

How Far Are AI Scientists from Changing the World?

A Survey of Autonomous Scientific Discovery

Qiujie Xie^{1,2,*}, Yixuan Weng^{2,*}, Minjun Zhu^{1,2,*},
Junying Fang^{3,2}, Fuchen Shen², Shulin Huang², Zhen Lin²,
Zhouyang Wang², Shichen Li², Jiahui Zhou^{4,2}, Zilan Mao²,
Sifan Liu², Zhenhao Liu², Dalv Yin², Fanjunduo Wei²,
Zhiyuan Ning², Zijie Yang⁵, Linyi Yang^{6,2}, Jian Wu², Yue Zhang^{2,†}

¹Zhejiang University ²School of Engineering, Westlake University

³Guangzhou Institute of Tropical and Marine Meteorology,
China Meteorological Administration

⁴School of Software, Dalian University of Technology

⁵School of Life Sciences, Westlake University ⁶University College London

*Equal contribution †Corresponding author

Abstract

This survey asks how far AI Scientist systems are from changing the world, and argues that their impact should be evaluated not by whether they produce plausible papers, but by whether they produce auditable research trajectories that can be inspected, reproduced, and challenged. We first provide a definition of AI Scientists, distinguish them from adjacent scientific AI systems, and organize the field around four core capabilities: knowledge acquisition, hypothesis generation, experimental validation, and evolution. We further synthesize representative systems by autonomy, execution grounding, auditability, human control, and evidence level. The resulting picture shows a field moving from prompt-to-paper automation toward auditable research laboratories, where executable grounding, provenance, and human-governed autonomy become central. We conclude by identifying open challenges toward trustworthy AI-driven science and providing a roadmap for integrating AI Scientists into the future landscape of scientific discovery.

1 Introduction

Scientific discovery has long advanced in step with the development of tools (Figure 1). While traditional instruments have expanded human perception into both macroscopic and microscopic scales (Krauss, 2025), the process of scientific discovery has remained largely human-driven (Gauch, 2003). Over time, this workflow has become increasingly strained by **finite human cognitive capacity** (Marois and Ivanoff, 2005) and the **overwhelming growth of scientific literature** (Szalay and Gray, 2006; Baker, 2016).

With the advent of computing technology, researchers have pursued the automation of scientific discovery since the 1950s. Early systems such as Logic Theorist (Newell and Simon, 1956) showed that rule-based reasoning and heuristic search could generate mathematical proofs. In subsequent decades, **expert systems** often translated domain knowledge into logical rules, enabling automated reasoning in well-defined scientific settings, such as chemical structure elucidation and analytical decision-making (Lindsay et al., 1993; McDermott, 1982).

As computational resources increased and digital data became more abundant, **data-driven approaches** began to complement the earlier rule-based systems (Langley, 1977). These approaches demonstrated the potential of automatic discovery within curated inputs and simplified assumptions (Langley, 1987). The emergence of **deep learning** in the early 2010s has significantly expanded the scope of automatic scientific discovery (King et al., 2009; LeCun et al., 2015). Foundation models, trained on large-scale scientific data, have demonstrated strong capabilities in both understanding and generating scientific content (Beltagy et al., 2019; Bi et al., 2024).

In parallel, the integration of these **foundation models** with **agentic frameworks** (Yao et al., 2023) and **automated experimental infrastructures** (Jumper et al., 2021; Mieszczynek et al., 2024; Abolhasani and Kumacheva, 2023) has enabled closed-loop systems capable of iteratively proposing, testing, and refining scientific hypotheses (Angello et al., 2024). This progress is propelling automated scientific discovery to a new level, giving rise to the concept of the AI Scien-

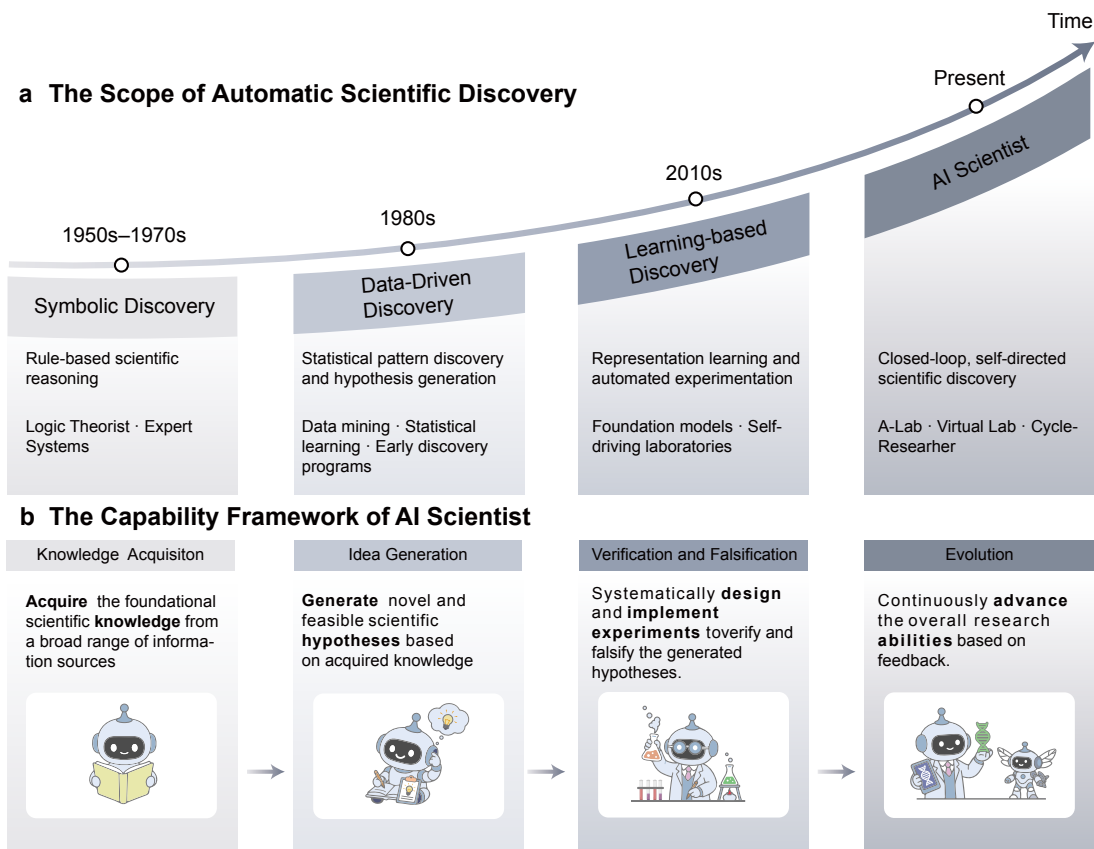


Figure 1: Timeline of automated scientific discovery. Automated scientific discovery has progressed from rule-based reasoning to closed-loop AI Scientist systems that integrate knowledge acquisition, hypothesis generation, validation, and evolution.

tist (Lu et al., 2024).

An **AI Scientist** is an autonomous, goal-directed system that coordinates multiple stages of the scientific discovery process (Kitano, 2021; Lu et al., 2024). In traditional AI breakthroughs, humans define the scientific objective and success criteria, while machine learning models serve as specialized tools within that workflow (Wang et al., 2023). An illustrative example is AlphaFold (Jumper et al., 2021), which solves the task of protein structure prediction with near-experimental accuracy. By contrast, an AI Scientist aims to coordinate research trajectories: it can identify knowledge gaps, formulate hypotheses, design experiments, interpret results, record failures, and iterate on its findings (Lu et al., 2024). This architecture can **accelerate hypothesis–test–revise cycles** and support systematic exploration of high-dimensional design spaces beyond human intuition, thus reshaping how scientific questions are explored (Romera-Paredes et al., 2024; Weng et al., 2026).

Over the past two years, the development of AI Scientist systems has accelerated across multiple disciplines (Szymanski et al., 2023; Oh et al.,

2025; Swanson et al., 2025; Yamada et al., 2025; Yang et al., 2023). Stanford University’s Virtual Lab (Swanson et al., 2025) organizes multiple large language model agents into a team structure and enables automated nanobody design, producing candidates with improved binding to recent SARS-CoV-2 variants. The AI Scientist-v2 (Yamada et al., 2025) demonstrates the capacity to automate a machine-learning research lifecycle from ideation to manuscript generation. In a controlled experiment coordinated with an ICLR 2025 workshop, one AI-generated manuscript achieved review scores comparable to accepted submissions (Yamada et al., 2025).

However, the rise of autonomous research agents is not without peril. The inherent limitations of foundation models may cause AI Scientists to hallucinate facts with high fluency or introduce systematic biases, thereby undermining the reliability and scientific value of their outputs (Gallegos et al., 2024; Zhu et al., 2025). Moreover, unlike human researchers, AI Scientists lack moral constraints (Tang et al., 2025b). They are incapable of making ethical judgments about the societal impli-

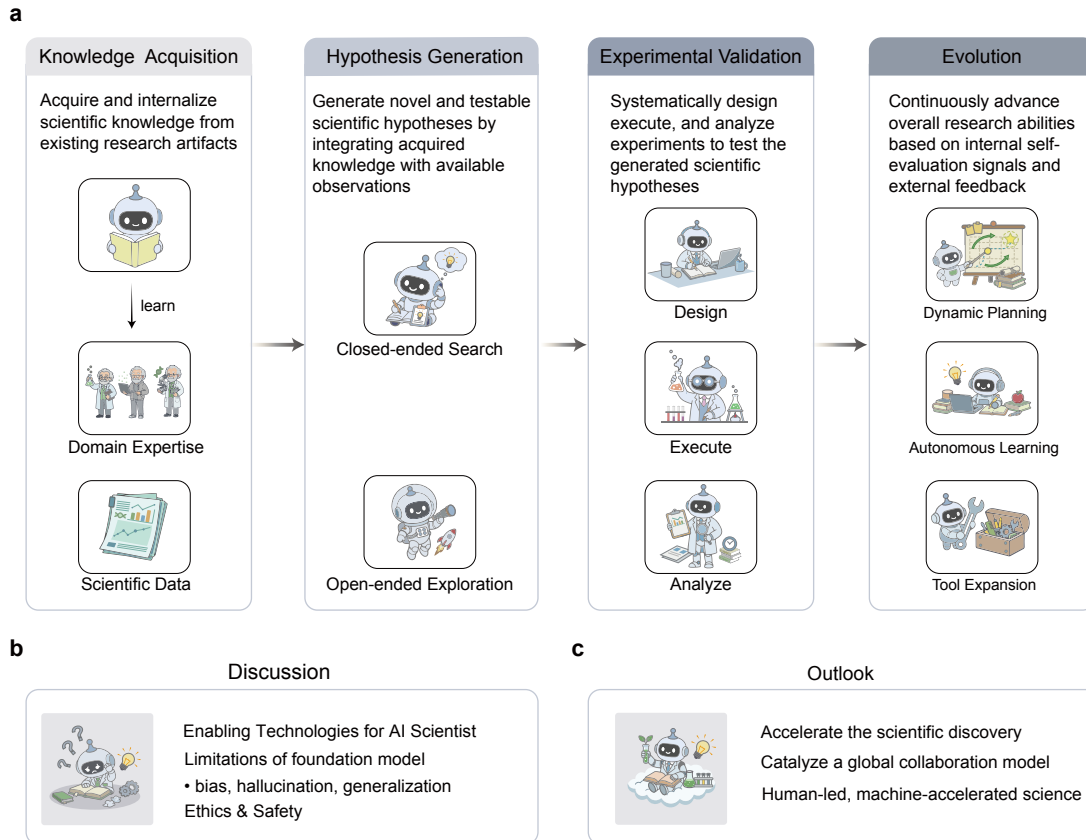


Figure 2: Overview of AI Scientist systems. This survey frames AI Scientists as closed-loop research systems that connect four core capabilities with enabling technologies, risks, and future directions.

cations of their work, raising concerns about the potential misuse of these systems (Bengio et al., 2025). As we stand on the threshold of this new era, we explore a central question: to what extent are AI Scientist systems capable of reshaping the scientific research paradigm, and how far are they from changing the world?

This survey makes four contributions (Figure 2): **First**, we provide a definition of AI Scientists and distinguish them from adjacent systems (Section 2). **Second**, we synthesize representative systems, clarifying what has been demonstrated (Sections 3–4). **Third**, we introduce a capability taxonomy across knowledge acquisition, hypothesis generation, experimental validation, and evolution, organizing relevant work into a unified framework for analyzing how AI Scientists coordinate scientific discovery (Sections 5–8). **Fourth**, we examine the technical and governance challenges that must be addressed as the field moves from prompt-to-paper automation toward auditable research laboratories (Sections 9–10). Table 5 summarizes how this survey differs from existing surveys.

2 Definition

An **AI Scientist** is an autonomous, goal-directed system that coordinates multiple substantive stages of the scientific discovery process, including knowledge acquisition, hypothesis generation, experimental design, execution, analysis, communication, and iterative refinement.

Conceptually, the AI Scientist represents a **shift in the scientific actor** (Figure 3): from a human scientist who orchestrates tools and instruments, to an AI system that assumes greater responsibility for planning, execution, and iteration under human supervision. They are systems that connect hypotheses, executable actions, failures, and revisions into a research process that can be inspected and governed (Table 1).

Specifically, an AI Scientist first acquires scientific knowledge from existing research artifacts to establish a working representation of a scientific domain. By integrating accumulated knowledge with new observations, it then generates novel and testable hypotheses, which are validated through scientific tools (Ding et al., 2025), automated experimental platforms (Roch et al., 2018), robotic lab-

Table 1: What counts as an AI Scientist?

System type	Status	Boundary test
Scientific tool	No	Solves one task under external orchestration.
Paper-writing agent	Usually no	Produces manuscripts without grounded validation.
AutoML pipeline	Adjacent	Optimizes predefined objectives, not research questions.
Self-driving lab	Partial	Executes experiments; needs reasoning and interpretation.
AI Scientist	Yes	Coordinates a hypothesis–test–revise trajectory.

oratories (Takahashi et al., 2025), code execution, or simulation. Finally, an evolution mechanism closes the loop by using self-evaluation signals, external feedback, and recorded failures to refine subsequent research decisions. While each capability is a mature research area in its own right, this review focuses on how these capabilities are integrated to endow AI agents with scientific autonomy and auditable research trajectories.

3 Survey Scope and Methodology

This survey examines AI-driven systems that automate or coordinate substantive stages of scientific research. We ask four questions: what distinguishes AI Scientists from adjacent systems; which stages of the scientific process they automate; what evidence supports claims of autonomy and discovery; and what technical, governance, and safety challenges remain.

We searched ACL Anthology, arXiv, Semantic Scholar, OpenReview, and project repositories using terms such as “AI Scientist”, “autonomous scientific discovery”, “scientific agents”, “automated research”, “self-driving laboratory”, and “hypothesis generation”. We followed citation links from representative systems and included recent preprints when they introduced a concrete system. We excluded systems limited to generic summarization, polishing, citation formatting, domain prediction, or hyperparameter optimization unless they were embedded in a broader research loop.

Each included system was coded by scientific stage, autonomy, human role, execution grounding, verification, auditability, output artifact, and evidence level. To avoid conflating visions with

demonstrated capabilities, we classify evidence from E0 conceptual proposals to E4 externally validated impact (Table 2). These levels describe what kind of support exists for a system’s claims. Many current systems operate at E2–E3, while E4 impact remains rare and domain-dependent.

4 System Map and Evidence Synthesis

Systems differ not only in capability, but also in what their demonstrations prove: some generate artifacts, some execute code or experiments, and a smaller set exposes a linked trajectory of decisions, actions, results, and revisions. Table 3 therefore assigns one evidence level to each entry using the cumulative rules in Table 2.

5 Knowledge Acquisition

The capability of knowledge acquisition lets an AI Scientist build and update a working representation of a domain before proposing research actions. It supplies the background state for the rest of the loop: what is known, what remains uncertain, which tools are available, and which observations should constrain later hypotheses.

Current systems draw on three main sources. **Structured expert knowledge** encodes rules, constraints, protocols, or toolchains, continuing the tradition of symbolic discovery systems such as Logic Theorist and DENDRAL (Newell and Simon, 1956; Lindsay et al., 1993). **Scientific data** and foundation models extend this base by learning from large-scale literature, images, sequences, and domain datasets (Beltagy et al., 2019; Taylor et al., 2022). Retrieval and tool use then allow systems such as Virtual Lab, Coscientist, and Dolphin to access **literature, online resources, and domain knowledge** during hypothesis generation and experimental planning (Swanson et al., 2025; Boiko et al., 2023; Yuan et al., 2025).

This capability involves more than reading papers into a context window. AI Scientist systems must turn heterogeneous evidence into a research context that can support later reasoning. Claims must be related to methods, datasets and open questions, and this context must remain revisable as new observations arrive. For instance, AI Scientist systems based on foundation models provide broad scientific priors from literature and domain data (Singhal et al., 2023; He et al., 2025b; Madani et al., 2023), while retrieval-augmented and agentic designs compensate for static model knowl-

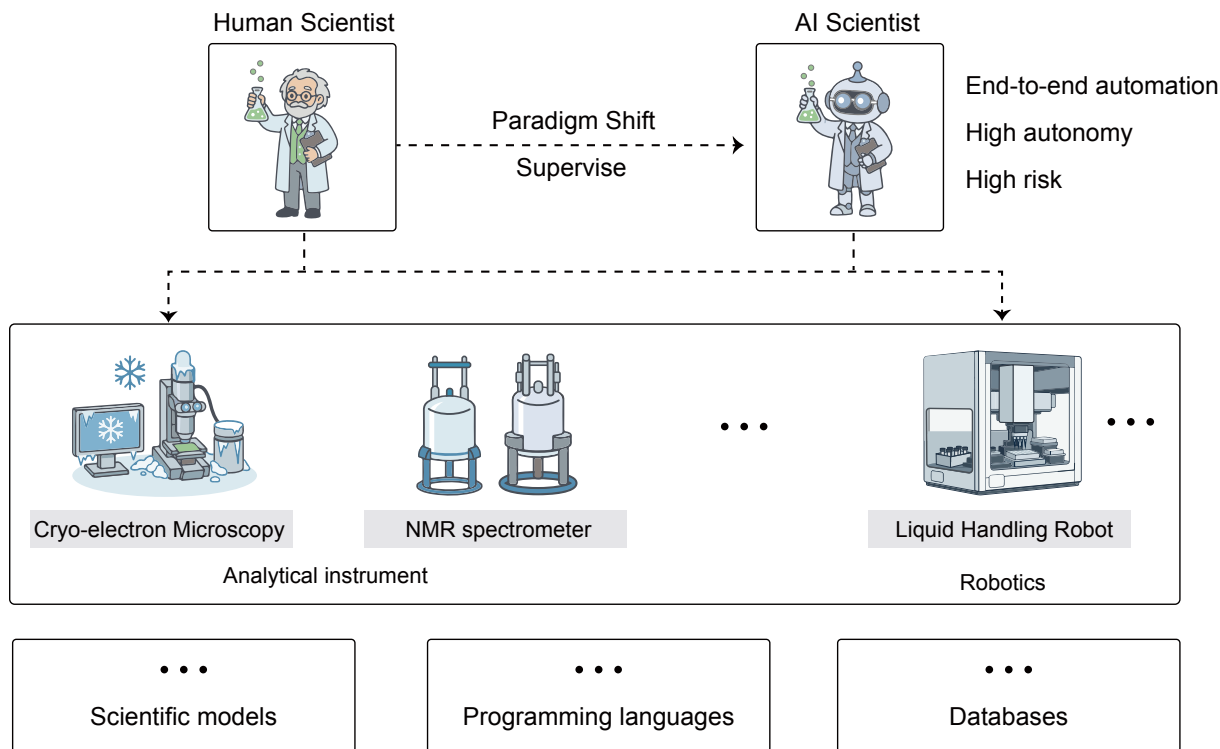


Figure 3: From human-in-the-loop science to AI Scientist-driven research. AI Scientist-driven research shifts human involvement from direct execution toward governance over autonomous systems that coordinate instruments, models and data.

Table 2: Evidence levels for AI Scientist claims.

Level	Name	Criterion
E0	Concept	Vision or architecture without a runnable system.
E1	Text demo	Ideas, reviews, reports, or papers without grounded execution.
E2	Executable prototype	Runs code, tools, simulations, or experiments with measured feedback.
E3	Auditable loop	Satisfies the E2 criterion and provides public records that connect the key research trajectory into an inspectable evidence chain.
E4	Independently validated result	Satisfies the E3 criterion and provides validation of the scientific result by a team independent of both the system developers and the authors of the original study.

edge by consulting external evidence during the research process (Lewis et al., 2020; Liu et al., 2024; Gupta et al., 2024). In systems such as Dolphin and Coscientist, retrieved papers, web resources, and databases become part of the scientific context for later ideation and planning, rather than passive background text (Yuan et al., 2025; Boiko et al., 2023).

Conceptually, knowledge acquisition marks the shift from accessing scientific text to forming a **usable research context**. An AI Scientist therefore differs from a literature tool when acquired knowledge is repeatedly reorganized and brought to bear on later research decisions.

6 Hypothesis Generation

Hypothesis generation is the capability to propose novel and testable research directions from acquired knowledge and observations. It connects the domain state built during knowledge acquisition to concrete questions that can be evaluated, refined, or rejected.

Earlier discovery systems searched **closed, hand-specified** spaces (Swanson, 1986; Lindsay et al., 1993), whereas foundation-model-based systems support **open-ended hypothesis generation** and iterative refinement using retrieved literature, tool outputs, and experimental feedback (Si et al., 2025; Lu et al., 2024; Boiko et al., 2023). Representative systems implement this process through two complementary mechanisms:

In **generative exploration**, systems such as The

Table 3: Representative AI Scientist systems. We compare demonstrated workflow, execution and records, human involvement, reported output units, and the highest evidence level fully supported by public evidence. NR means that no overall output count is reported. Detailed rationales appear in Appendix Tables 4.

System	Workflow	Execution evidence	Human role	Reported scale	Level
The AI Scientist (2024)	idea → experiment → paper	code, logs, generated artifacts	topic; code template in focused mode	NR	E2
AI Scientist-v2 (2025)	tree search → experiment → paper	code, search nodes, results	topic and experimental environment	3 workshop-evaluated papers	E2
DeepScientist (2026)	goal → hypothesis → iterative validation	code, Findings Memory, iteration records	research goal and operating environment	~5,000 ideas; ~1,100 validations	E3
Virtual Lab (2025)	question → design → screen → wet lab	discussions, code, scores, wet-lab data	feedback and wet-lab validation	92 designs; 2 improved	E3
A-Lab (2023)	plan → robotic synthesis → optimization	instrument results and failure analyses	targets, constraints, same-team reanalysis	57 targets; 353 recipes; 36 syntheses	E3
Coscientist (2023)	retrieval → planning → robotic experiment	retrieval, code, robotics records	task specification and supervision	6 chemistry task classes	E2
SciToolAgent (2025)	decompose → plan → tool use	tool calls and returned results	problem specification	4 case-study families; hundreds of tools	E2
ToolUniverse (2025)	execution infrastructure for upstream agents	API and tool-call records	upstream system supplies decisions	600+ tools	E2
autoresearch (2026)	edit → timed run → retain/discard	code, continuous logs, metrics	initial code, instructions, metric	NR	E2
AutoResearch Claw (2026)	topic → experiment → analysis → report	code, failures, citations, cross-run memory	seven intervention modes	25 ARC-Bench topics	E3
Claw AI Lab (2026)	team formation → experiment → paper	code, data, checkpoints, rollback	role configuration and monitoring	5 internal cases	E3

AI Scientist and Coscientist combine language models with literature retrieval, code execution, and tool use to propose candidate ideas and revise them through novelty and feasibility checks (Lu et al., 2024; Boiko et al., 2023).

In **structured selection**, multi-agent or reviewer-style workflows assign distinct roles to idea generators, critics, and domain-specific evaluators, allowing large candidate pools to be filtered before costly validation. For example, PrimeGen decomposes primer design into constraint interpretation, candidate generation, and thermodynamic and off-target evaluation (Wang et al., 2026), while CycleResearcher uses automated review feedback to iteratively improve research proposals (Weng et al., 2025). Together, these mechanisms frame hypothesis generation as a search process guided by memory, constraints, and evaluation signals.

Candidate hypotheses are typically assessed for novelty, feasibility, clarity, and scientific value (Si et al., 2025; Guo et al., 2025); although automated reviewers and agentic critics can scale this assessment, human expert judgment remains important when evaluation depends on tacit domain knowledge or ambiguous evidence (Weng et al., 2025;

Awasthi et al., 2025).

Conceptually, hypothesis generation is the step where an AI Scientist moves from representing existing knowledge to opening a space of possible discoveries. It is the transformation of a knowledge state into research directions that can organize attention, resources, and subsequent tests.

7 Experimental Validation

Experimental validation is the capability to design, execute, and analyze experiments that test generated hypotheses. It turns an idea into evidence by translating abstract claims into executable artifacts, selecting appropriate tools or instruments, and feeding observations back into the hypothesis–test–revise loop (Ding et al., 2025; Yamada et al., 2025; Si et al., 2026).

Current systems validate the generated hypotheses through three linked stages: **experiment design, execution, and result analysis**:

First, a system **translates a hypothesis into testable artifacts**, such as measurable objectives, experimental conditions, code, or machine-executable protocols; for example, a materials hypothesis may be converted into synthesis param-

ters, measurement plans, and stability targets (Angello et al., 2024; Szymanski et al., 2023; Starace et al., 2025).

Second, **the system selects an appropriate environment for testing the hypothesis and executes the experiment.** Depending on the research domain, it may involve code and benchmarks, simulation or theory-coupled workflows, or robotic platforms and scientific instruments (Starace et al., 2025; Liang et al., 2025; Mandal et al., 2025). Execution may therefore be **physical**, as in A-Lab and autonomous microscopy (Szymanski et al., 2023; Mandal et al., 2025); **virtual**, as in Virtual Lab and AMASE’s experiment–theory loop (Swanson et al., 2025; Liang et al., 2025); or **substitutional**, with AI agents approximating human participants in preliminary social-science studies (Hutson and Mastin, 2023; Grossmann et al., 2023).

Third, the system **interprets the returned observations**, including negative results and failure records, as evidence for revising subsequent decisions. This feedback mechanism makes systems such as ChemCrow and A-Lab scientifically meaningful: tool outputs and instrument records serve not merely as final products, but as constraints on the next round of reasoning (Bran et al., 2024; Szymanski et al., 2023).

Conceptually, experimental validation is what separates an AI Scientist from a paper-writing or ideation system. It moves the system from a textual loop of plausible claims into an empirical loop where the world can revise, or falsify those claims.

8 Evolution

Evolution is the capability to improve research behavior through self-evaluation and external feedback (Weng et al., 2025; Yuan et al., 2024; Yamada et al., 2025). As scientific discovery is iterative, an AI Scientist must support both short-horizon refinement within a project and long-horizon accumulation across projects (Novikov et al., 2025; Tang et al., 2025a; Shi et al., 2025).

Current AI Scientist systems pursue evolution through three complementary mechanisms operating across different timescales:

Within a single project, self-reflection allows a system to act as both generator and reviewer, iteratively revising hypotheses, code or manuscripts in response to failed checks and evaluator feedback, as demonstrated by CycleResearcher and AI Scientist-v2 (Romera-Paredes et al., 2024; Weng

et al., 2025; Yamada et al., 2025).

Across iterations and projects, planning and memory mechanisms allocate resources among candidate directions and preserve previous artifacts, failed trials and feedback, enabling subsequent searches to build on accumulated experience (Tang et al., 2025a; Novikov et al., 2025).

Tool expansion provides a third route by increasing a system’s effective research capacity without modifying its model weights: systems such as SciToolAgent and ToolUniverse connect agents to scientific databases, simulators, theorem provers, laboratory interfaces, and standardized APIs (Ding et al., 2025; Gao et al., 2025). Together, these mechanisms make evolution both cognitive and infrastructural, determined by how effectively a system incorporates feedback, preserves experience and acquires reusable scientific capabilities.

Conceptually, evolution gives AI Scientists a cumulative dimension. A system that only completes one project remains a powerful automation pipeline; a system that learns from failures, incorporates feedback, and refines future search begins to resemble a developing scientific actor. The long-term importance of evolution is therefore the possibility that repeated research attempts become reusable experience.

9 Discussion

9.1 Enabling technologies for AI Scientist

Their emergence is driven by the co-evolution of three enabling technologies that operate together in a closed loop:

First, by integrating information across literature, **foundation models** can identify research questions (Lahat et al., 2023), formulate hypotheses (Si et al., 2025), and translate ideas into executable artifacts (Jiang et al., 2026).

Second, **automated experimental infrastructure** such as self-driving laboratories (Abolhasani and Kumacheva, 2023) and robotics (Cooper et al., 2025), enables repeatable experimentation and produce high-frequency feedback that can be fed back to foundation models for refinement.

Third, **agentic framework** endows foundation models with the capacity for proactive exploration (Yao et al., 2023). Through task planning, tool use, memory retrieval, and multi-agent collaboration, agents can coordinate iterative design–execute–analyze cycles, enabling end-to-end research workflows (Yamada et al., 2025).

Together, these technologies form a scalable loop: foundation models propose and interpret, automated laboratories execute and generate data, and agents orchestrate and verify, thereby giving the rise to AI Scientist systems (Lu et al., 2024).

9.2 Limitations of current AI Scientist

Current AI Scientist systems inherit limitations from their underlying foundation models, reducing their reliability in automated discovery workflows (Wang et al., 2023). Training-data **bias** may favor well-studied domains and well-resourced institutions while marginalizing underrepresented fields and regions (Gallegos et al., 2024; Resnik and Hosseini, 2025). **Hallucinations** remain a persistent risk (Ji et al., 2023), particularly in high-stakes domains such as chemistry, where factual errors may misdirect physical experiments or produce unsafe protocols (Tang et al., 2025b; Leong et al., 2025). Combined with **poorly calibrated uncertainty**, which obscures the distinction between speculation and well-supported conclusions (Ye et al., 2024), these weaknesses constrain their trustworthiness as research contributors.

Beyond model-level weaknesses, current systems remain **limited across the scientific discovery cycle** (Weng et al., 2025). During knowledge acquisition, they may summarize literature without adequately assessing methodological quality (Kusumegi et al., 2025; Stokel-Walker and Van Noorden, 2023). In hypothesis generation, they perform well when objectives and automated evaluators are clearly defined (Romera-Paredes et al., 2024), but their ability to formulate open-ended problems with less standardized evaluation criteria remains uncertain (Novikov et al., 2025). Physical validation can also be costly, particularly in resource-constrained fields, while long-term evolution remains unstable and prone to local optima.

Domain coverage is similarly imbalanced. Most current systems focus on fields with machine-readable data, well-defined objectives, and automation-friendly evaluation, such as machine learning, chemistry, and materials science. In areas such as the social sciences (Grossmann et al., 2023), however, research often depends on qualitative evidence, contextual interpretation, and competing conceptual frameworks, making autonomous “optimize-and-verify” workflows difficult to formulate without standardized benchmarks (Sourati and Evans, 2023).

9.3 Security risks

AI Scientist systems present security risks. In sensitive fields such as biology and chemistry, the potential for AI Scientist systems to be misused is a significant concern (Urbina et al., 2022). As autonomous agents, these systems **lack moral judgment**. They could intentionally be employed to create dangerous substances, manipulate biological data, or design unsafe medical treatments (Bengio et al., 2025). Moreover, these systems may contain **security vulnerabilities**, such as algorithmic instability or susceptibility to hacking, which could lead to serious dangers (Debenedetti et al., 2024). Therefore, ensuring the security of AI Scientist systems is a key challenge to their widespread application (He et al., 2025a).

10 Outlook

Scientific production. The long-term impact of AI Scientists is a reorganization of scientific discovery itself. As models, tools, and laboratories become more tightly coupled, science may move from a project-then-paper rhythm toward more continuous closed-loop discovery, where hypotheses, experiments, measurements, and revisions unfold as a persistent process (Stocker et al., 2025; Musslick et al., 2025). This would not remove human creativity, but it could change where it is most needed: less in routine screening and baseline verification, and more in framing consequential questions and interpreting unexpected results.

Scientific evaluation. If AI Scientist systems can rapidly generate manuscript-like artifacts, the paper may lose some of its role as the primary evidence of scientific work. Evaluation will need to move closer to the research process itself: the hypotheses considered, data used, protocols executed, failures retained, and human interventions made along the way. This shift does not make papers irrelevant; rather, it makes them summaries of an underlying trajectory that should itself be inspectable, reproducible, and open to challenge.

Human role and societal pathway. AI Scientists may develop along two complementary paths. For individual researchers, they can become personalized research partners that help track literature, explore ideas, run baselines, and preserve project memory. For society, they can become large-scale discovery infrastructure for problems that require sustained search across diverse disciplines. In both paths, their value will depend on expanding the

scale of exploration while preserving human responsibility for problem framing, governance, and societal purpose.

11 Conclusion

This paper presented a systematic survey of existing research on AI Scientist systems, providing a comprehensive overview of advances in the field. We first critically analyzed current research and highlighted gaps AI Scientist systems must address before they can make ground-breaking discoveries that solve grand challenges across medicine, energy, and the environment, thereby changing the world and reshaping the scientific research paradigm. Finally, we concluded the survey by outlining future directions essential for closing the gap and advancing the field toward truly autonomous scientific intelligence.

Limitations

This survey is limited by the fast pace of AI Scientist research. Several systems discussed here are recent preprints or open-source artifacts, and their reported capabilities may change as code, evaluations, and independent replications appear. Our evidence levels are therefore intended as a snapshot of demonstrated support rather than a permanent ranking. In addition, the survey emphasizes systems that coordinate scientific workflows; it does not exhaustively cover all foundation models for science, all self-driving laboratory platforms, or all scientific information-retrieval tools.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62336006); the National Natural Science Foundation of China Key Project under Grant No. 625B2152 (Research on Evaluation Paradigms and Closed-Loop Optimization Mechanisms for the Automated Discovery of AI Algorithms), and by the 2025 Dean's Special "PhD Student Project" (WU2025D007) of the School of Engineering, Westlake University. Linyi Yang was supported by the High-level Special Funds (No.G03050K007) of Southern University of Science and Technology.

References

Milad Abolhasani and Eugenia Kumacheva. 2023. The rise of self-driving labs in chemical and materials sciences. *Nature Synthesis*, 2(6):483–492.

Atilla Kaan Alkan, Shashwat Sourav, Maja Jablonska, Simone Astarita, Rishabh Chakrabarty, Nikhil Garuda, Pranav Khetarpal, Maciej Pióro, Dimitrios Tanoglidis, Kartheik G. Iyer, Mugdha S. Polimera, Michael J. Smith, Tirthankar Ghosal, Marc Huertas-Company, Sandor Kruk, Kevin Schawinski, and Ioana Ciucă. 2025. [A survey on hypothesis generation for scientific discovery in the era of large language models](#). *Preprint*, arXiv:2504.05496.

Nicholas H Angello, David M Friday, Changhyun Hwang, Seungjoo Yi, Austin H Cheng, Tiara C Torres-Flores, Edward R Jira, Wesley Wang, Alán Aspuru-Guzik, Martin D Burke, Charles M Schroeder, Ying Diao, and Nicholas E Jackson. 2024. Closed-loop transfer enables artificial intelligence to yield chemical knowledge. *Nature*, 633(8029):351–358.

Raghav Awasthi, Atharva Bhattad, Sai Prasad Ramachandran, Shreya Mishra, Ashish K Khanna, Jacek B Cywinski, Kamal Maheshwari, Dwarikanath Mahapatra, Izabella DiRosa, Anabelle Cohen, Hajra Arshad, Aarit Atreja, Asma Alshukaili, Aryan Vohra, Nishant Singh, Francis A Papay, Ashish Atreja, Rahul Kashyap, and Piyush Mathur. 2025. Human evaluation of large language models in healthcare: gaps, challenges, and the need for standardization. *npj Health Systems*, 2(1):40.

Monya Baker. 2016. [1,500 scientists lift the lid on reproducibility](#). *Nature*, 533:452–454.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. [Scibert: A pretrained language model for scientific text](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620. Association for Computational Linguistics.

Yoshua Bengio, Michael Cohen, Damiano Fornasiero, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, Marc-Antoine Rondeau, Pierre-Luc St-Charles, and David Williams-King. 2025. Superintelligent agents pose catastrophic risks: Can scientist ai offer a safer path? *arXiv preprint arXiv:2502.15657*.

Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. 2024. OceanGPT: A large language model for ocean science tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3357–3372.

Federico Bianchi, Owen Queen, Nitya Thakkar, Eric Sun, and James Zou. 2025. Exploring the use of ai authors and reviewers at agents4science. *Nature Biotechnology*, pages 1–4.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

- Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldasari, Andrew D White, and Philippe Schwaller. 2024. [Augmenting large language models with chemistry tools](#). *Nature Machine Intelligence*, 6:525–535.
- Qiguang Chen, Mingda Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, Mengkang Hu, Yimeng Zhang, Yihao Liang, Yuhang Zhou, Jiaqi Wang, Zhi Chen, and Wanxiang Che. 2025. [Ai4research: A survey of artificial intelligence for scientific research](#). *Preprint*, arXiv:2507.01903.
- Andrew I Cooper, Patrick Courtney, Kourosh Darvish, Moritz Eckhoff, Hatem Fakhroldeen, Andrea Gabrielli, Animesh Garg, Sami Haddadin, Kanako Harada, Jason Hein, Maria Hübner, Dennis Knobbe, Gabriella Pizzuto, Florian Shkurti, Ruja Shrestha, Kerstin Thurow, Rafael Vescovi, Birgit Vogel-Heuser, Ádám Wolf, and 4 others. 2025. Accelerating discovery in natural science laboratories with ai and robotics: Perspectives and challenges. *Science Robotics*, 10(106):eadv7932.
- Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. 2024. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents. *Advances in Neural Information Processing Systems*, 37:82895–82920.
- Keyan Ding, Jing Yu, Junjie Huang, Yuchen Yang, Qiang Zhang, and Huajun Chen. 2025. Scitoolagent: a knowledge-graph-driven scientific agent for multitool integration. *Nature Computational Science*, 5(10):962–972.
- Steffen Eger, Yong Cao, Jennifer D’Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, Chenghua Lin, Nafise Sadat Moosavi, Wei Zhao, and Tristan Miller. 2025. [Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation](#). *Preprint*, arXiv:2502.05151.
- Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational linguistics*, 50(3):1097–1179.
- Shanghai Gao, Richard Zhu, Pengwei Sui, Zhenglun Kong, Sufian Aldogom, Yepeng Huang, Ayush Noori, Reza Shamji, Krishna Parvataneni, Theodoros Tsiligkaridis, and Marinka Zitnik. 2025. Democratizing ai scientists using tooluniverse. *arXiv preprint arXiv:2509.23426*.
- Hugh G Gauch. 2003. *Scientific method in practice*. Cambridge University Press.
- Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. 2023. Ai and the transformation of social science research. *Science*, 380(6650):1108–1109.
- Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Myles Kim, Corey M Williams, Stefan Bekiranov, and Aidong Zhang. 2025. Ideabench: Benchmarking large language models for research idea generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5888–5899.
- Sonakshi Gupta, Akhlak Mahmood, Pranav Shetty, Aishat Adeboye, and Rampi Ramprasad. 2024. Data extraction from polymer literature using large language models. *Communications materials*, 5(1):269.
- Feng He, Tianqing Zhu, Dayong Ye, Bo Liu, Wanlei Zhou, and Philip S Yu. 2025a. The emerged security and privacy of llm agent: A survey with case studies. *ACM Computing Surveys*, 58(6):1–36.
- Yong He, Pan Fang, Yongtao Shan, Yuanfei Pan, Yanhong Wei, Yichang Chen, Yihao Chen, Yi Liu, Zhenyu Zeng, Zhan Zhou, Feng Zhu, Edward C Holmes, Jieping Ye, Jun Li, Yuelong Shu, Mang Shi, and Zhaorong Li. 2025b. Generalized biological foundation model with unified nucleic acid and protein language. *Nature Machine Intelligence*, 7(6):942–953.
- Matthew Hutson and Ashley Mastin. 2023. Guinea pigbots. *Science*, 381(6654):121–123.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.
- Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2026. [A survey on large language models for code generation](#). *ACM Transactions on Software Engineering and Methodology*, 35(2):1–72.
- John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, and 15 others. 2021. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589.
- Andrej Karpathy. 2026. [autoresearch](#). GitHub repository.
- Ross D King, Jem Rowland, Stephen G Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N Soldatova, Andrew Sparkes, Kenneth E Whelan, and Amanda Clare. 2009. The automation of science. *Science*, 324(5923):85–89.

- Hiroaki Kitano. 2021. Nobel turing challenge: creating the engine for scientific discovery. *NPJ systems biology and applications*, 7(1):29.
- Alexander Krauss. 2025. New scientific fields are triggered by powerful new methods. *Humanities and Social Sciences Communications*, 12(1):1–12.
- Keigo Kusumegi, Xinyu Yang, Paul Ginsparg, Mathijs de Vaan, Toby Stuart, and Yian Yin. 2025. Scientific production in the era of large language models. *Science*, 390(6779):1240–1243.
- Adi Lahat, Eyal Shachar, Benjamin Avidan, Zina Shatz, Benjamin S Glicksberg, and Eyal Klang. 2023. Evaluating the use of large language model in identifying top research questions in gastroenterology. *Scientific Reports*, 13(1):4164.
- P Langley. 1987. *Scientific discovery: Computational explorations of the creative processes*. MIT Press.
- Pat Langley. 1977. Bacon: A production system that discovers empirical laws. In *IJCAI*, page 344.
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.
- Shi Xuan Leong, Caleb E Griesbach, Rui Zhang, Kourosh Darvish, Yuchi Zhao, Abhijoy Mandal, Yunheng Zou, Han Hao, Varinia Bernales, and Alán Aspuru-Guzik. 2025. Steering towards safe self-driving laboratories. *Nature Reviews Chemistry*, 9(10):707–722.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Haotong Liang, Chuangye Wang, Heshan Yu, Dylan Kirsch, Rohit Pant, Austin McDannald, A Gilad Kusne, Ji-Cheng Zhao, and Ichiro Takeuchi. 2025. Real-time experiment-theory closed-loop interaction for autonomous materials science. *Science Advances*, 11(27):eadu7426.
- Robert K Lindsay, Bruce G Buchanan, Edward A Feigenbaum, and Joshua Lederberg. 1993. Dendral: a case study of the first expert system for scientific hypothesis formation. *Artificial intelligence*, 61(2):209–261.
- Jiaqi Liu, Shi Qiu, Mairui Li, Bingzhou Li, Haonian Ji, Siwei Han, Xinyu Ye, Peng Xia, Zihan Dong, Meng Chen, Congyu Zhang, Letian Zhang, Guiming Chen, Haoqin Tu, Xinyu Yang, Lu Feng, Xujiang Zhao, Haifeng Chen, Jiawei Zhou, and 17 others. 2026. *Autoresearchclaw: Self-reinforcing autonomous research with human-ai collaboration*. Preprint, arXiv:2605.20025.
- Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. *The ai scientist: Towards fully automated open-ended scientific discovery*. arXiv preprint arXiv:2408.06292v3.
- Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. Llm4sr: A survey on large language models for scientific research. arXiv preprint arXiv:2501.04306.
- Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, James S Fraser, and Nikhil Naik. 2023. Large language models generate functional protein sequences across diverse families. *Nature biotechnology*, 41(8):1099–1106.
- Indrajeet Mandal, Jitendra Soni, Mohd Zaki, Morten M Smedskjaer, Katrin Wondraczek, Lothar Wondraczek, Nitya Nand Gosvami, and NM Anoop Krishnan. 2025. Evaluating large language model agents for automation of atomic force microscopy. *Nature Communications*, 16(1):9104.
- René Marois and Jason Ivanoff. 2005. Capacity limits of information processing in the brain. *Trends in cognitive sciences*, 9(6):296–305.
- John McDermott. 1982. R1: A rule-based configurer of computer systems. *Artificial intelligence*, 19(1):39–88.
- Pawel Mieszczanek, Peter Corke, Courosh Mehanian, Paul D Dalton, and Dietmar W Huttmacher. 2024. Towards industry-ready additive manufacturing: Ai-enabled closed-loop control for 3d melt electrowriting. *Communications Engineering*, 3(1):158.
- Sebastian Musslick, Laura K Bartlett, Suyog H Chandramouli, Marina Dubova, Fernand Gobet, Thomas L Griffiths, Jessica Hullman, Ross D King, J Nathan Kutz, Christopher G Lucas, Suhas Mahesh, Franco Pestilli, Sabina J Sloman, and William R Holmes. 2025. Automating the practice of science: Opportunities, challenges, and implications. *Proceedings of the National Academy of Sciences*, 122(5):e2401238121.
- Allen Newell and Herbert Simon. 1956. The logic theory machine—a complex information processing system. *IRE Transactions on information theory*, 2(3):61–79.
- Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. 2025. *Alphaevolve: A coding agent*

- for scientific and algorithmic discovery. *Preprint*, arXiv:2506.13131.
- Junhyuk Oh, Gregory Farquhar, Iurii Kemaev, Dan A Calian, Matteo Hessel, Luisa Zintgraf, Satinder Singh, Hado Van Hasselt, and David Silver. 2025. [Discovering state-of-the-art reinforcement learning algorithms](#). *Nature*, 648(8093):312–319.
- Shuo Ren, Pu Jian, Zhenjiang Ren, Chunlin Leng, Can Xie, and Jiajun Zhang. 2025. Towards scientific intelligence: A survey of llm-based scientific agents. *arXiv preprint arXiv:2503.24047*.
- David B Resnik and Mohammad Hosseini. 2025. The ethics of using artificial intelligence in scientific research: new guidance needed for a new tool. *AI and Ethics*, 5(2):1499–1521.
- Loïc M Roch, Florian Häse, Christoph Kreisbeck, Teresa Tamayo-Mendoza, Lars PE Yunker, Jason E Hein, and Alán Aspuru-Guzik. 2018. Chemos: orchestrating autonomous experimentation. *Science Robotics*, 3(19):eaat5559.
- Bernardino Romera-Paredes, Mohammadamin Barekatin, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475.
- Atiqer Rahman Sarkar, Yao-Shun Chuang, Noman Mohammed, and Xiaoqian Jiang. 2024. De-identification is not enough: a comparison between de-identified and synthetic clinical notes. *Scientific reports*, 14(1):29669.
- Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyan Wang, Yubin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2025. [Continual learning of large language models: A comprehensive survey](#). *ACM Computing Surveys*, 58(5):1–42.
- Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2025. [Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers](#). In *The Thirteenth International Conference on Learning Representations*.
- Chenglei Si, Zitong Yang, Yejin Choi, Emmanuel Candès, Diyi Yang, and Tatsunori Hashimoto. 2026. [Towards execution-grounded automated ai research](#). In *Forty-third International Conference on Machine Learning*.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Jamshid Sourati and James A Evans. 2023. Accelerating science with human-aware artificial intelligence. *Nature human behaviour*, 7(10):1682–1696.
- Springer Nature. 2025. [Editorial policies](#).
- Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. 2025. [PaperBench: Evaluating AI’s ability to replicate AI research](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 56843–56873. PMLR.
- Markus Stocker, Lauren Snyder, Matthew Anfuso, Oliver Ludwig, Freya Thießen, Kheir Eddine Farfar, Muhammad Haris, Allard Oelen, and Mohammad Yaser Jaradeh. 2025. Rethinking the production and publication of machine-readable expressions of research findings. *Scientific Data*, 12(1):677.
- Chris Stokel-Walker and Richard Van Noorden. 2023. What chatgpt and generative ai mean for science. *Nature*, 614(7947):214–216.
- Don R Swanson. 1986. Undiscovered public knowledge. *The Library Quarterly*, 56(2):103–118.
- Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. 2025. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 646(8085):716–723.
- Alexander Szalay and Jim Gray. 2006. Science in an exponential world. *Nature*, 440(7083):413–414.
- Nathan J Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E Kumar, Tanjin He, David Milsted, Matthew J McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. 2023. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624(7990):86.
- Lauren Takahashi, Mikael Kuwahara, and Keisuke Takahashi. 2025. Ai and automation: democratizing automation and the evolution towards true ai-autonomous robotics. *Chemical Science*, 16(35):15769–15780.
- Xiangru Tang, Tianyu Hu, Muyang Ye, Yanjun Shao, Xunjian Yin, Siru Ouyang, Wangchunshu Zhou, Pan Lu, Zhuosheng Zhang, Yilun Zhao, Arman Cohan, and Mark Gerstein. 2025a. Chemagent: Self-updating memories in large language models improves chemical reasoning. In *The Thirteenth International Conference on Learning Representations*.
- Xiangru Tang, Qiao Jin, Kunlun Zhu, Tongxin Yuan, Yichi Zhang, Wangchunshu Zhou, Meng Qu, Yilun Zhao, Jian Tang, Zhuosheng Zhang, Arman Cohan, Dov Greenbaum, Zhiyong Lu, and Mark Gerstein.

- 2025b. Risks of ai scientists: prioritizing safeguarding over autonomy. *Nature Communications*, 16(1):8317.
- Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.
- Gary Tom, Stefan P. Schmid, Sterling G. Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M. Rajaonson, Marta Skreta, Naruki Yoshikawa, Samantha Corapi, Gun Deniz Akkoc, Felix Strieth-Kalthoff, Martin Seifrid, and Alán Aspuru-Guzik. 2024. [Self-driving laboratories for chemistry and materials science](#). *Chemical Reviews*, 124(16):9633–9732.
- Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. 2022. Dual use of artificial-intelligence-powered drug discovery. *Nature machine intelligence*, 4(3):189–191.
- Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, and 11 others. 2023. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.
- Yi Wang, Yuejie Hou, Lin Yang, Shisen Li, Weiting Tang, Hui Tang, Qiushun He, Siyuan Lin, Yanyan Zhang, Xingyu Li, Shiwen Chen, Yusheng Huang, Lingsong Kong, Huijun Zhang, Duncan Yu, Feng Mu, Huanming Yang, Jian Wang, Nattiya Hirankarn, and Meng Yang. 2026. [Accelerating primer design for amplicon sequencing using large language model-powered agents](#). *Nature Biomedical Engineering*, 10:338–353.
- Jiaqi Wei, Yuejin Yang, Xiang Zhang, Yuhan Chen, Xiang Zhuang, Zhangyang Gao, Dongzhan Zhou, Guangshuai Wang, Zhiqiang Gao, Juntao Cao, Zijie Qiu, Ming Hu, Chenglong Ma, Shixiang Tang, Junjun He, Chunfeng Song, Xuming He, Qiang Zhang, Chenyu You, and 8 others. 2025. [From ai for science to agentic science: A survey on autonomous scientific discovery](#). *Preprint*, arXiv:2508.14111.
- Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. [Cycleresearcher: Improving automated research via automated review](#). In *The Thirteenth International Conference on Learning Representations*.
- Yixuan Weng, Minjun Zhu, Qiuji Xie, QiYao Sun, Zhen Lin, Sifan Liu, and Yue Zhang. 2026. [Deepscientist: Advancing frontier-pushing scientific findings progressively](#). In *The Fourteenth International Conference on Learning Representations*.
- Fan Wu, Cheng Chen, Zhenshan Tan, Taiyu Zhang, Xinzheng Xu, Yanyu Qian, Dingcheng Gao, Lanyun Zhu, Qi Zhu, Yi Tan, Deyi Ji, Guosheng Lin, Tianrun Chen, Deheng Ye, and Fayao Liu. 2026. [Claw ai lab: An autonomous multi-agent research team](#). *Preprint*, arXiv:2605.22662.
- Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*.
- Zijie Yang, Yukai Wang, and Lijing Zhang. 2023. [Ai becomes a masterbrain scientist](#). *bioRxiv*, pages 2023–04.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.
- Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*, 37:15356–15385.
- Jiakang Yuan, Xiangchao Yan, Bo Zhang, Tao Chen, Botian Shi, Wanli Ouyang, Yu Qiao, Lei Bai, and Bowen Zhou. 2025. [Dolphin: Moving towards closed-loop auto-research through thinking, practice, and feedback](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21768–21789, Vienna, Austria. Association for Computational Linguistics.
- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. 2024. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*.
- Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024. [A comprehensive survey of scientific large language models and their applications in scientific discovery](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8783–8817, Miami, Florida, USA. Association for Computational Linguistics.
- Tianshi Zheng, Zheyang Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. [From automation to autonomy: A survey on large language models in scientific discovery](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17733–17750, Suzhou, China. Association for Computational Linguistics.
- Lianhao Zhou, Hongyi Ling, Cong Fu, Yepeng Huang, Michael Sun, Wendi Yu, Xiaoxuan Wang, Xiner Li, Xingyu Su, Junkai Zhang, Xiusi Chen, Chenxing Liang, Xiaofeng Qian, Heng Ji, Wei Wang, Marinka Zitnik, and Shuiwang Ji. 2025. [Autonomous](#)

agents for scientific discovery: Orchestrating scientists, language, code, and physics. *Preprint*, arXiv:2510.09901.

Kunlun Zhu, Jiaxun Zhang, Ziheng Qi, Nuoxing Shang, Zijia Liu, Peixuan Han, Yue Su, Haofei Yu, and Jiaxuan You. 2025. [Safescientist: Enhancing ai scientist safety for risk-aware scientific discovery](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2317, Suzhou, China. Association for Computational Linguistics.

A Detailed Evidence and Mechanism Comparisons

The main text reports the highest evidence level fully supported for each representative entry. Table 4 gives the public evidence supporting that level and the evidence missing for the next level. This distinction prevents isolated logs, same-team checks, or infrastructure capabilities from being interpreted as complete auditable research loops or independent validation.

B Hypothesis Quality Criteria

Hypothesis generation is often the first capability that makes an AI Scientist appear creative, but a generated hypothesis is not valuable merely because it sounds novel. In an AI Scientist workflow, a hypothesis is a commitment that shapes later retrieval, experiment design, resource allocation, and interpretation (Swanson, 1986; Lu et al., 2024). Its quality should therefore be judged not only as an isolated sentence, but as an element in a research trajectory that can be inspected and challenged.

Novelty concerns whether the hypothesis departs meaningfully from existing work. This does not mean rewarding superficial novelty or unusual wording. A useful hypothesis should expose a genuine gap, combine evidence in a way that was not already obvious, or formulate a question that redirects attention toward an underexplored mechanism or design space (Swanson, 1986; Romera-Paredes et al., 2024). Recent studies of AI-generated research ideas often evaluate novelty against expert judgments and prior literature (Si et al., 2025; Guo et al., 2025; Awasthi et al., 2025).

For AI Scientist systems, however, novelty should be accompanied by provenance: the system should show which papers, observations, or failures motivated the proposed direction, so that reviewers can distinguish a real research opportunity from a recombination of familiar claims.

Feasibility asks whether the hypothesis can be pursued with available methods, data, instruments, and budgets. A hypothesis may be imaginative but scientifically unhelpful if it cannot be turned into a testable plan. Feasibility is domain dependent: in machine learning it may depend on datasets, baselines, and compute (Starace et al., 2025; Yamada et al., 2025); in chemistry or materials science it may depend on synthesis protocols, instrument availability, and safety constraints (Bran et al., 2024; Boiko et al., 2023; Szymanski et al.,

2023); in social science it may depend on participant access, measurement validity, and ethical approval (Grossmann et al., 2023). This criterion should not force AI Scientists toward only low-risk ideas, but it should require them to make the path from claim to validation explicit.

Clarity and testability connect hypothesis generation to experimental validation. A clear hypothesis specifies the variables, expected relationships, assumptions, and conditions under which it should hold. A testable hypothesis further identifies what evidence would support it, weaken it, or falsify it. Vague statements such as “method X may improve task Y” are easy to generate but difficult to evaluate. By contrast, a testable hypothesis can be translated into metrics, controls, benchmarks, protocols, or measurement plans (Starace et al., 2025). This is why hypothesis quality cannot be separated from the later validation stage: a good hypothesis already carries the shape of a possible test.

Scientific value captures why the hypothesis matters. Some generated hypotheses are feasible and clear but scientifically minor: they may optimize a narrow benchmark, repeat a familiar ablation, or produce results whose interpretation would not change understanding. A valuable hypothesis should help resolve uncertainty, reveal a mechanism, improve a meaningful capability, or open a path to downstream applications (Si et al., 2025; Romera-Paredes et al., 2024). In this sense, the “excitement” criterion used in some idea-generation evaluations should be interpreted carefully. The point is not rhetorical excitement, but whether the hypothesis would justify the attention, cost, and risk of the subsequent research process.

Grounding is especially important for AI Scientist systems because fluent hypotheses can hide weak evidence. A grounded hypothesis should be connected to retrieved literature, domain constraints, prior failures, or empirical observations (Lewis et al., 2020; Liu et al., 2024; Gupta et al., 2024; Yuan et al., 2025). It should also indicate where expert judgment remains necessary, particularly when the hypothesis depends on tacit domain knowledge or ambiguous evidence (Awasthi et al., 2025).

C Risk and Governance Considerations

The risks of AI Scientist systems are not limited to incorrect outputs. Because these systems can retrieve information, plan actions, call tools, gener-

Table 4: Evidence-level rationales. Each system receives one level based on public evidence. The final column identifies the missing evidence required for the next level.

System	Level	Supporting evidence	Evidence missing for the next level
The AI Scientist (2024)	E2	Code experiments produce metrics, journals, plots, and manuscripts.	Public links among rejected hypotheses, failed branches, interventions, and final claims.
AI Scientist-v2 (2025)	E2	Agentic tree search runs code experiments and produces measured results and papers.	A complete trace of explored and rejected branches, failures, decisions, and conclusions.
DeepScientist (2026)	E3	Findings Memory connects ideas, staged validation, and reported results.	Independent validation of the reported scientific improvements.
Virtual Lab (2025)	E3	Discussions, code, screening scores, and wet-lab data form an inspectable artifact chain.	Validation or replication outside the original collaboration.
A-Lab (2023)	E3	Robotic synthesis, characterization, active learning, and failure analysis form an inspectable loop.	Replication by a team independent of the original study.
Coscientist (2023)	E2	Retrieval, planning, code, and robotics provide measurable feedback across six task classes.	Public end-to-end trajectories including failed paths and interventions.
SciToolAgent (2025)	E2	Knowledge-graph retrieval selects and executes scientific tools with recorded returns.	A complete hypothesis-to-validated-conclusion trajectory.
ToolUniverse (2025)	E2	Scientific APIs and tool-call interfaces provide executable infrastructure.	An auditable end-to-end trajectory from a concrete host system; infrastructure autonomy is not rated.
autoresearch (2026)	E2	Code edits, fixed-budget runs, metrics, and retain/discard/crash status are logged.	A broader scientific trajectory beyond optimizing one fixed metric.
AutoResearchClaw (2026)	E3	Failure recovery, result reporting, interventions, and cross-run memory connect execution and reporting.	Independent validation beyond benchmark and model-based evaluation.
Claw AI Lab (2026)	E3	Monitoring, artifact inspection, execution harnesses, and rollback expose a controllable trajectory.	Independent validation beyond internal cases and model-based judges.

ate code, operate experimental infrastructure, and produce publishable artifacts, their risks span technical reliability, security, scientific institutions, and social distribution. Governance must therefore address the research process as a whole rather than only screening final manuscripts.

Reliability risks originate from the foundation models and tools on which AI Scientists depend. Hallucination can introduce false claims into literature reviews, protocols, or generated explanations (Ji et al., 2023). Bias in training data can shape which problems appear promising, often favoring well-represented domains and well-resourced communities (Gallegos et al., 2024). Out-of-distribution generalization remains particularly important at the scientific frontier, where new molecules, instruments, datasets, or conceptual frameworks may differ from the model’s training distribution (Wang et al., 2023). Poor uncertainty calibration further makes it difficult to distinguish speculation from well-supported conclusions (Ye et al., 2024). These risks are amplified in closed-loop systems because an early error can influence later hypothesis generation, experimental design, and interpretation.

Security and dual-use risks become more serious as AI Scientists gain autonomy and tool ac-

cess. In biology, chemistry, medicine, and cybersecurity, systems that accelerate discovery can also lower barriers to harmful procedures or unsafe designs (Urbina et al., 2022; Bengio et al., 2025; Tang et al., 2025b). Agentic systems introduce additional vulnerabilities: prompt injection, insecure tool permissions, adversarial instructions, or compromised dependencies may alter research actions or leak sensitive information (Debenedetti et al., 2024; He et al., 2025a). These risks require more than content filtering. They call for sandboxed execution, restricted tool permissions, risk screening before sensitive actions, red-team evaluation, and logs that make the system’s actions auditable after the fact.

Publication and evaluation risks arise because AI Scientist systems can generate manuscript-like artifacts at low cost. If such artifacts are evaluated only through traditional peer review, submission volume may increase faster than human reviewing capacity, and fluent writing may obscure weak evidence (Bianchi et al., 2025). Authorship and responsibility also become more complex. Current editorial norms generally hold that AI tools should not be listed as authors because they cannot assume responsibility for the work (Springer Nature, 2025). Human researchers and institutions must remain accountable for claims, data use and experimental

Table 5: Comparison with related surveys. Columns indicate whether each survey centrally covers AI/LLM science, scientific agents, self-driving labs (SDLs), the AI Scientist boundary, the four-capability loop, and auditable research trajectories.

Survey	Primary focus	AI/LLM	Agents	SDLs	Boundary	Loop	Audit
Wei et al. (2025)	Agentic science	✓	✓	✓	✗	✗	✗
Zheng et al. (2025)	LLMs in discovery	✓	✓	✗	✗	✗	✗
Chen et al. (2025)	AI for research	✓	✗	✗	✗	✗	✗
Eger et al. (2025)	LLM-assisted science	✓	✗	✗	✗	✗	✗
Luo et al. (2025)	LLMs for research	✓	✗	✗	✗	✗	✗
Alkan et al. (2025)	Hypothesis generation	✗	✗	✗	✗	✗	✗
Zhou et al. (2025)	Autonomous agents	✗	✓	✓	✗	✗	✗
Zhang et al. (2024)	Scientific LLMs	✓	✗	✗	✗	✗	✗
Tom et al. (2024)	Self-driving labs	✗	✗	✓	✗	✗	✗
Ren et al. (2025)	Scientific agents	✗	✓	✗	✗	✗	✗
Leong et al. (2025)	Safe self-driving labs	✗	✗	✓	✗	✗	✗
This survey	AI Scientist systems	✓	✓	✓	✓	✓	✓

safety. This reinforces the need to evaluate AI-generated research through its trajectory: sources, decisions, experiments, failures, and human interventions should be available for inspection.

Privacy and data governance risks are central when AI Scientists operate over medical, biological, educational, or proprietary data. Even de-identified data may leak sensitive information or be combined with external sources in ways that create re-identification risks (Sarkar et al., 2024). AI Scientist systems also tend to accumulate logs, intermediate artifacts, and tool outputs, which may contain private information even when the final paper does not. Responsible deployment should therefore apply data minimization, access control, institutional review, provenance tracking, and careful rules for what can be stored, shared, or reused across projects.

Institutional and educational risks concern how AI Scientists may reshape the scientific ecosystem. If access to advanced models, compute, proprietary datasets, and automated laboratories is concentrated in a small number of institutions, AI-driven science may widen existing inequalities (Resnik and Hosseini, 2025). At the level of training, excessive reliance on automated systems may weaken the development of human skills in hypothesis formulation, experimental design, data analysis, and critical review. These risks do not imply that AI Scientists should be rejected. They imply that scientific institutions need clearer boundaries between automation and human responsibility, especially in education, authorship, funding, and high-stakes deployment.

These considerations suggest that governance should be built into the architecture of AI Scientist

systems. A responsible system should disclose AI involvement, preserve provenance, record human decisions, restrict unsafe tool use, and make its research trajectory available for review. The goal is to ensure that acceleration remains compatible with accountability, reproducibility, and public trust.