How Far Are AI Scientists from Changing the World? A Survey of Autonomous Scientific Discovery

Anonymous ACL submission

Abstract

AI Scientists are emerging as autonomous systems that coordinate multiple stages of scientific discovery, from knowledge acquisition and hypothesis generation to validation and iterative refinement. This survey asks how far such 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., 2023b) and **automated experimental infrastructures** (Jumper et al., 2021; Mieszczanek 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 Scientist (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

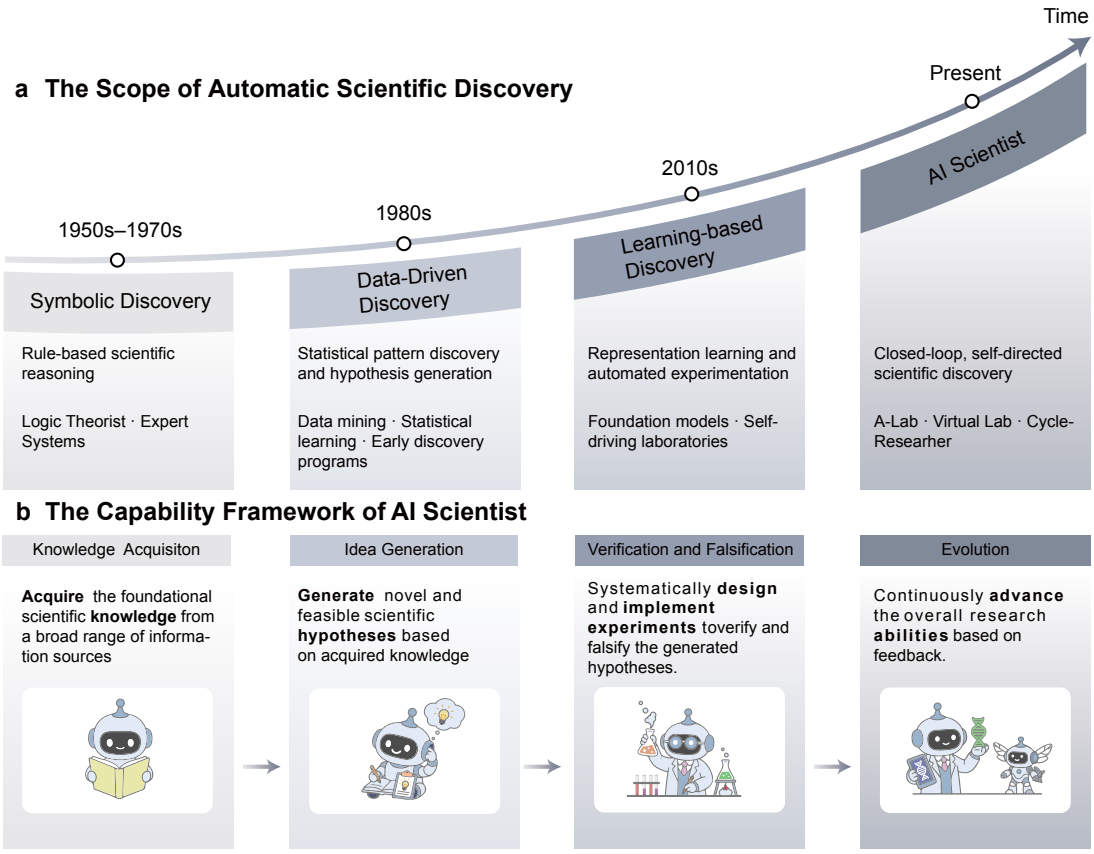

Figure 1: Timeline of automated scientific discovery. Automated scientific discovery has progressed from rule-based reasoning to closed-loop AI Scientist systems that integrate knowledge acquisition, hypothesis generation, validation, and evolution.

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., 2025a). Moreover, unlike human researchers, AI Scientists lack moral constraints (Tang et al., 2025b). They are incapable of making ethical judgments about the societal implications 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):

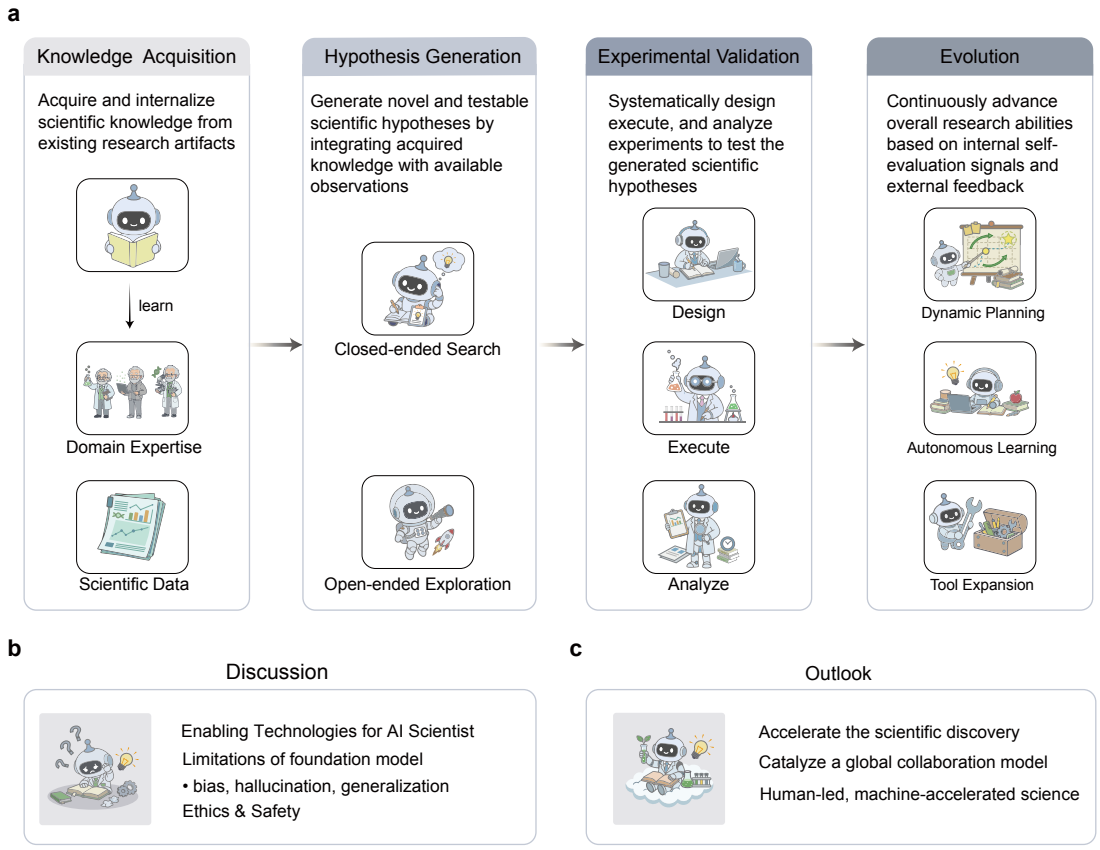

Figure 2: Overview of AI Scientist systems. This survey frames AI Scientists as closed-loop research systems that connect four core capabilities with enabling technologies, risks, and future directions.

First, we provide a definition of AI Scientists and distinguish them from adjacent systems (Section 2). Second, 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). Third, we synthesize representative systems by autonomy, execution grounding, verification, auditability, human role, output artifact, and evidence level, clarifying what has been demonstrated rather than merely proposed (Sections 3–4). 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, where the central object of evaluation is the inspectable research trajectory (Sections 9–10). Table 4 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. A mature AI Scientist approaches end-to-end closed-loop research under human governance. This definition intentionally avoids treating every scientific AI tool as an AI Scientist: the key distinction is not whether a system uses a foundation model, but whether it organizes a research trajectory across multiple scientific stages.

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, decisions, executable actions, measurements, 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 domain. By integrating accumulated knowledge with new observations, it then generates novel and testable hypotheses, which are validated through scientific

Table 1: Boundary between AI Scientist and adjacent systems. A system qualifies when it coordinates a hypothesis–test–revise trajectory rather than only executing a single scientific task.

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.

tools (Ding et al., 2025), automated experimental platforms (Roch et al., 2018), robotic laboratories (Takahashi et al., 2025), code execution, or simulation. Finally, an evolution mechanism closes the loop by using internal 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, augment, 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, benchmark, or auditable artifact. 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 logs, failures, and intervention points. Table 3 summarizes this map.

The pattern is bounded but clear. AI Scientists matter most when they shift automation from isolated tasks to executable and inspectable research trajectories. E1 artifacts are weak evidence for discovery; E2 systems ground claims in execution; E3 adds auditability and intervention; E4 requires external replication, domain acceptance, or durable use.

5 Knowledge Acquisition

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 ex-

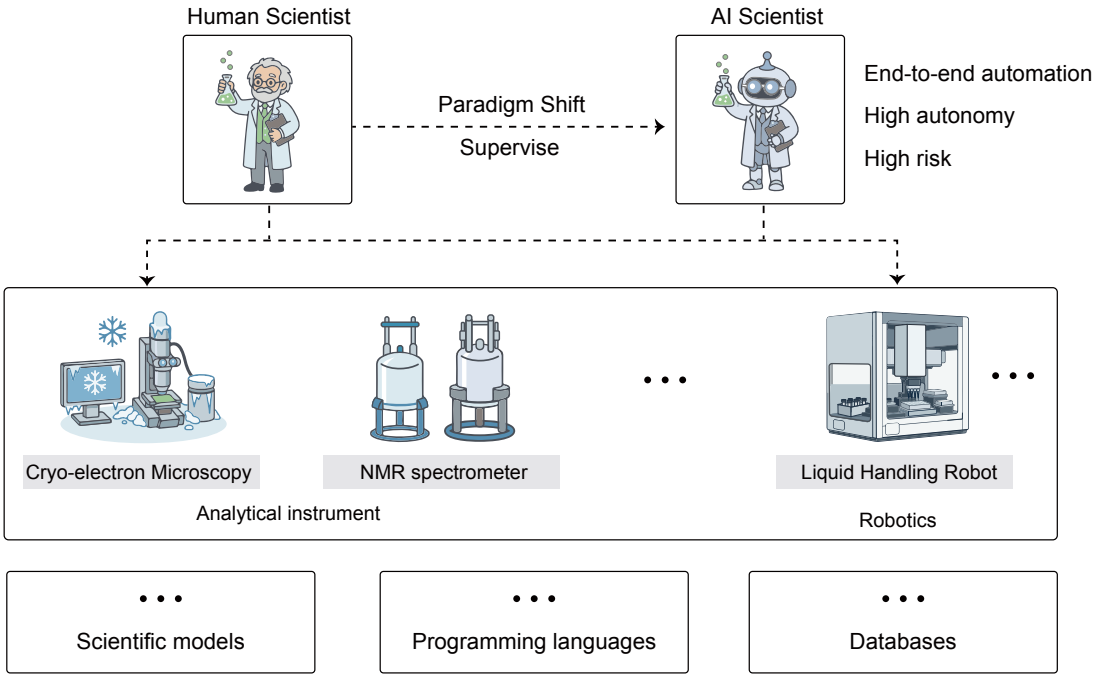

Figure 3: From human-in-the-loop science to AI Scientist-driven research. AI Scientist-driven research shifts human involvement from direct execution toward governance over autonomous systems that coordinate instruments, models and data.

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**	Exposes logs, artifacts, failures, provenance, interventions, or checks.
E4	**Validated impact**	Shows independently verified or adopted scientific impact.

perimental 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, constraints, tools, and open questions, and this context must remain revisable as new observations arrive. 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 knowledge 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, code, and database outputs 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.** Scientific knowledge is not a static corpus to be summarized once, but a changing background against which hypotheses become meaningful and experiments become interpretable. An AI Scientist therefore differs from a literature tool when acquired knowledge is not merely reported, but 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). Foundation-model-based systems

Table 3: Representative AI Scientist systems. We compare systems by loop coverage, grounding, auditability, human control, and evidence level.

System	Loop	Grounding	Audit trail	Human control	Evid.
The AI Scientist (2024)	idea → exp. → paper	code	logs; reviews	topic setup	E2
AI Scientist-v2 (2025)	idea → exp. → paper	code	logs; tree search	topic setup	E2–E3
DeepScientist (2026)	goal → idea → exp.	code	memory; logs	human tasks	E3–E4
Virtual Lab (2025)	question → design → screen	tools; wet lab	agent trace	scientist steering	E3–E4
A-Lab (2023)	design → synth. → analysis	robotic lab	instrument records	goal setting	E4
Coscientist (2023)	plan → execute	web; code; robotics	limited trace	supervision	E2–E3
SciToolAgent (2025)	plan → tool use	tools	tool trace	problem setup	E2
ToolUniverse (2025)	tool ecosystem	APIs	interface trace	integration	E2
auto-research (2026)	edit → run → select	code	experiment log	instruction; seed	E2
ResearchClaw (2026)	idea → exp. → report	code	failure recovery	interventions	E3
Claw AI Lab (2026)	roles → execution	code; data	dashboard; rollback	PI dashboard	E3

expand this space by generating **open-ended hypotheses** and iteratively refining them with retrieved literature, tool outputs, or experimental feedback (Si et al., 2024; Lu et al., 2024; Boiko et al., 2023). Evaluation typically considers novelty, feasibility, clarity, and scientific value (Si et al., 2024; Guo et al., 2025); automated reviewers and agentic critics can scale this process, but human expert judgment remains important when hypotheses depend on tacit domain context or ambiguous evidence (Weng et al., 2025; Awasthi et al., 2025).

Representative systems show two complementary mechanisms. One mechanism is **generative exploration**: systems such as The AI Scientist and Coscientist use language models, literature retrieval, code execution, and tool calls to propose candidate ideas and revise them against novelty or feasibility checks (Lu et al., 2024; Boiko et al., 2023). Another is **structured selection**: multi-agent or reviewer-style workflows assign roles to idea generators, critics, and domain-specific evaluators, so that a large candidate pool can be filtered before costly validation. PrimeGen, for instance, decomposes primer design into constraint interpretation, candidate generation, and thermodynamic/off-target evaluation (Wang et al., 2025), while CycleResearcher uses automated review feedback to improve research proposals across iterations (Weng et al., 2025). This makes hypothesis generation a search process with memory, constraints, and evaluation signals.

Conceptually, hypothesis generation is the step where an AI Scientist moves from representing existing knowledge to opening a space of possible discoveries. Its significance is not the production of more ideas, but 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 hypotheses through experiment **design, execution, and result analysis.** Computational systems generate code, define metrics, and run controlled evaluations (Starace et al., 2025); laboratory systems synthesize machine-executable protocols and operate instruments or robotic platforms (Boiko et al., 2023; Ruan et al., 2024; Szymanski et al., 2023). **Execution may be physical**, as in A-Lab and instrument-level 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**, where AI agents approximate human participants for preliminary social-science exploration (Hutson and Mastin, 2023; Grossmann et al., 2023).

The validation mechanism typically has three linked stages. First, the system **translates a hypothesis into testable artifacts**, such as measurable objectives, experimental conditions, or protocols; for example, a materials hypothesis may become synthesis parameters, measurement plans, and stability targets (Angello et al., 2024; Szymanski et al., 2023). Second, it **selects an em-**

pirical setting: code and benchmarks for computational research, simulation or theory-coupled loops for expensive physical domains, and robotic or instrument-level automation for laboratory settings (Starace et al., 2025; Liang et al., 2025; Mandal et al., 2025). Third, **it interprets observations,** including negative results and failure records, as evidence for revising the next decision. This analysis step is what makes systems such as ChemCrow and A-Lab scientifically meaningful: tool outputs and instrument records are not just terminal products, but constraints on the next round of reasoning (Bran et al., 2023; 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 resist, revise, or falsify those claims. The scientific force of this capability lies in converting linguistic possibility into evidence, so that generated hypotheses become part of a testable research process.

8 Evolution

Evolution is the capability to improve research behavior through internal self-evaluation and external feedback (Weng et al., 2025; Yuan et al., 2024; Yamada et al., 2025). Because 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., 2024).

Current systems pursue evolution through autonomous learning, tool expansion, and planning with memory. **Self-reflection** treats the system as both generator and reviewer, enabling iterative critique of hypotheses, code, experiments, or manuscripts (Romera-Paredes et al., 2024; Weng et al., 2025). **Tool-based systems** such as SciToolAgent and ToolUniverse expand capability by connecting agents to scientific tools, databases, solvers, and standardized APIs (Ding et al., 2025; Gao et al., 2025). **Planning and memory mechanisms,** as in AI Scientist-v2 and AlphaEvolve, allocate resources across candidate directions and reuse past artifacts, failures, and evaluator feedback (Yamada et al., 2025; Novikov et al., 2025).

In practice, evolution operates at several timescales. **Within a single project,** self-reflection and reviewer agents revise hypotheses, code, and analyses after failed checks, as in CycleResearcher and AI Scientist-v2 (Weng et al., 2025; Yamada et al., 2025). **Across projects,** memory mechanisms preserve failed trials, experimental records, and selected artifacts so that future searches do not restart from scratch (Tang et al., 2025a; Novikov et al., 2025). **Tool expansion adds a third route:** instead of changing model weights, systems can grow their effective research capacity by gaining access to external scientific resources, such as databases, simulators, theorem provers, laboratory interfaces, and domain tools (Ding et al., 2025; Gao et al., 2025). These mechanisms make evolution partly cognitive and partly institutional: progress depends on how the system preserves experience, absorbs feedback, and decides which scientific capacities to reuse or acquire.

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 not simply better performance on the next run, but the possibility that repeated research attempts become reusable experience.

9 Discussion

Enabling technologies. AI Scientist systems arise from the closed-loop coupling of foundation models, automated experimental infrastructure, and agentic orchestration (Rood et al., 2024; Leong et al., 2025; Yao et al., 2023a). Foundation models integrate literature, datasets, and experimental records to support question identification, hypothesis generation, code writing, and interpretation (Xu et al., 2024; Liu et al., 2025; Si et al., 2024; Jiang et al., 2026). Automated laboratories and robotics turn experiments into digitized, repeatable workflows (Abolhasani and Kumacheva, 2023; Cooper et al., 2025). Agentic systems organize multi-step scientific activity by linking planning, evidence retrieval, role coordination, and revision (Yamada et al., 2025). The strongest systems are therefore integrated research environments, not merely larger language models.

Current limitations. Current AI Scientist systems inherit hallucination, bias, poor out-of-distribution generalization, and weak uncertainty calibration from foundation models (Ji et al., 2023; Gallegos et al., 2024; Wang et al., 2023; Ye et al.,

2024). They also face capability-specific limits: knowledge acquisition can miss source quality and conflicting evidence; hypothesis generation remains strongest when objectives and evaluators are explicit; experimental validation often works in simplified computational or laboratory settings; and evolution can optimize local scores rather than durable knowledge (Kusumegi et al., 2025; Romera-Paredes et al., 2024; Starace et al., 2025; Novikov et al., 2025). Domain coverage is similarly uneven, favoring fields with machine-readable data and standardized evaluation over domains requiring contextual interpretation (Grossmann et al., 2023; Sourati and Evans, 2023).

Security and ethics risks. Autonomous tool use can lower the barrier to unsafe procedures or dual-use discoveries in biology, chemistry, medicine, and cybersecurity (Urbina et al., 2022; Bengio et al., 2025; Tang et al., 2025b). Agent systems also introduce vulnerabilities such as prompt injection, insecure tool access, and unstable behavior under adversarial instructions (Debenedetti et al., 2024; He et al., 2025a). Beyond technical safety, low-cost manuscript generation may strain peer review, authorship and responsibility remain human obligations, sensitive data create privacy risks, and unequal access to compute, data, and laboratories may concentrate scientific power (Bianchi et al., 2025; Springer Nature, 2025; Resnik and Hosseini, 2025; Sarkar et al., 2024).

10 Outlook

Scientific production. The long-term impact of AI Scientists is not simply automatic paper production, but 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, uncertainty reported, 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 disciplines, from health to materials and sustainability. In both paths, their value will depend less on replacing human scientists than on expanding the scale of exploration while preserving human responsibility for problem framing, governance, and societal purpose.

11 Conclusion

AI Scientists mark a shift from using AI as an isolated scientific tool toward using AI as a coordinator of research trajectories. This survey has argued that such systems should not be evaluated only by the plausibility of their final papers or outputs, but by whether their knowledge acquisition, hypothesis generation, experimental validation, and evolution can be inspected, reproduced, challenged, and governed. The next stage of AI Scientist research requires more than stronger models. It requires auditable infrastructure, execution-grounded evaluation, human-governed autonomy, and institutional norms that preserve scientific accountability.

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.

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, and 1 others. 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, and 1 others. 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.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *EMNLP*.

Yoshua Bengio, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, and 1 others. 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 Baldassari, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*.

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 Fakhruldeen, Andrea Gabrielli, Animesh Garg, Sami Haddadin, Kanako Harada, Jason Hein, and 1 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. 2026. 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.

Shanghua Gao, Richard Zhu, Pengwei Sui, Zhenglun Kong, Sufian Aldogom, Yepeng Huang, Ayush Noori, Reza Shamji, Krishna Parvataneni, Theodoros Tsiligkaridis, and 1 others. 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, and 1 others. 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, and 1 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, and 1 others. 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.

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, and 1 others. 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, Congyu Zhang, Letian Zhang, Guiming Chen, Haoqin Tu, Xinyu Yang, Lu Feng, Xujiang Zhao, Haifeng Chen, Jiawei Zhou, Xiao Wang, and 16 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.

Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, and 1 others. 2025. A generalist medical language model for disease diagnosis assistance. *Nature medicine*, 31(3):932–942.

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, and 1 others. 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 Hutmacher. 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, and 1 others. 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 Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, and 1 others. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475.

Jennifer E Rood, Anna Hupalowska, and Aviv Regev. 2024. Toward a foundation model of causal cell and tissue biology with a perturbation cell and tissue atlas. *Cell*, 187(17):4520–4545.

Yixiang Ruan, Chenyin Lu, Ning Xu, Yuchen He, Yixin Chen, Jian Zhang, Jun Xuan, Jianzhang Pan, Qun Fang, Hanyu Gao, and 1 others. 2024. An automatic end-to-end chemical synthesis development platform powered by large language models. *Nature communications*, 15(1):10160.

Atiquer 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, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2024. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5):1–42.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Chenglei Si, Zitong Yang, Yejin Choi, Emmanuel Candès, Diyi Yang, and Tatsunori Hashimoto. 2026. Towards execution-grounded automated ai research. *arXiv preprint arXiv:2601.14525*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 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, and 1 others. 2025. Paperbench: Evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*.

Markus Stocker, Lauren Snyder, Matthew Anfuso, Oliver Ludwig, Freya Thießen, Kheir Eddine Farfar, Muhammad Haris, Allard Oelen, and Mohamad Yaser Jaradeh. 2025. Rethinking the production and publication of machine-readable expressions of research findings. *Scientific Data*, 12(1):677.

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, and 1 others. 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, and 1 others. 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, and 1 others. 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, and 1 others. 2023. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Yi Wang, Yuejie Hou, Lin Yang, Shisen Li, Wenting 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 Mengcheng Yang. 2025. 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, Juntai 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, Qiujie 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, Xinzhen 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.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, and 1 others. 2024. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188.

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, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023b. 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, Shiyang Feng, 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. *arXiv preprint arXiv:2501.03916*.

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. *arXiv preprint arXiv:2406.10833*.

Tianshi Zheng, Zheye 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. *arXiv preprint arXiv:2505.13259*.

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. 2026. 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. 2025a. 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.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025b. Deepreview: Improving llm-based paper review with human-like deep thinking process. *arXiv preprint arXiv:2503.08569*.

A 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., 2024; 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., 2023; 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 (Ding et al., 2025; 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., 2024; 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). The strongest AI Scientist systems should therefore output not only a candidate idea, but also a justification trace: why this hypothesis was proposed, what evidence supports it, what assumptions it depends on, and what would count as a meaningful test (Boiko et al., 2023; Weng et al., 2025).

B Enabling Technologies

AI Scientist systems arise from the co-evolution of several technologies that were previously developed in relative isolation (King et al., 2009; Kitano, 2021; Lu et al., 2024). The central point is not that any single component has become sufficiently powerful, but that foundation models, scientific infrastructure, and agentic orchestration can now

Table 4: 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. (2026)	LLM-assisted science	✓	✗	✗	✗	✗	✗
Luo et al. (2025)	LLMs for research	✓	✗	✗	✗	✗	✗
Alkan et al. (2025)	Hypothesis generation	✗	✗	✗	✗	✗	✗
Zhou et al. (2026)	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**	✓	✓	✓	✓	✓	✓

be coupled into a closed loop. In this loop, models propose and interpret, instruments or computational environments execute and generate observations, and agents coordinate the sequence of decisions that turns isolated actions into a research process (Angello et al., 2024; Musslick et al., 2025).

Foundation models provide broad scientific priors for language, code, images, sequences, and other scientific representations (Beltagy et al., 2019; Taylor et al., 2022; Bi et al., 2024; Singhal et al., 2023; Madani et al., 2023; Xu et al., 2024; He et al., 2025b; Liu et al., 2025). Their role is not limited to writing text. They help identify research questions, summarize literature, formulate candidate hypotheses, translate ideas into executable artifacts, and interpret results. This breadth makes them attractive as the cognitive core of AI Scientist systems. At the same time, their limitations shape the entire workflow. Hallucination, bias, outdated knowledge, weak out-of-distribution generalization, and poor uncertainty calibration can all propagate from early knowledge acquisition into later experimental decisions (Ji et al., 2023; Gallegos et al., 2024; Wang et al., 2023; Ye et al., 2024). For this reason, foundation models must be coupled with retrieval, verification, and human oversight rather than treated as self-sufficient scientific authorities.

Scientific data and retrieval infrastructure gives AI Scientist systems access to knowledge beyond static pre-training. Scientific discovery depends on current papers, databases, protocols, experimental records, code repositories, and negative results. Retrieval-augmented and tool-using systems can consult these sources during the research process, allowing the working context to be revised as new evidence appears (Lewis et al., 2020; Liu et al., 2024; Gupta et al., 2024; Yuan et al., 2025; Boiko et al., 2023). This is crucial because scientific knowledge is not a closed corpus. The quality of this infrastructure directly affects the quality of downstream hypotheses: incomplete retrieval may hide conflicting results, poor metadata may distort provenance, and uneven coverage may steer systems toward well-indexed domains while neglecting less digitized fields (Stocker et al., 2025).

Automated experimental infrastructure turns parts of scientific validation into executable workflows. In computational domains, this may involve code generation, benchmark execution, simulation, and statistical analysis (Starace et al., 2025; Yamada et al., 2025). In laboratory domains, self-driving laboratories and robotic platforms can execute protocols, collect measurements, and feed results back into the search process (Jumper et al., 2021; Mieszczanek et al., 2024; Abolhasani and Kumacheva, 2023; Boiko et al., 2023; Bran et al., 2023; Szymanski et al., 2023; Cooper et al., 2025). This infrastructure is what allows AI Scientist systems to move beyond text-level plausibility. A hypothesis becomes scientifically meaningful only when the system can connect it to observations, failures, and revised decisions (Angello et al., 2024). However, automation remains uneven across disciplines. Fields with standardized datasets, machine-readable outputs, and repeatable protocols are better suited to current closed-loop systems than fields where evidence is qualitative, contextual, or ethically constrained (Grossmann et al., 2023; Sourati and Evans, 2023).

Agentic orchestration coordinates the scientific

workflow across time (Yao et al., 2023b,a; Yamada et al., 2025). Agents decompose goals, retrieve evidence, call tools, assign roles, execute code, record results, and revise plans. Multi-agent designs can separate generation, criticism, planning, and verification, approximating some of the division of labor found in human research teams (Swanson et al., 2025; Weng et al., 2025). Memory mechanisms can preserve failed trials and intermediate artifacts so that later iterations do not restart from scratch (Tang et al., 2025a; Novikov et al., 2025). Tool-oriented systems further expand what agents can do by connecting them to scientific databases, solvers, and standardized APIs (Ding et al., 2025; Gao et al., 2025). These mechanisms are essential for moving from a collection of tools to an AI Scientist system. Yet they also introduce new failure modes: long-horizon plans may drift, tool calls may be unsafe, and local optimization may produce artifacts that satisfy internal evaluators without advancing scientific understanding.

Audit and evaluation infrastructure is therefore an enabling technology in its own right. If AI Scientists are evaluated only by their final manuscripts, the community cannot reliably determine whether the underlying process was valid (Zhu et al., 2025b; Bianchi et al., 2025). A mature system should record the hypotheses considered, sources retrieved, experiments attempted, failures encountered, human interventions made, and revisions introduced (Stocker et al., 2025; Musslick et al., 2025). Such records transform AI-driven research from a black-box artifact generator into an inspectable research trajectory. This requirement also connects directly to the theme-track motivation of this survey: the future of AI Scientists depends not only on stronger models, but on new forms of scientific accountability.

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, generate 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 access. 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, experimental safety, and corrections. 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 cen-

tral 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.

Taken together, 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 not to slow scientific discovery for its own sake, but to ensure that acceleration remains compatible with accountability, reproducibility, and public trust.