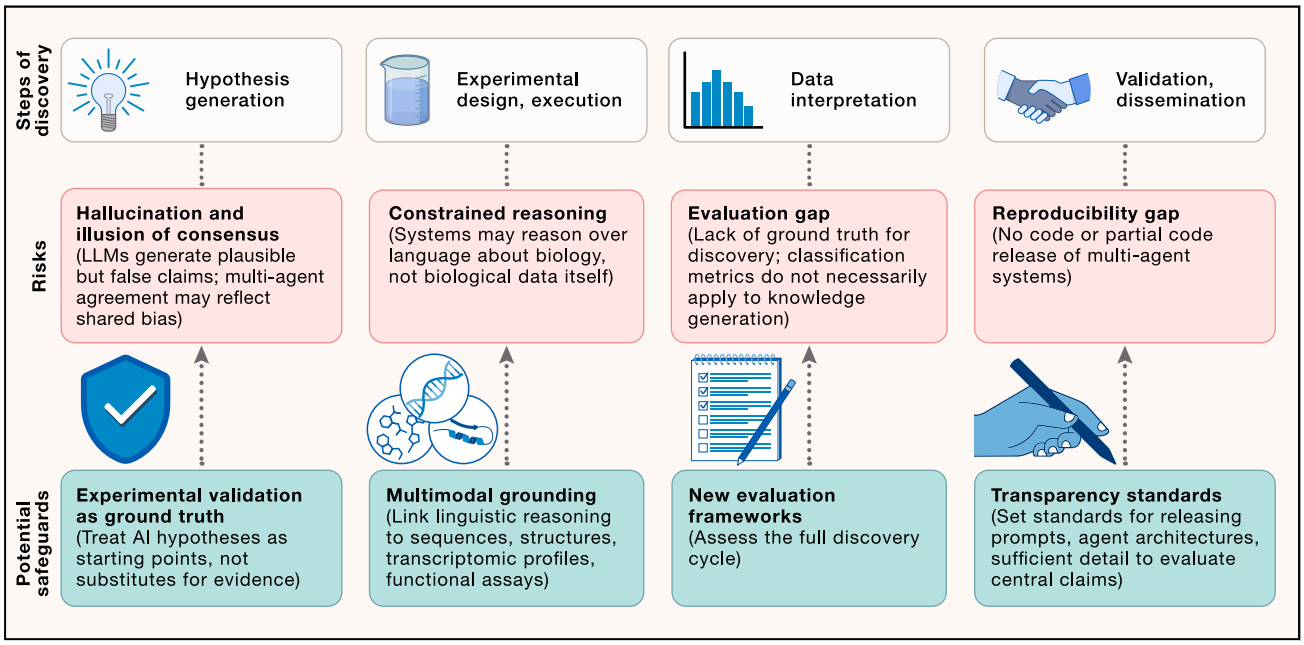


Figure 3. Risks and proposed safeguards of agentic AI across the scientific discovery cycle

analytical tool. Successful execution of code still does not guarantee scientifically sound conclusions.

At the same time, these limitations are not unique to AI. Human scientists are likewise susceptible to cognitive biases, including confirmation bias, groupthink, publication incentives, and orthodoxies within disciplines that can shape how evidence is interpreted. Scientific progress has never relied on the infallibility of individual investigators, but rather on the self-correcting capability of the scientific enterprise. Independent replication, peer review, open debate, and competition among research groups are mechanisms through which erroneous conclusions can be identified and corrected over time. As autonomous AI systems become increasingly integrated into scientific discovery, AI-generated hypotheses should be viewed as starting points for rigorous empirical testing instead of substitutes for experimental evidence.

Gap between biological knowledge and reality

A challenge that is distinct from hallucination is the fundamental incompleteness and context-dependence of biological knowledge itself. Even when a model generates claims faithfully grounded in the literature without fabrication, those claims may not hold in the experimental context in which a hypothesis is tested. A gene's function may differ drastically across cell types, organisms, or disease states, and a model that synthesizes findings across various contexts may produce a hypothesis that is well supported by published evidence but empirically irrelevant in a specific tested setting. Biological systems often resist the simplified pathway representations used to describe them (such as feedback loops, compensatory mechanisms, and redundancy), and even an accurate synthesis of available knowledge may inaccurately predict how a system will respond to perturbation. Fidelity to the scientific record is not equivalent to fidelity to biological reality, which is much more complex. Internal consistency and accurate citation are necessary but insufficient conditions for a hypothesis to be biologically correct.

Evaluation

The metrics by which earlier biomedical AI was assessed—sensitivity, specificity, area under the receiver operating characteristic (ROC) curve—were inherited from classification tasks with well-defined "ground truth," in which each prediction can be compared against a known correct answer, such as an expert clinician's diagnosis, a pathologist's read, or a verified clinical outcome.[3,4,8] For scientific agentic AI, the "ground truth" must be established through the very experiments that the hypothesis motivates. This is a conceptual difference between AI for diagnosis versus AI for discovery. Diagnostic AI is fundamentally a prediction problem with known labels, whereas scientific AI is a hypothesis-generation problem in which the correctness of the claims is initially unknowable. Before experimental testing, researchers must rely on proxy indicators of quality—such as novelty, biological plausibility, and experimental tractability—to prioritize hypotheses for validation.

Existing benchmarks for biomedical AI agents fall into two categories. The first evaluates analytical competence: the ability to execute well-defined computational tasks such as querying biological databases, reasoning over sequences, or prioritizing genetic variants, where each question has a known correct answer. BixBench,[69] for example, assesses LLM-based agents on bioinformatics-analysis tasks with real data, and BiomniEval1 extends this approach to 433 queries across 10 biomedical task types with direct comparisons to human expert performance. These benchmarks are the most systematic evaluation

Please cite this article in press as: Guo et al., Toward autonomous science with agentic artificial intelligence, Cell (2026), https://doi.org/10.1016/j.cell.2026.08.052

frameworks for biomedical agents to date, but they test whether an agent can solve a given problem correctly, not necessarily whether it can decide what problems to solve. The second category would be evaluation of hypothesis generation and experimental design, and this has not yet been thoroughly benchmarked. No existing framework systematically assesses whether an agent can generate a novel hypothesis with scientific promise.

A third frontier would be evaluating the full closed-loop discovery cycle. In closed-loop robotic laboratories, the unit of evaluation would be the complete autonomous cycle: hypothesis generation, experimental planning, robotic execution, assay quality control, data analysis, the decision to continue or stop, and subsequent hypothesis revision. Minimum reporting should include the number of autonomous cycles, the number of experiments attempted and successfully executed, the number of instrument or protocol failures, the number of human interventions, and the extent to which each next experiment depended on prior results. These metrics would make it possible to distinguish a genuinely autonomous discovery trajectory from a conventional wet-lab study assisted by AI at isolated points.

As an initial step toward greater comparability and transparency, we propose that published studies adopt a minimum reporting standard that includes the total number of hypotheses generated, the number prioritized for experimental testing, and the number ultimately validated. The standard should also require that the full hypothesis pool and the system's original ranking be disclosed prior to human selection, so that the contribution of autonomous prioritization can be distinguished from human curation. Consistent reporting will enable the scientific community to begin estimating the precision and yield of AI-assisted hypothesis generation.

Reproducibility

Co-Scientist[12] was published without publicly released source code, preventing independent investigators from reproducing its reported behavior, probing failure modes, or determining how much of its performance derives from the underlying foundation model versus the proposed multi-agent reasoning architecture. The authors provided pseudocode (plain-language description of an algorithm's logic, written in human-readable form that resembles code, instead of in programming language syntax) and system prompts as supplementary material, but stated that full source code release is precluded by the system's integration with proprietary infrastructure. Google DeepMind has since made the Co-Scientist system available to individual researchers through Hypothesis Generation, a Google-hosted tool accessed via a gated interest form. For Robin,[13] the orchestration code and underlying literature-search framework were released as open source; the hosted deployment of these agents is now operated by Edison Scientific, FutureHouse's commercial spinout. The released components can be run independently, and reconstructing the full pipeline as described in the paper depends on those hosted services. Biomni was released as open source, including its environment (Biomni-E1), agent architecture (Biomni-A1), and reinforcement learning (RL)-trained models (Biomni-R0).[14]

If agentic scientific systems are to be evaluated with the same rigor expected of other research tools, we believe that publications should be required to provide sufficient methodological detail (such as prompts, agent architectures, and model versions) to enable their central claims to be independently evaluated. In experimental biology, a publication is required to provide sufficient methodological detail (including reagents, protocols, and analysis methods such as code) to enable independent replication. Comparable standards need to be set for agentic AI. This requirement extends naturally to the orchestration layer that drives autonomous laboratories. When an OpenClaw-like operating system schedules and executes physical experiments, reproducibility depends on disclosing not only prompts and model versions but also the instrument calls, parameters, error recovery, and decision logs that produced the reported results. Open, inspectable orchestration frameworks are advantageous here precisely because they allow an independent group to replay the agent's execution trace against the same or equivalent hardware. The senescence study[54] is evidence that its closed-loop procedure can be inspected and repeated in this manner. We recognize that full code release may be impractical for systems deeply integrated with proprietary infrastructure and that an absolutist requirement could discourage industry labs from publishing in scientific journals altogether. The appropriate standard may therefore need to be tiered, requiring enough transparency for independent evaluation of central claims even when full reproducibility is not achievable. We believe that the reasoning process, not just the answer, is central to every scientific publication. If key scientific contributions of these systems lie in how they reason, rather than simply what they predict, then transparency should extend beyond the final outputs to the reasoning framework itself.

Bias amplification and the narrowing of scientific breadth

Agentic scientific systems are trained on existing biomedical literature and will inherit its strengths, as well as its limitations. They may be more likely to generate hypotheses centered on well-studied genes, pathways, and diseases than on poorly characterized biology with fewer published evidence. They may also inherit systematic imbalances within the scientific literature itself. For example, a known source of bias in biomedical research is publication bias, which is the tendency for positive results to be published more readily than negative or null results.[70,71] Models trained on this literature may inadvertently reinforce these patterns, giving disproportionate weight to well-trodden paths while overlooking underexplored directions.

If many research groups deploy similar agentic systems trained on similar overlapping literature and datasets, there is a risk that scientific exploration would become concentrated within similar regions of biological hypothesis space. Rather than independently generating diverse lines of inquiry, some worry that there is the risk that researchers may increasingly prioritize similar genes, pathways, and mechanisms, while reinforcing existing scientific attention rather than broadening it. Whether this will occur in practice remains to be seen. But some of the most transformative scientific advances have emerged from serendipitous or unexpected observations or unconventional hypotheses that lie outside of the prevailing paradigms.[72] Historically, the most impactful ideas were often initially

Please cite this article in press as: Guo et al., Toward autonomous science with agentic artificial intelligence, Cell (2026), https://doi.org/10.1016/j.cell.2026.08.052

regarded as implausible, precisely because they conflicted with existing scientific consensus.

Dual use

Agentic scientific systems capable of autonomously generating biological hypotheses, proposing mechanistic pathways, and designing experiments inevitably raise concerns about use and misuse. These same capabilities to accelerate beneficial biomedical discovery could theoretically be used to advance harmful biological research. It is, therefore, important for governance to evolve alongside technical capability.[73] Robin utilizes commercially available, "off-the-shelf" LLMs that have already undergone extensive safety alignment.[13] These models have been subjected to red-teaming (in which experts deliberately attempt to elicit unsafe or harmful responses), as well as RL from human feedback (RLHF), which is a training process that encourages helpful behavior while discouraging the generation of potentially dangerous content. By relying on these safety-aligned foundation models, rather than unrestricted custom language models, Robin inherits a layer of existing biosafety safeguards.[13] Agentic systems capable of protocol generation and automated analysis could lower barriers to misuse; Biomni's capability in cloning protocol generation[14] may increase the accessibility of customized genetic manipulation.

Traditional dual-use risks in biotechnology have been constrained by an expertise barrier: the tacit knowledge required to design and execute dangerous experiments was acquired through years of specialized training and embedded within institutional cultures of responsibility. Biological agentic AI systems may produce harmful or operationally relevant outputs in a much higher percentage of misuse-relevant biological prompts as compared to standalone versions of the same underlying models.[74] The agent's planner may break a harmful request into individually permissible-appearing subtasks that individually pass the base model's safety checks. Agentic LLMs can outperform the median human expert (with doctoral-level training) on tasks such as DNA fragment design and writing scripts to operate liquid-handling robots for DNA assembly.[75] In wet-lab validation, scripts generated by an OpenAI o4-mini-high agent successfully assembled DNA on a laboratory robot, confirmed by whole-plasmid sequencing.[75] These findings mark a shift from an earlier red-team study[76] that found no statistically significant uplift from 2023-era LLMs for biological attack planning. There is fear that the newer models may be changing this risk profile and that the architectural features of the agentic paradigms are systemically lowering the expertise barrier for misuse through task decomposition, tool integration, and iterative planning. As these models become increasingly powerful, questions of who can access them, with what scientific directions and oversight, will be as important as questions of model performance. We propose that the biomedical community adopt governance frameworks modeled on existing dual-use precedents in synthetic biology, which include institutional review for hypothesis-generation in high-consequence areas such as pathogen-related research, and audit trails for system-generated proposals in high-risk areas, similar to screening frameworks that DNA synthesis providers already apply to synthetic nucleic acid orders.[73]

Cognitive offloading and selective deskilling

As AI agents assume a greater share of literature synthesis, hypothesis generation, experimental design, and data interpretation, there is concern that this may influence the expertise of the scientists that utilize them. Cognitive offloading itself is not inherently problematic, since scientific progress has usually benefited from tools that reduce the cognitive and technical burden. The bigger concern is selective deskilling or "never-skilling"[77]: if scientists increasingly delegate the processes through which mechanistic intuition and judgment are developed, their own skills in these may erode. This risk is particularly relevant for trainees, whose struggles with literature, data, and failed experiments are often precisely how scientific intuition and judgment are forged. A future challenge in education will be determining which cognitive tasks can be safely offloaded and which should be deliberately effortful. After all, in scientific training as in other creative pursuits—*the struggle is the point.*[78]

OUTLOOK

We believe scientific agentic AI is developing to be a powerful extension of human scientific judgment, instead of as the replacement of human creativity. The Co-Scientist, Robin, and Biomni publications do not show that these systems possess scientific judgment in the fullest sense: the ability to weigh competing priorities, recognize limitations of available evidence, take responsibility for consequences of a line of inquiry, or to recognize when a statistically compelling result has not yet achieved scientific confidence.[12,13] What they do demonstrate is an ability to compress the cycle between research question and a testable hypothesis and execute complex analytical workflows between the experimental result and its provisional interpretation, in ways that potentially accelerate the iterative cycle on which scientific discovery has always depended. Current AI agents do not yet have the scientific intuition or intellectual courage to take the data into unexpected directions, including abandoning the original question entirely[79] or making abductive "jumps."[42] Systems built primarily to reason through the synthesis of existing knowledge may be better suited to interpolation within established conceptual frameworks than to recognize when those frameworks should be abandoned or reformulated. Many transformative advances in science have resulted from treating an anomalous result as evidence that the underlying model may be wrong to begin with. It remains unknown whether advances in agentic AI will enable systems to recognize when an experiment's most important implications lie outside the question it was originally designed to answer.

The most powerful scientific systems of the coming decades are unlikely to consist of humans or machines operating independently, but instead of collaborative ecosystems in which each contributes complementary strengths. AI offers generative capacity: the ability to explore a hypothesis space larger than any individual investigator could survey and compress the cycle time from question to experimentally testable idea. Human scientists will continue to supply the judgment on which questions matter, the ethical values that constrain how these questions can be pursued, and the heavy weight of accountability that none of these AI systems are designed to bear. Realizing this

complementary relationship will also require deliberate investment in training scientists to use, interrogate, and supervise these systems, so that human oversight remains substantive rather than nominal as the tools grow more capable.

In summary, this perspective traced the history from supervised classifiers[3,4,8,9] to AlphaFold's generative leap,[18,21] and now to three distinct agentic paradigms: multi-agent tournament systems (Co-Scientist), multi-agent systems with autonomous data analysis (Robin), and a general-purpose agent operating through an integrated code-execution environment (Biomni). This history reflects an outward expansion of the role of AI from recognizing patterns, to predicting biological structure, to generating and evaluating hypotheses within broadly defined scientific problems. Each advance has built upon those that preceded it: today's hypothesis-generating agents draw upon a scientific ecosystem made possible by advances such as AlphaFold, and AlphaFold itself was enabled by decades of crystallographic data that earlier classifiers helped to process and annotate. The next stage, whatever form the autonomous laboratory ultimately takes,[52,53] will likewise build upon the agentic reasoning systems emerging today. That next stage is already coming to view. The pace at which such systems become routine may depend on the emergence of open, standardized interfaces between reasoning agents and laboratory automation, to enable interoperability, reproducibility, and rapid iteration across experimental platforms. Ensuring that these technologies advance science responsibly will require more than improvements in model capability. It will require thoughtful choices about evaluation, reproducibility, governance, and preservation of scientific diversity, to ensure that increasingly capable AI will broaden the creativity, rigor, and accessibility of biomedical research.

In the future, the wide adaptation of agentic AI may change which skills determine scientific success. As literature synthesis, coding, data analysis, and even hypothesis generation become increasingly automated, proficiency in these tasks may become less differentiating. Instead, advantage may shift toward scientists who can effectively direct and collaborate with AI systems: asking productive questions, interrogating outputs, and recognizing hallucinations and subtle errors. We believe that the ability to work effectively with AI may itself become a core scientific competency, and increasing automation can make deep human expertise even more valuable. As the generation of plausible hypotheses becomes increasingly accessible and inexpensive, the limiting resource will shift from producing ideas to judging and testing them. The human intuition to recognize the value of an unexpected observation will be even more important. The scientists who thrive in our agentic era will not be the ones who delegate the most tasks to AI but the ones who know best when to trust or distrust it.

ACKNOWLEDGMENTS

This work was supported by grants from the National Medical Research Council, Singapore (MOH-001689-00, MOH-000655-00, and MOH-001014-00); Duke-NUS Medical School (Duke-NUS/RSF/2021/0018, 05/FY2020/EX/15-A58, and 05/FY2022/EX/66-A128); the Agency for Science, Technology and Research, Singapore (A20H4g2141 and H20C6a0032); and Research to Prevent Blindness. L.Y.G. acknowledges support from the NIH Director's Early Independence Award (DP5 0D039444), Foundation Fighting Blindness, American Society of Gene and Cell Therapy, and NIH P30 EY001583. D.S.J.T. acknowledges support from the Medical Research Council Clinician Scientist Fellowship (UKRI2441) and Sight Research UK (SRUK) Translational Research Award (reference: TRN008_Ting_Birmingham).

DECLARATION OF INTERESTS

A.Z. is the CEO of Insilico Medicine, and A.A. is the president of Insilico Medicine.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

The authors used Claude Opus 5, GPT-5.6 Sol, and Biomni to proofread grammar and syntax, improve readability, and generate a list of technical terminology that the authors then defined in the Glossary. Any AI-assisted edits were reviewed by the authors, who take full responsibility for the content of this publication.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2026.08.052.

REFERENCES

1. Teo, Z.L., Thirunavukarasu, A.J., Elangovan, K., Cheng, H., Moova, P., Soetikno, B., Nielsen, C., Pollreisz, A., Ting, D.S.J., Morris, R.J.T., et al. (2025). Generative artificial intelligence in medicine. Nat. Med. *31*, 3270–3282. https://doi.org/10.1038/s41591-025-03983-2.
2. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., and Ting, D.S.W. (2023). Large language models in medicine. Nat. Med. *29*, 1930–1940. https://doi.org/10.1038/s41591-023-02448-8.
3. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA *316*, 2402–2410. https://doi.org/10.1001/jama.2016.17216.
4. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature *542*, 115–118. https://doi.org/10.1038/nature21056.
5. Keskin, O., Tuncbag, N., and Gursoy, A. (2016). Predicting Protein-Protein Interactions from the Molecular to the Proteome Level. Chem. Rev. *116*, 4884–4909. https://doi.org/10.1021/acs.chemrev.5b00683.
6. Sun, T., Zhou, B., Lai, L., and Pei, J. (2017). Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinf. *18*, 277. https://doi.org/10.1186/s12859-017-1700-2.
7. Wang, L., Wang, H.-F., Liu, S.-R., Yan, X., and Song, K.-J. (2019). Predicting Protein-Protein Interactions from Matrix-Based Protein Sequence Using Convolution Neural Network and Feature-Selective Rotation Forest. Sci. Rep. *9*, 9848. https://doi.org/10.1038/s41598-019-46369-4.
8. Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P.T., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol. *36*, 983–987. https://doi.org/10.1038/nbt.4235.
9. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315. https://doi.org/10.1038/ng.2892.
10. Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J.K., Brock, K., Gal, Y., and Marks, D.S. (2021). Disease variant prediction with deep generative models of evolutionary data. Nature *599*, 91–95. https://doi.org/10.1038/s41586-021-04043-8.

11. Gao, H., Hamp, T., Ede, J., Schraiber, J.G., McRae, J., Singer-Berk, M., Yang, Y., Dietrich, A.S.D., Fiziev, P.P., Kuderna, L.F.K., et al. (2023). The landscape of tolerated genetic variation in humans and primates. Science *380*, eabn8153. https://doi.org/10.1126/science.abn8197.

12. Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Sirkovic, P., Myaskovsky, A., Glowaty, G., Weissenberger, F., Orlandi, A., Popovici, D., et al. (2026). Accelerating scientific discovery with Co-Scientist. Nature *655*, 487–496. https://doi.org/10.1038/s41586-026-10644-y.

13. Ghareeb, A.E., Chang, B., Mitchener, L., Yiu, A., Szostkiewicz, C.J., Shved, D., Gyimesi, G.J., Laurent, J.M., Wright, S.M., Razzak, M.T., et al. (2026). A multi-agent system for automating scientific discovery. Nature *655*, 497–505. https://doi.org/10.1038/s41586-026-10652-y.

14. Huang, K., Zhang, S., Wang, H., Qu, Y., Lu, Y., Li, R., Roohani, Y., Qiu, L., Cao, S., Li, G., et al. (2026). Autonomous biomedical research with an artificial intelligence agent. Science *393*, eadz4351. https://doi.org/10.1126/science.adz4351.

15. Teng, C.W., Patel, S.D., Barkmeier, A.J., Liu, T.Y.A., Myung, D., Henderer, J., Liu, J., Hansen, E., and Al-Aswad, L.A. (2026). Autonomous Artificial Intelligence in Diabetic Retinopathy Testing—Lessons Learned on Successful Health System Adoption. Ophthalmol. Sci. *6*, 100935. https://doi.org/10.1016/j.xops.2025.100935.

16. Tropsha, A., Isayev, O., Varnek, A., Schneider, G., and Cherkasov, A. (2024). Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. Nat. Rev. Drug Discov. *23*, 141–155. https://doi.org/10.1038/s41573-023-00832-0.

17. Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. Nature *577*, 706–710. https://doi.org/10.1038/s41586-019-1923-7.

18. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

19. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science *373*, 871–876. https://doi.org/10.1126/science.abj8754.

20. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science *379*, 1123–1130. https://doi.org/10.1126/science.ade2574.

21. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. *50*, D439–D444. https://doi.org/10.1093/nar/gkab1061.

22. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature *630*, 493–500. https://doi.org/10.1038/s41586-024-07487-w.

23. Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. (2024). scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat. Methods *21*, 1470–1480. https://doi.org/10.1038/s41592-024-02201-0.

24. Theodoris, C.V., Xiao, L., Chopra, A., Chaffin, M.D., Al Sayed, Z.R., Hill, M.C., Mantineo, H., Brydon, E.M., Zeng, Z., Liu, X.S., and Ellinor, P.T. (2023). Transfer learning enables predictions in network biology. Nature *618*, 616–624. https://doi.org/10.1038/s41586-023-06139-9.

25. Watson, J.L., Juergens, D., Bennett, N.R., Trippe, B.L., Yim, J., Eisenach, H.E., Ahern, W., Borst, A.J., Ragotte, R.J., Milles, L.F., et al. (2023). De novo design of protein structure and function with RFdiffusion. Nature *620*, 1089–1100. https://doi.org/10.1038/s41586-023-06415-8.

26. Bianchi, J., Hirt, J., Vogt, M., and Vetsch, J. (2025). Data Extractions Using a Large Language Model (Elicit) and Human Reviewers in Randomized Controlled Trials: A Systematic Comparison. Cochrane Evid. Synth. Methods *3*, e70033. https://doi.org/10.1002/cesm.70033.

27. Wierenga, R.P., Golas, S.M., Ho, W., Coley, C.W., and Esvelt, K.M. (2023). PyLabRobot: An open-source, hardware-agnostic interface for liquid-handling robots and accessories. Device *1*, 100111. https://doi.org/10.1016/j.device.2023.100111.

28. Si, C., Yang, D., and Hashimoto, T. (2024). Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. Preprint at arXiv. https://doi.org/10.48550/arXiv.2409.04109.

29. Si, C., Hashimoto, T., and Yang, D. (2025). The Ideation-Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas. Preprint at arXiv. https://doi.org/10.48550/arXiv.2506.20803.

30. Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., Zhou, W., Huang, J., and Tang, Y. (2012). Prediction of Drug-Target Interactions and Drug Repositioning via Network-Based Inference. PLoS Comput. Biol. *8*, e1002503. https://doi.org/10.1371/journal.pcbi.1002503.

31. Guney, E., Menche, J., Vidal, M., and Barábasi, A.-L. (2016). Network-based in silico drug efficacy screening. Nat. Commun. *7*, 10331. https://doi.org/10.1038/ncomms10331.

32. Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K.N., et al. (2006). The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. Science *313*, 1929–1935. https://doi.org/10.1126/science.1132939.

33. Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K., et al. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell *171*, 1437–1452.e17. https://doi.org/10.1016/j.cell.2017.10.049.

34. Wu, Z., Li, W., Liu, G., and Tang, Y. (2018). Network-Based Methods for Prediction of Drug-Target Interactions. Front. Pharmacol. *9*, 1134. https://doi.org/10.3389/fphar.2018.01134.

35. Waring, M.J., Arrowsmith, J., Leach, A.R., Leeson, P.D., Mandrell, S., Owen, R.M., Pairaudeau, G., Pennie, W.D., Pickett, S.D., Wang, J., et al. (2015). An analysis of the attrition of drug candidates from four major pharmaceutical companies. Nat. Rev. Drug Discov. *14*, 475–486. https://doi.org/10.1038/nrd4609.

36. Mao, Y., and Finnemann, S.C. (2021). Acute RhoA/Rho Kinase Inhibition Is Sufficient to Restore Phagocytic Capacity to Retinal Pigment Epithelium Lacking the Engulfment Receptor MerTK. Cells *10*, 1927. https://doi.org/10.3390/cells10081927.

37. Minami, Y., Song, Y.S., Ishibazawa, A., Omae, T., Ro-mase, T., Ishiko, S., and Yoshida, A. (2019). Effect of ripasudil on diabetic macular edema. Sci. Rep. *9*, 3703. https://doi.org/10.1038/s41598-019-40194-5.

38. Yamagishi-Kimura, R., Honjo, M., and Aihara, M. (2025). Neuroprotective effect of ripasudil on retinal ganglion cells via an antioxidative mechanism. Jpn. J. Ophthalmol. *69*, 823–832. https://doi.org/10.1007/s10384-025-01243-x.

39. Baker, D.J., Frommer, L.M., Uslu, U., Patel, K.K., Zhu, D., Engel, N.W., George, J.M., Zhao, W., Kim, S.I., Sun, L., et al. (2026). AI-driven discovery of GPNMB CAR T cells as a multi-cancer therapy. Cell *189*, 3871–3882.e12. https://doi.org/10.1016/j.cell.2026.06.002.

40. Zemp, F.J., Breckenridge, Z., Song, H., Gill, G.S., Louie, T.L., Narta, K., Liu, H., Suh, Y., Guignard, L., Mandujano-Tinoco, E.A., et al. (2026). GPNMB-directed CAR T cell therapy against MiT/TFE-family fusion-driven solid tumors. Nat. Cancer *7*, 1189–1207. https://doi.org/10.1038/s43018-026-01194-3.

41. Savage, N., Grewal, S., Shaikh, M.V., Zemp, F.J., Mckenna, D., Mikolajewicz, N., Najem, H., Pyczek, J., Wei, J., Taleb, M.A.B., et al. (2026). Dual tumour–myeloid targeting of glioblastoma with GPNMB CAR-T cells. Nature *656*, 1013–1022. https://doi.org/10.1038/s41586-026-10641-1.

42. Zahavy, T. (2026). Position: LLMs can't jump. In Proceedings of the 43rd International Conference on Machine Learning, *306* (PMLR).

43. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling Laws for Neural Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2001.08361.

44. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de L., Hendricks, L.A., Welbl, J., Clark, A., et al. (2022). Training Compute-Optimal Large Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2203.15556.

45. Chen, Z., Wang, S., Xiao, T., Wang, Y., Chen, S., Cai, X., He, J., and Wang, J. (2025). Revisiting Scaling Laws for Language Models: The Role of Data Quality and Training Strategies. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, W. Che, J. Nabende, E. Shutova, and M.T. Pilehvar, eds. (Association for Computational Linguistics), pp. 23881–23899. https://doi.org/10.18653/v1/2025.acl-long.1163.

46. Snell, C., Lee, J., Xu, K., and Kumar, A. (2024). Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. Preprint at arXiv. https://doi.org/10.48550/arXiv.2408.03314.

47. Karten, S., Zhang, J., Upaa, T., Feng, R., Li, W., Shi, C., Jin, C., and Vodrahalli, K. (2026). Continual Harness: Online Adaptation for Self-Improving Foundation Agents. Preprint at arXiv. https://doi.org/10.48550/arXiv.2605.09998.

48. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of Go with deep neural networks and tree search. Nature *529*, 484–489. https://doi.org/10.1038/nature16961.

49. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., and Xie, W. (2024). Development of a large-scale medical visual question-answering dataset. Commun. Med. *4*, 277. https://doi.org/10.1038/s43856-024-00709-2.

50. Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., Liu, Z., Chen, X., Davison, B.D., Ren, H., et al. (2024). A generalist vision–language foundation model for diverse biomedical tasks. Nat. Med. *30*, 3129–3141. https://doi.org/10.1038/s41591-024-03185-2.

51. Liechti, R., George, N., Götz, L., El-Gebali, S., Chasapi, A., Crespo, I., Xenarios, I., and Lemberger, T. (2017). SourceData: a semantic platform for curating and searching figures. Nat. Methods *14*, 1021–1022. https://doi.org/10.1038/nmeth.4471.

52. Abolhasani, M., and Kumacheva, E. (2023). The rise of self-driving labs in chemical and materials sciences. Nat. Synth. *2*, 483–492. https://doi.org/10.1038/s44160-022-00231-0.

53. Tom, G., Schmid, S.P., Baird, S.G., Cao, Y., Darvish, K., Hao, H., Lo, S., Pablo-García, S., Rajaonson, E.M., Skreta, M., et al. (2024). Self-Driving Laboratories for Chemistry and Materials Science. Chem. Rev. *124*, 9633–9732. https://doi.org/10.1021/acs.chemrev.4c00055.

54. Tang, Q., Xiao, D., Veviorskiy, A., Xin, Y., Lok, S.W.Y., Pulous, F.E., Zhang, P., Zhu, Y., Ma, Y., Hu, X., et al. (2025). AI-Driven Robotics Laboratory Identifies Pharmacological TNIK Inhibition as a Potent Senomorphic Agent. Aging Dis. *17*, 432–451. https://doi.org/10.14336/AD.2024.1492.

55. Zhavoronkov, A., Galkin, F., Chen, S., Ren, F., Aliper, A., Durymanov, M., Sidorenko, D., Cui, H., Han, J.-DJ, Xu, H., et al. (2026). Integration of proteomic aging clocks in a phase 2a clinical trial supports simultaneous geroprotective assessment. Nat. Biotechnol. https://doi.org/10.1038/s41587-026-03286.

56. Fushimi, K., Nakai, Y., Nishi, A., Suzuki, R., Ikegami, M., Nimura, R., Tomono, T., Hidese, R., Yasueda, H., Tagawa, Y., and Hasunuma, T. (2025). Development of the autonomous lab system to support biotechnology research. Sci. Rep. *15*, 6648. https://doi.org/10.1038/s41598-025-89069-y.

57. King, R.D., Whelan, K.E., Jones, F.M., Reiser, P.G.K., Bryant, C.H., Muggleton, S.H., Kell, D.B., and Oliver, S.G. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. Nature *427*, 247–252. https://doi.org/10.1038/nature02236.

58. King, R.D., Rowland, J., Oliver, S.G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L.N., et al. (2009). The Automation of Science. Science *324*, 85–89. https://doi.org/10.1126/science.1165620.

59. Williams, K., Bilsland, E., Sparkes, A., Aubrey, W., Young, M., Soldatova, L.N., De Grave, K., Ramon, J., de Clare, M., Sirawaraporn, W., et al. (2015). Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. J. R. Soc. Interface *12*, 20141289. https://doi.org/10.1098/rsif.2014.1289.

60. Qiu, Y., Huang, Z., Wang, Z., Liu, H., Qiao, Y., Hu, Y., Sun, S., Peng, H., Xu, R.X., and Sun, M. (2025). BioMARS: A Multi-Agent Robotic System for Autonomous Biological Experiments. Preprint at arXiv. https://doi.org/10.48550/arXiv.2507.01485.

61. Wang, W., Swain, S., Lee, J., Lin, Z., Canales, B., Aljović, A., Liu, Y., Li, Q., Marin-Llobet, A., Liu, M., et al. (2025). Agentic Lab: An Agentic-physical AI system for cell and organoid experimentation and manufacturing. Preprint at bioRxiv. https://doi.org/10.1101/2025.11.11.686354.

62. Angers, K., Darvish, K., Yoshikawa, N., Okhovatian, S., Bannerman, D., Yakavets, I., Shkurti, F., Aspuru-Guzik, A., and Radisic, M. (2025). RoboCulture: A Robotics Platform for Automated Biological Experimentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2505.14941.

63. Dai, T., Vijayakrishnan, S., Szczypiński, F.T., Ayme, J.-F., Simaei, E., Fellowes, T., Clowes, R., Kotopanov, L., Shields, C.E., Zhou, Z., et al. (2024). Autonomous mobile robots for exploratory synthetic chemistry. Nature *635*, 890–897. https://doi.org/10.1038/s41586-024-08173-7.

64. Yu, Y., Mai, Y., Zheng, Y., and Shi, L. (2024). Assessing and mitigating batch effects in large-scale omics studies. Genome Biol. *25*, 254. https://doi.org/10.1186/s13059-024-03401-9.

65. Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., and Irizarry, R.A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. *11*, 733–739. https://doi.org/10.1038/nrg2825.

66. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Trans. Inf. Syst. *43*, 1–55. https://doi.org/10.1145/3703155.

67. Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature *630*, 625–630. https://doi.org/10.1038/s41586-024-07421-0.

68. Topaz, M., Roguin, N., Gupta, P., Zhang, Z., and Peltonen, L.-M. (2026). Fabricated citations: an audit across 2.5 million biomedical papers. Lancet *407*, 1779–1781. https://doi.org/10.1016/S0140-6736(26)00603-3.

69. Mitchener, L., Laurent, J.M., Andonian, A., Tenmann, B., Narayanan, S., Wellawatte, G.P., White, A., Sani, L., and Rodriques, S.G. (2025). BixBench: a Comprehensive Benchmark for LLM-based Agents in Computational Biology. Preprint at arXiv. https://doi.org/10.48550/arXiv.2503.00096.

70. Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. Scientometrics *90*, 891–904. https://doi.org/10.1007/s11192-011-0494-7.

71. Hopewell, S., Loudon, K., Clarke, M.J., Oxman, A.D., and Dickersin, K. (2009). Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst. Rev. *2009*, MR000006. https://doi.org/10.1002/14651858.MR000006.pub3.

72. Henderson, J.W. (1997). The Yellow Brick Road to Penicillin: A Story of Serendipity. Mayo Clin. Proc. *72*, 683–687. https://doi.org/10.4065/72.7.683.

73. Hoffmann, S.A., Diggans, J., Densmore, D., Dai, J., Knight, T., Leproust, E., Boeke, J.D., Wheeler, N., and Cai, Y. (2023). Safety by design: Biosafety and biosecurity in the age of synthetic genomics. iScience *26*, 106165. https://doi.org/10.1016/j.isci.2023.106165.

74. Provatas, K.A., Self, A., Mouratidis, I., and Georgakopoulos-Soares, I. (2026). BioVeil MATRIX: Uncovering and categorizing vulnerabilities of agentic biological AI scientists. Preprint at arXiv. https://doi.org/10.48550/arXiv.2605.00927.

75. Liu, A.B., Nedungadi, S., Cai, B., Kleinman, A., Bhasin, H., and Donoughe, S. (2026). ABC-Bench: An Agentic Bio-Capabilities Benchmark for Biosecurity. Preprint at arXiv. https://doi.org/10.48550/arXiv.2606.11150.

76. Mouton, C.A., Lucas, C., and Guest, E. (2024). RAND Corporation. The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study. RR-A2977-2. https://doi.org/10.7249/RRA2977-2.

77. Ke, Y., Jin, L., Ong, J.C.L., Thirunavukarasu, A.J., Car, J., Cheung, C.Y., Tham, Y.C., Ting, D.S.W., Ong, M.E.H., Compton, S., et al. (2026). AI-induced never-skilling in medical education. Nat. Med. *32*, 1997–2006. https://doi.org/10.1038/s41591-026-04438-y.

78. Wang, V. (2025). The Dartmouth. Reflection: The struggle is the point. http://www.thedartmouth.com/article/2025/10/wang-reflection-the-struggle-is-the-point.

79. Elemento, O. (2026). AI systems devise hypotheses and ways to test them. Nature *655*, 313–314. https://doi.org/10.1038/d41586-026-01873-2.

Cell, Volume *189*

Supplemental information

Toward autonomous science

with agentic artificial intelligence

Lucie Y. Guo, Darren S.J. Ting, Xinyi Su, Alex Aliper, Alex Zhavoronkov, and Daniel S.W. Ting

GLOSSARY

Agent: A software system that can perceive inputs, reason about them, and take actions to achieve a goal, often by calling external tools, querying databases, or executing code. Unlike a conventional model that produces a single output, an agent interacts with its environment iteratively.

Benchmark: A standardized set of tasks or datasets used to evaluate and compare the performance of AI systems against each other or against human experts.

Closed-loop system: A system in which the output of one stage feeds back as input to the next, enabling iterative self-correction. In the autonomous laboratory, AI-generated hypotheses lead to robotic experiments, whose results are analyzed and used to generate improved hypotheses, with minimal human intervention.

Code-as-action: An agent design in which each step of a plan is expressed as executable code, allowing the agent to interact with tools and databases programmatically and compose workflows dynamically without predefined templates.

Convolutional neural network (CNN): A neural network architecture designed to process grid-structured data such as images, extracting spatial features like edges and textures through learned filters.

Deep learning: A branch of machine learning that uses multi-layered neural networks to learn hierarchical patterns from data, enabling tasks such as image recognition and language processing.

Drug repurposing: The identification of new therapeutic uses for existing, already-approved drugs, leveraging established safety profiles to accelerate clinical translation.

Dual-use: Technologies or research that can be applied for both beneficial and harmful purposes. In biology, dual-use research of concern involves knowledge, tools, or techniques that could be misapplied to cause harm (e.g., engineering pathogens).

Elo-based ranking: A method for ranking competitors from pairwise comparisons, originally developed to rank chess players; applied here to rank competing hypotheses by their performance in head-to-head evaluations.

Forward pass: A single pass of input data through a neural network to produce an output, without iterative refinement or self-correction.

Foundation model: A large model trained on broad, diverse data (often via self-supervised learning) that can be adapted to many downstream tasks with minimal task-specific training.

Generative AI: AI systems that produce new content (such as text, images, molecular structures, protein designs) instead of only classifying or labeling existing data.

Ground truth: A known correct answer against which a model's predictions can be compared. In diagnostic AI, ground truth is established by expert labels; in scientific discovery, it must be established through the very experiments the hypothesis motivates.

Hallucination: The generation of plausible-sounding but factually incorrect or fabricated content by an AI model — including nonexistent citations, false factual claims, or erroneous data — presented with the appearance of authority.

Inference: The process of using a trained model to generate outputs for new, unseen inputs.

Large language model (LLM): A type of foundation model trained on large quantities of text to predict and generate human language; underlies systems such as GPT, Claude, and Gemini.

Monte Carlo Tree Search: A search algorithm that evaluates possible actions by building a decision tree through many random simulations, used in game-playing AI such as AlphaGo to search through vast possibility spaces.

Multi-agent system: An AI architecture in which multiple specialized agents, each assigned distinct functional roles (e.g., generating, critiquing, ranking, or analyzing), collaborate or compete to solve a problem.

Multi-modal AI: AI systems that integrate and reason across multiple data types simultaneously (e.g., text, images, genomic sequences, structured databases).

Open source: Software whose source code is made publicly available for anyone to inspect, modify, and redistribute, enabling independent verification and community contribution.

Orchestration layer: A software layer that coordinates communication and execution between reasoning agents, laboratory instruments, data systems, and analysis pipelines, enabling end-to-end automation of experimental workflows.

Parameters: The internal numerical values within a model that store what it has learned during training; larger models have more parameters and typically greater capacity.

Protein language model: A model trained on large collections of protein sequences (analogous to how LLMs are trained on text) that learns to represent and predict structural and functional properties of proteins.

Publication bias: The tendency for studies with positive or statistically significant results to be published more readily than those with negative or null findings, distorting the published evidence base.

QSAR models: Computational models that predict the biological activity or properties of a molecule from its chemical structure, widely used in drug discovery to screen candidate compounds.

Red-teaming: A safety evaluation practice in which experts deliberately attempt to elicit unsafe, harmful, or unintended responses from an AI system to identify and address vulnerabilities before deployment.

Reinforcement learning (RL): A machine learning paradigm in which an agent learns by trial and error, receiving rewards or penalties for its actions, to maximize cumulative reward over time.

Retrieval-augmented planning: A planning strategy in which an agent recalls past experiences or references external data to inform its decisions, rather than relying solely on knowledge encoded during pre-training.

Scaling laws: The empirical observation that AI system performance improves in predictable ways as models are trained on more data, with more computing power, and with more parameters.

Scientific reasoning: The iterative cognitive process of formulating hypotheses, designing experiments to test them, interpreting evidence, and refining understanding. This is the process that agentic AI systems now begin to approximate, rather than merely performing fixed prediction tasks

Senescence: A state in which cells permanently stop dividing but remain metabolically active, contributing to aging and age-related disease. Senomorphic interventions modify the harmful secretions of senescent cells without killing them.

Supervised learning: A machine learning paradigm in which a model is trained on labeled examples (known input-output pairs) to learn a mapping from inputs to desired outputs.

Test-time compute: The computational effort a model expends while solving a specific problem after training, such as exploring alternatives, evaluating evidence, and refining potential solutions before committing to an answer.

Token: The basic unit of text that a language model processes, such as a word, part of a word, or a punctuation mark.

Tournament-style self-critique: An architecture in which multiple AI agents generate candidate hypotheses and then compete in pairwise contests, with a judge agent ranking them, to iteratively refine and select the most promising proposals.

Training-time compute: The computational resources used to build and train a model before it is deployed.

Transformer: A neural network architecture that processes all elements of an input simultaneously (rather than sequentially), enabling efficient modeling of long-range dependencies in sequences.

Variant caller: A computational tool that identifies genetic variants (e.g., differences between a sample's genome and a reference genome) from DNA sequencing data.