Review

A Brief History of AI for Scientific Discovery: Open Research, Metrics, and Autonomous Agents

Surasak Phetmanee *

Posted Date: 23 March 2026

doi: 10.20944/preprints202603.1694.v1

Keywords: autonomous agents; open research; scientific discovery

Review

A Brief History of AI for Scientific Discovery: Open Research, Metrics, and Autonomous Agents

Surasak Phetmanee

Department of Electrical and Computer Engineering, Faculty of Engineering, Thammasat School of Engineering, Thammasat University, Thailand; psurasak@engr.tu.ac.th

Abstract

The history of artificial intelligence for scientific discovery is not a two year story about chatbots learning to write papers. It is a sixty year story about science repeatedly handing its bottlenecks to machines—first inference, then search, then measurement, then the full workflow—only to discover that each delegation solves one problem and exposes a harder one underneath. This paper traces that history from DENDRAL (1965) through the construction of open scholarly infrastructure (arXiv, Google Scholar, ORCID), the oracle breakthroughs of AlphaFold, and the current era of LLM driven autonomous research agents. Three interlocking threads are followed including AI as research instrument, AI for research infrastructure, and the reshaping of scholarly profiles and incentives by machine readable metrics. The central tension throughout is between automation and augmentation between building systems that replace human researchers and tools that amplify human creativity and judgement. The paper presents that the most consequential development is not any single tool but the emergence of an interconnected ecosystem where AI agents, preprint platforms, open source codebases, and citation infrastructure form a feedback loop that is fundamentally restructuring who can do science, how fast discoveries propagate, and what counts as a valid scientific contribution.

Keywords: autonomous agents; open research; scientific discovery

1. Introduction

In 1965, a Nobel laureate geneticist walked into a computer scientist's office at Stanford and proposed that a machine should learn to read mass spectra—not because anyone thought computers would become scientists, but because there were too many possible molecular structures for any chemist to hold in mind. Sixty years later, an AI system in Tokyo generated a complete scientific paper in twelve minutes for less than the price of a sandwich. Between those two moments lies a history that is not really about artificial intelligence at all. It is about a recurring crisis in science itself: too many possibilities, too much literature, too little time and an argument, still unresolved, over whether the solution is to build better tools or better replacements.

The geneticist was Joshua Lederberg, who had won the Nobel Prize at thirty three for discovering bacterial conjugation. The computer scientist was Edward Feigenbaum, fresh from studying how people learn and looking for a scientific problem on which to test his ideas about induction. Their collaboration produced DENDRAL, a system for inferring molecular structures from mass spectrometry data that became one of the first major AI application to a real empirical science [21]. The Tokyo system was The AI Scientist, released by Sakana AI in August 2024, which chained together literature search, hypothesis generation, code writing, experiment execution, and manuscript drafting into a single automated pipeline [23]. The paper it produced cost approximately fifteen dollars and contained citation errors, shallow experiments, and hallucinated results. It also passed a simulated peer review.

Between DENDRAL and The AI Scientist, three rivers converged. The first was the dream of using computers as research instruments—systems that could form hypotheses and interpret experiments. The second was the construction of research infrastructure that could search, sort, and map a scientific literature too large for any scholar to survey. The third was the creation of open scholarly plumbing—preprint servers, author identifiers, citation profiles, and open access mandates—that made scientific output machine readable, machine searchable, and ultimately machine consumable. The modern agentic systems of 2024–2026 sit at the point where those three rivers meet.

This paper follows all three threads. It is organised across five eras, but the chronology serves an argument rather than merely a timeline. The argument is this: the history of AI for scientific discovery is the story of science repeatedly handing its bottlenecks to machines, only to discover that each delegation exposes a harder problem underneath, so that the deepest question—*is the machine a tool or a colleague?*—is never settled, only restated at higher stakes

2. The Expert Systems Era (1965–1989)

2.1. The Birth of Knowledge Engineering

The partnership that produced DENDRAL was not inevitable. Feigenbaum, a student of Herbert Simon at Carnegie Tech, had arrived at Stanford in 1965 to help found its computer science department. He was searching for a domain in which to study machine induction—how a program might learn from examples rather than from explicit instruction. Lederberg, working on NASA's search for extraterrestrial life, had a concrete problem: if a mass spectrometer were sent to Mars, who would interpret the spectra? Not who in the human sense—there would be no chemist on Mars. The question was whether a program could bottle enough chemical intuition to reason where no expert could go [22].

The resulting system encoded domain specific knowledge—chemical bond rules, fragmentation heuristics, structural constraints—into a search procedure that could propose molecular structures consistent with observed spectra. The key insight was that general purpose problem solving methods, which AI researchers had been pursuing since the Logic Theorist and GPS, were insufficient for real scientific problems. What worked was *knowledge*: the detailed, domain specific expertise of chemists like Carl Djerassi, who contributed the chemical rules that made DENDRAL's inferences scientifically credible [21].

This was a foundational lesson that every subsequent generation of AI for science has had to relearn. Science punishes unguided generality. DENDRAL succeeded not by being clever in the abstract but by being knowledgeable in the particular. The same principle applies, in transformed form, to today's LLM agents, which still require retrieval pipelines, tool access, and domain scaffolding to produce anything beyond plausible sounding prose.

DENDRAL's ripple effects were substantial. It inspired MYCIN, Shortliffe's rule based system for diagnosing infectious diseases [33], and helped launch the expert systems industry of the 1980s. Meta-DENDRAL, a later extension, even learned new chemical rules from data—an early form of automated knowledge acquisition [3]. But in the public narrative, DENDRAL belonged to Feigenbaum and Lederberg. Bruce Buchanan, who actually built the system and developed the knowledge engineering methodology, and Djerassi, whose chemical expertise made the rules credible, were overshadowed. This is a pattern that recurs throughout the history: the visionary and the domain celebrity receive credit; the builder and the domain contributor are footnoted. Table 1 summarises the foundational systems that established AI as a legitimate instrument for scientific research.

Table 1. Early AI systems for scientific discovery.

System	Year	Function
DENDRAL	1965	Molecular structure inference from mass spectra
Meta-DENDRAL	1976	Automated chemical rule learning from data
MYCIN	1976	Infectious disease diagnosis
BACON / DALTON	1987	Rediscovery of scientific laws from data
Robot Scientist Adam	2009	Autonomous hypothesis testing in yeast genomics

2.2. *Modelling Scientific Discovery*

A parallel strand of work asked a more radical question. If DENDRAL could reason *within* chemistry, could a program model the *logic of scientific discovery itself*? Herbert Simon and Patrick Langley, with Gary Bradshaw and Jan Zytkow, built systems like BACON and DALTON that rediscovered known scientific laws—the ideal gas law, Kepler's third law—from data [18]. The results sounded modest: the programs rediscovered what was already known. But that was precisely the point. If discovery had computational structure, then the creative processes of science were not ineffable genius but something that could, in principle, be formalised and implemented.

This line of work mattered more than practically. Working scientists did not recognise their own practice in the rediscovery of seventeenth century gas laws. But the conviction that discovery has algorithmic structure—that it is not magic but method—runs as a hidden thread all the way to today's research agents. Every system that claims to generate hypotheses is implicitly testing Simon and Langley's bet. This was augmentation dressed in the language of intelligence—the first instance of a pattern that would recur in every subsequent era.

2.3. *The Shift from Rules to Data*

The expert systems boom of the early 1980s, fuelled partly by geopolitical anxiety over Japan's Fifth Generation Computer project, created a commercial ecosystem that collapsed by the decade's end [5]. Teknowledge, IntelliCorp, and Symbolics all failed or contracted. The resulting AI winter (roughly 1987–1993) was partly a market crash and partly a signal that hand crafted rules could not keep pace with the rate at which scientific domains changed. But the deeper reason the era ended was not that expert systems failed at what they did. It was that the nature of the scientific problem shifted beneath them. By the late 1980s, science was becoming a data flood. The Human Genome Project was launching. Digital sky surveys were beginning. Mass spectrometry itself was generating data at rates no rule base could absorb. We cannot write rules fast enough when the data doubles every few years. The bottleneck had moved from expert reasoning in narrow domains to pattern extraction from vast digital corpora.

3. Science Becomes Searchable (1989–2004)

3.1. *From Rules to Patterns*

A single formulation captured the shift. In 1996, Usama Fayyad, Gregory Piatetsky Shapiro, and Padhraic Smyth defined Knowledge Discovery in Databases (KDD) as the process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [6]. That definition with its emphasis on extraction from large corpora rather than reasoning from handcrafted cases marked a conceptual break. Discovery was no longer something a program did by consulting its rules. It was something a program found by sifting through enough data.

KDD is now subsumed under the broader labels of data science and machine learning, and its originators rarely appear in AI for science narratives. They were industry adjacent (Fayyad at NASA's Jet Propulsion Laboratory, Piatetsky Shapiro at GTE Labs) rather than in high prestige academic departments, which may explain their lower visibility in a field that privileges Stanford MIT Cambridge lineages. But their conceptual contribution was the bridge between the expert systems era and the data driven era that followed.

3.2. *The Preprint Revolution*

On 14 August 1991, a particle physicist at Los Alamos National Laboratory Paul Ginsparg set up an email server—`hep-th@xxx.lanl.gov`—so that physicists could share papers before journals published them. He did it to speed distribution and to level the playing field for researchers outside elite institutions who received physical preprints months late if at all [8]. Ginsparg could not have known that he was building the most consequential piece of scientific infrastructure since the journal itself. His server, which migrated to the web and became arXiv, now holds over 2.5 million papers and has spawned an entire ecosystem: bioRxiv (2013) for biology, medRxiv for health sciences, TechRxiv for engineering, and, most recently, AgentRxiv for agent generated research. ArXiv did not automate discovery. What it did was make science machine consumable. Pre publication circulation of research, it created the searchable, linkable, freely accessible substrate on which every modern AI research agent depends. Without arXiv, The AI Scientist has no papers to read. The entire preprint ecosystem might not exist.

This is the quiet infrastructure story that makes the loud agent story possible. The agents of 2025 work only because the librarians, physicists, and policy makers of the 1990s and 2000s built the substrate. That ratchet—making science incrementally more machine readable—has been monotonic across the entire history. No era has reversed it.

3.3. The Data Flood

The era's other defining development was the explosion of digital scientific data. The Human Genome Project (1990–2003) demonstrated that biology was becoming an information science [17]. DNA microarrays generated thousands of measurements per experiment. Digital sky surveys produced terabytes of astronomical observations. The pattern that had motivated DENDRAL recurred at vastly greater scale.

The Fourth Paradigm, a manifesto published by Microsoft Research in 2009, gave this shift a name and a self conscious identity: data intensive discovery as a mode of science alongside theory, experiment, and simulation [10]. As an inflection point, the Fourth Paradigm was more diagnosis than cause—the real structural shifts were already underway in repositories, search, data growth, and automation. But it canonised a worldview that would shape the next two decades: science is now a data problem, and computation is not merely a tool but a way of knowing. By 2004, science was searchable. What nobody anticipated was what would happen when it also became *rankable*. A search engine and a single number were about to reshape academic careers.

4. When Scholarship Became a Score (2004–2016)

4.1. From Citation Counts to Incentive Systems

Google Scholar launched in November 2004, making scholarly literature findable through the same search interface people used for everything else. Jorge Hirsch, a condensed matter physicist at UC San Diego, proposed the h-index in 2005 as a modest improvement over crude citation counts—a single number capturing both productivity and impact [11]. Google Scholar Citations, launched in 2011, made those metrics visible and trackable to individual researchers at scale. ORCID, which began registry services in October 2012, solved the long standing problem of author name ambiguity. These developments created what might be called the metrics machine including a searchable, countable, publicly visible, and gameable scholarly substrate.

The paradox is that Hirsch proposed the h-index as a useful characterisation, not as the seed of a global incentive machine. His paper is measured and modest. But the timing was perfect: Google Scholar had just made citation data free and universal, and the infrastructure to compute h-indices was suddenly available to everyone. Within a decade, the h-index had become the most consequential metric in global academic culture—shaping hiring, tenure, funding, and self worth.

The second order effects were unforeseen and substantial. A culture of quantified productivity created demand for tools that could help researchers publish more, faster exactly the demand that AI writing assistants and literature agents would later fill. It also created conditions for gaming such as citation cartels, strategic self citation, and in a development that would have been unimaginable in 2005 AI generated preprints designed to inflate bibliometric signals. Retraction Watch's 2025 exposé on h-index manipulation via self citation through preprints [30] is the late but logical consequence of the infrastructure shift that Hirsch set in motion.

Meanwhile, the *reproducibility crisis* was eroding confidence in the literature that these metrics measured. The Open Science Collaboration's 2015 attempt to replicate 100 psychology studies found that fewer than 40 per cent reproduced successfully [28]. Similar problems surfaced in cancer biology, economics, and medicine. This crisis created demand for tools that could check, verify, and systematise the scientific literature. The niche that AI powered screening and review tools would fill.

4.2. The Robot in the Laboratory

Against this backdrop of infrastructure construction, a quieter development occurred that would prove more important than it seemed at the time. In 2009, Ross King and colleagues at Aberystwyth University reported in *Science* that Robot Scientist "Adam" had autonomously generated hypotheses about yeast functional genomics, designed experiments to test them, executed the experiments using laboratory robotics, and had its conclusions confirmed by human scientists [15]. Michael Schmidt and Hod Lipson published their distillation of natural laws from experimental data [32], underscoring that autonomous discovery was maturing along multiple fronts.

Adam, the robot, closed the experimental loop. Earlier systems advised, predicted, or extracted patterns. Adam actually *did science* at least in a constrained domain. King, who had trained in microbiology before pivoting to computer science at the Turing Institute in Glasgow, brought unusual credibility with both communities. Larisa Soldatova's development of the EXPO ontology for recording Adam's experimental processes deserves particular note: it was the formal record keeping, not just the hypothesis testing, that made Adam's claim to reproducible automated science credible [16].

That question whether a system operating in a constrained domain is genuinely discovering or merely executing a well specified protocol remains unresolved and recurs with each new generation of tools. It is the same question asked of DENDRAL in the 1970s, of AlphaFold in the 2020s, and of LLM research agents today. The goalposts move, but the challenge to scientific legitimacy is permanent.

4.3. The Scholarly Stack Takes Shape

The era's enduring contribution was not any single system but the stack: open, linked, searchable, metrified science. By 2016, a researcher's work was findable (Google Scholar), citable (CrossRef DOIs), attributable (ORCID), measurable (h-index, citation counts), and increasingly pre publishable (arXiv, bioRxiv). Semantic Scholar, launched by the Allen Institute for AI in 2015, applied machine learning to scholarly search itself marking the moment when AI became not just a tool for doing science but a tool for navigating science.

This infrastructure was not built for AI agents. It was built for human researchers trying to manage an exploding literature. But it had the unintended consequence of making the scholarly record legible to software at scale—a precondition for everything that followed. Table 2 traces the emergence of the open repositories including preprint servers, data repositories, identifiers, and search platforms.

Table 2. Open research repositories.

Component	Year	Function	Access
Preprint servers			
arXiv	1991	Preprints: physics, CS, mathematics	Open access
SSRN	1994	Preprints: social sciences, law, economics	Free[a]
bioRxiv	2013	Preprints: biology	Open access
PeerJ	2013	Open access publishing and preprints	Open access
Preprints.org	2016	Preprints: multidisciplinary	Open access
medRxiv	2019	Preprints: health sciences	Open access
TechRxiv	2018	Preprints: engineering	Open access
AgentRxiv	2025	Preprints: agent generated research	Open access
Data and code repositories			
Dryad	2008	Curated research data archiving	Open access
figshare	2012	Figures, datasets, media, and code	Open access
Zenodo	2013	General purpose research archive (CERN/EU)	Open access
OSF	2013	Project management, preregistration, data sharing	Open access
Identifiers, search, and metrics			
CrossRef	2000	Persistent DOIs for publications	Free
Google Scholar	2004	Universal citation search and profiles	Free
h-index	2005	Single number impact metric	—
Google Scholar Citations	2011	Public, trackable author profiles	Free
ORCID	2012	Unique author identifiers	Free
Semantic Scholar	2015	ML powered paper discovery	Free
OpenAlex	2022	Open catalogue of the global research	Open access[b]

[a] Acquired by Elsevier in 2016; free to read, author upload requires registration.

[b] Successor to Microsoft Academic Graph; fully open data and API.

5. Science in the Age of Oracles (2016–2022)

5.1. AlphaFold

AlphaFold's performance at CASP14 in 2020 was not just impressive; it was destabilising. For the first time, an AI system had not merely helped scientists work faster. It had produced a scientific answer of extraordinary value that decades of human effort had failed to reach. Protein structure prediction, a problem that had resisted conventional approaches for fifty years, was solved for single chain proteins at experimental accuracy [13]. The significance was not only scientific but rhetorical. Before AlphaFold, AI for science often sounded like workflow help such as pattern mining, expert support, literature navigation. After AlphaFold, AI could credibly claim to have *leapfrogged* human capability on a grand challenge. This shifted expectations. People stopped asking whether AI could assist scientists and started asking where else it might surpass them.

But AlphaFold also established a specific mode of AI for science that later developments would both inherit and react against the *oracle* paradigm. We pose a well defined question (what is this protein's structure?), and the model returns a high value answer. The oracle mode was successful in its own domain. Its limitation, visible only in retrospect, was that it answered questions without asking them. It solved without formulating. It predicted without explaining mechanisms. The next generation of systems would try to cross that boundary not by answering harder questions but by learning to ask them.

5.2. ASReview and the Case for Augmentation

Rens van de Schoot's path to ASReview was unconventional.[1] ASReview—an ope source framework for systematic reviews, where the system iteratively learns which papers a reviewer is likely to find relevant and prioritises them accordingly, published in *Nature Machine Intelligence* in 2021 [37].

ASReview occupies a distinctive position in the AI for science landscape. It is neither an oracle that solves problems nor an agent that runs the whole research loop. It is a tool that augments human decision making at the literature navigation bottleneck, reducing screening labour by orders of magnitude while keeping humans firmly in the loop. It is now used by hundreds of thousands of researchers worldwide and has become a standard tool in evidence based medicine. ASReview matters for this history not because it is the most technically ambitious system but because it is the most convincing demonstration that augmentation can work at scale without requiring full automation. Its emphasis on transparency, reproducibility, and human oversight represents one pole of the field's central tension and, so far, the pole with the stronger practical track record.

5.3. The Open Science Policy

The period also saw open science infrastructure mature from aspiration to mandate. UNESCO's 2021 Recommendation on Open Science established an international normative framework, framing openness as a response to global inequity and the digital divide [35]. The OSTP Nelson Memo of August 2022 required immediate public access to all US federally funded research from 2026, dropping the previous twelve month embargo [26]. The EU's Plan S took effect. The COVID-19 pandemic accelerated these trends including preprints in biomedicine surged as researchers needed rapid dissemination, and BenevolentAI's identification of baricitinib as a potential COVID treatment became a widely cited success for drug repurposing [41].

These policies were direct preconditions for the agentic era. They ensured that the raw material AI agents would need—open papers, open data, open code—would be increasingly available. Every increment of openness creates the conditions for the next generation of AI tools, which in turn create demand for further openness. This is a genuine positive feedback loop—the machine-readable ratchet—though it coexists with countervailing forces such as publisher paywalls, intellectual property restrictions, and the economic interests of companies that monetise research access.

6. The Emergence of Scientific Agents (2022–Present)

6.1. Scientific Agents

The publication of the GPT-4 Technical Report in March 2023 was not the first scientific agent. It was the moment when the field had to take seriously that a general purpose language model could read, summarise, reason, write code, and produce long text at a level sufficient to scaffold multi step research workflows [27]. Before GPT-4 models, scientific AI was typically modular like one model for structures, another for screening, another for search. After GPT-4, it became plausible to build a single language orchestration layer that could call tools, interpret papers, generate code, and draft manuscripts. The agent became a believable software form for science.

Agentic AI frameworks—LangChain, AutoGen (AG2), CrewAI—emerged as critical enabling technology, making it possible to chain language model calls with tool use, code execution, and retrieval augmented generation (combining model output with real time information retrieval) into multi step workflows. The AI Scientist, Denario, Agent Laboratory, and FutureHouse's suite are all built on such frameworks [12]. Their existence is contingent on this software infrastructure as much as on the foundation models themselves. Table 3 identifies the open source orchestration frameworks.

[1] Biographical details in this paragraph are drawn from van de Schoot's personal account: https://www.rensvandeschoot.com/story-of-asreview/.

Table 3. Agentic AI frameworks enabling multi step research workflows.

Framework	Function	Access
LangChain	LLM orchestration and tool chaining	Open source
LangGraph	Stateful multi agent graph workflows	Open source
AutoGen (AG2)	Multi agent conversation framework	Open source
CrewAI	Role based multi agent workflows	Open source
HuggingFace	Model hub, datasets, and inference API	Open source

6.2. The Prototype Researcher

On 12 August 2024, a paper appeared on arXiv claiming to describe the first comprehensive framework for fully automatic scientific discovery [23]. It had been written by another AI system. The AI Scientist, built by a small team at Sakana AI in Tokyo—Chris Lu, David Ha, and collaborators—chained together idea generation, literature search, experiment planning, code writing, experiment execution, visualisation, paper drafting, and automated peer review into a single open source pipeline.

The cost was approximately six to fifteen dollars per paper, depending on the foundation model used (see the details). The reception was polarised. Enthusiasts saw proof of concept for autonomous science. Critics, including a thorough evaluation by Baumgart et al. in *ACM SIGIR Forum*, found significant shortcomings: citation errors, missing references, shallow experimental design, and papers that mimicked the surface form of research without its depth [1]. The AI Scientist has become a Rorschach test for the field about what we see in it depends on whether we think the bottleneck in science is volume or judgement. What made the moment historically significant was not the system's quality—which was, by most independent assessments, uneven—but the fact that it existed at all. It crystallised the field's central tension. It was neither a tool that a scientist uses (like ASReview) nor an oracle that answers a well defined question (like AlphaFold). It was a system that claimed to perform the *whole research act*. Whether that claim was premature or prescient is the era's defining argument.

6.3. The Tool Landscape

The AI Scientist was neither the first nor the last entrant. By early 2026, the ecosystem had differentiated into recognisable categories.

End to end research agents attempt the full research workflow. Besides The AI Scientist, these include FutureHouse's Kosmos [7], Agent Laboratory from Johns Hopkins [31], AI Researcher from HKUDS (accepted at NeurIPS 2025), the Denario project from Cambridge and the Flatiron Institute [39], and LabClaw from Stanford and Princeton, which demonstrated a skill library of over 240 biomedical laboratory procedures [42]. FutureHouse, founded in late 2023 by Sam Rodriques and Andrew White with philanthropic backing from Eric Schmidt, represents the most well resourced experiment in purpose AI for science. Its May 2025 announcement of a promising treatment for dry age macular degeneration—discovered by AI agents with humans running the bench work—is the most concrete claim to date of an AI agent driven biological discovery. In November 2025, it spun out Edison Scientific as a for profit raising $70 million.

Literature and systematic review tools address the navigation bottleneck. ASReview remains the gold standard for systematic review [37]. Elicit, Semantic Scholar, ScholarCopilot [34], Scite.ai, ResearchRabbit, Litmaps, and Connected Papers serve overlapping niches. Knowledge mapping like OpenKnowledgeMaps, VOSviewer [36], CiteSpace [4]—visualise research landscapes. *Domain specific systems*—AlphaFold, ChemCrow [2], BenevolentAI—solve particular problems.

These systems position themselves along the automation augmentation axis. Agent Laboratory frames itself as a collaborator that lets humans focus on creative ideation. Denario rejects full automation as its goal. ASReview insists on keeping humans in the loop. The diversity is not merely branding; it reflects genuinely different views about what part of science can and should be delegated to machines. The open source fraction is high—most academic tools are released on GitHub under MIT, Apache 2, or similar licences—and tools compose into workflows through shared infrastructure.

Denario uses AG2 and LangGraph; Agent Laboratory chains arXiv retrieval, HuggingFace model access, execution, and LaTeX generation; ScholarCopilot is trained on 500,000 arXiv papers. Table 4 lists the current generation of research agents that attempt to automate the full scientific workflow; notably, all but one release their code under open source. Table 5 maps the augmentation side of literature tools, knowledge mapping platforms, and domain oracles.

Table 4. Research agents.

System	Function	Access
The AI Scientist	Full pipeline: ideation to manuscript	Open source
Kosmos (FutureHouse)	Literature synthesis, hypothesis, protocol	Partial[a]
Agent Laboratory	Code, experiments, manuscript drafting	Open source
Denario	Data analysis, simulation workflows	Open source
AI Researcher (HKUDS)	Automated research generation	Open source
LabClaw	240 and more biomedical lab procedures	Open source

[a] Individual agents are open source; the integrated Kosmos platform is proprietary.

6.4. The Integrity Crisis

The same infrastructure that enables AI assisted research also enables fraud. The integrity crisis of 2024–2026 has multiple dimensions. The Problematic Paper Screener has identified thousands of tortured phrases—awkward synonyms generated by paraphrasing software to evade plagiarism detection—across the published literature. Clear Skies analytics estimates that one in fifty papers shows patterns [38]. An analysis of AI conference review cycles found that a substantial fraction of peer reviews appeared to be AI generated [19]. Separately, AI detection tools identified over 100 hallucinated citations in papers accepted at major venues [9]. Some submitted manuscripts now contain prompt injection attacks in white font, instructing LLM reviewers to give positive assessments [20]. Metrics gaming has intensified. Retraction Watch documented in 2025 how Google Scholar h-indices could be inflated through strategic self citation via preprints posted to repositories like TechRxiv [30]. This is not a new problem about citation manipulation predates AI but AI tools have industrialised it. The same open infrastructure that makes science machine readable makes it machine gameable.

The main ethical problem, then, is not AI replaces scientists. It is that AI and open infrastructure jointly lower the cost of producing credible scientific artefacts faster than institutions can evaluate them. The peer review system, already strained, faces pressures it was not designed to withstand.

6.5. The Democratisation

AI for science tools are built in Western and East Asian institutions. FutureHouse is in San Francisco. Sakana AI is in Tokyo. Agent Laboratory is at Johns Hopkins. ASReview is at Utrecht. The most influential systems emerge from teams with access to compute, data, and talent concentrations that researchers in Thailand, Nigeria, Brazil, or India cannot easily match.

The tools are more accessible than ever. Open source models (DeepSeek, Llama, Qwen), falling inference costs, and free or low cost platforms mean that a researcher with a laptop and API access can now deploy something approximating an AI research assistant. DeepSeek R1's release in January 2025—an open source model rivalling frontier American models at lower cost, produced by a Chinese lab working under US chip export restrictions demonstrated that efficiency innovation can partially compensate for hardware constraints and accelerate global access.

The practical question is whether this constitutes genuine democratisation or merely its appearance. A researcher at Thammasat University[2] can now run the same literature agent as one at Stanford.

[2] Thammasat University, founded in 1934, is Thailand's second oldest university. The author is affiliated with the Department of Electrical and Computer Engineering in the Thammasat School of Engineering: https://engr.tu.ac.th.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2026 doi:10.20944/preprints202603.1694.v1

Table 5. Literature tools, knowledge mapping, and domain oracles.

System	Function	Access
Literature and systematic review		
ASReview	Systematic review screening	Open source
Elicit	Literature search, claim extraction	Freemium
ScholarCopilot	Citation suggestion, draft assistance	Open source
Semantic Scholar	ML powered paper discovery	Free
Scite.ai	Smart citation context analysis	Freemium
ResearchRabbit	Citation based paper discovery	Free
Litmaps	Citation network visualisation	Freemium
Connected Papers	Visual literature exploration	Free
GScholarLens	Position weighted h-index computation	Open source
Knowledge mapping		
OpenKnowledgeMaps	Visual research landscape overview	Open source
VOSviewer	Bibliometric network mapping	Open source
CiteSpace	Emerging trend detection in literature	Open source
Domain oracles		
AlphaFold	Protein structure prediction	Open source
ChemCrow	Chemistry tool augmented reasoning	Open source
BenevolentAI	Drug repurposing and target discovery	Proprietary

But the two researchers face very different institutional landscapes for turning that capability into career advancement and scientific impact. The career infrastructure—conference access, mentorship networks, journal prestige, funding systems—remains concentrated in the Global North. Open source tools lower one barrier while leaving others intact. UNESCO's framing of open science as a response to global inequity [35] captures the aspiration. Whether that aspiration is being realised is an empirical question that the field has not yet adequately investigated.

7. The Oldest Argument

7.1. Automation, Augmentation, and the Unstable Proportion

Are we building tools or colleagues? For sixty years, the answer has been both, in unstable proportions. The automation—FutureHouse/Kosmos, Sakana AI, Periodic Labs—frames the goal as building systems that conduct research. The augmentation—ASReview, ScholarCopilot, most working scientists—insists that human judgement must remain central. So far, augmentation has the stronger practical track record in everyday research; automation has the louder rhetoric and the higher ceiling of ambition. The historical pattern from every previous era is instructive. The ambitious automation vision captures attention, but the practical augmentation tools capture adoption. DENDRAL was received as knowledge engineering, not as an automated scientist. Robot scientist Adam made headlines but remained a narrow proof of concept. AlphaFold was an oracle, not a colleague. The average bench biologist or social scientist in 2026 uses AI for literature synthesis and drafting assistance, not for autonomous research campaigns. The augmentation has succeeded in the current market.

This is not a prediction that automation will never arrive. It is a caution against assuming that the most technically ambitious systems represent the most likely future. The history suggests that the real transformation comes not when machines can do what scientists do but when the research process itself is redesigned around AI capabilities when the unit of scientific work is no longer a paper but something else entirely.

doi:10.20944/preprints202603.1694.v1

7.2. Preprints, Metrics, and the Scholar Profile

No account of AI for scientific discovery is complete without addressing the infrastructure that shapes academic careers, because that infrastructure is now inseparable from the tools.

Google Scholar, Scopus, and Web of Science measure different things and count different sources [24]. Google Scholar is the most inclusive and the most easily gamed. The h-index has known limitations: it penalises early career researchers, rewards fields with higher citation norms, and has been gamed through various mechanisms [40]. Tools like GScholarLens, which computes a position weighted h-index giving more credit for first and last authorship [14], represent attempts to improve on Hirsch's original metric. But the deeper problem is structural. Once a metric becomes a target, it ceases to be a good measure—Goodhart's Law applied to scholarship—and AI tools that optimise for paper production risk turning this from a warning into a description. The ethical strategies for improving scholarly visibility—open access publishing, strategic preprinting, collaborative networks, data and code sharing, reproducibility signals—remain sound [25]. They operate in an environment increasingly distorted by AI volume, and their effectiveness depends on the integrity of the evaluation infrastructure that surrounds them.

7.3. Historical and Future

The current moment most resembles the early electrification of manufacturing in the 1890s–1920s. When electric motors first arrived in factories, they were installed as direct replacements for steam engines—a single large motor driving the same system of belts and shafts. Productivity gains were modest. The real transformation came a generation later, when factories were redesigned around distributed electric motors, each powering an individual machine. AI for science is at the *single large motor* stage. Current agents are bolted onto the existing research workflow like literature review, hypothesis, experiment, paper. The genuine transformation will come when the research process itself is reorganised around what AI makes possible continuously updated knowledge graphs, automated experimental programmes, living systematic reviews, forms of scientific communication. A cautionary echo is the expert systems of the late 1980s: ambitious claims, heavy investment driven by competitive anxiety, commercial spin offs from academic research, and growing concern about whether the technology actually delivered. The current field is more technically grounded foundation models are genuinely more capable than MYCIN but the risk of an expectations correction remains real.

7.4. Open Problems

Several questions remain unresolved.

Can autonomous agents produce genuine scientific novelty? FutureHouse's Kosmos claims to have reproduced findings from unpublished manuscripts and generated novel contributions, but independent validation at scale is still early. Independent evaluations of The AI Scientist found serious weaknesses [1]. Whether these are engineering bugs fixable with better models, or symptoms of a deeper incapacity, is the field's central question.

How should AI generated scientific claims be evaluated? No consensus exists. The field needs something equivalent to what CASP was for protein structure prediction: a blind, structured, community validated benchmark for research agent output [12]. The Deep Research Agents survey identified benchmark misalignment as a central problem including existing metrics track paper surface quality rather than epistemic value.

Can peer review survive? The system is under pressure from multiple directions simultaneously. Whether AI assistance can shore up a collapsing system or will accelerate its breakdown is genuinely uncertain.

What counts as authorship? If an agent writes the paper, runs the experiments, and generates the hypotheses, who is the author? Current norms (the International Committee of Medical Journal Editors and the Committee on Publication Ethics) require human accountability, but these norms were written for a world where the question did not arise. The emergence of agent native venues like AgentRxiv suggests the field may bifurcate rather than resolve this question.

Can these tools prevent the hollowing out of scientific craft? If AI tools optimise for producing papers rather than knowledge, the field risks a Goodhart's Law catastrophe in which the metric ceases to be a good measure precisely because it has become the target. The early evidence—floods of formulaic submissions, metrics gaming, papers that mimic the surface form of research—suggests this risk is not hypothetical. Empirical work showing that papers and patents are becoming less disruptive over time, even as volume increases [29], lends weight to this concern.

8. Conclusions

The deepest pattern in this history is the *migrating bottleneck*. Every generation of AI for science solved one bottleneck only to expose the next. DENDRAL solved the inference bottleneck in narrow domains. KDD and arXiv addressed the data and literature bottlenecks. Google Scholar and the h-index solved the navigation and visibility bottleneck while creating an incentive bottleneck. AlphaFold solved a grand prediction bottleneck. Today's research agents attempt the full workflow. But beneath all of these, the capacity to evaluate, interpret, and trust scientific claims has only grown more acute. The bottlenecks do not replace each other they accumulate. This is why progress in AI for science is real but lopsided. Every gain in speed, access, or automation has come with a new dispute about trust, value, and who gets left behind.

The deepest lesson from this history is that every previous generation was confident it understood the trajectory, and every one was wrong about the timeline and the mechanism even when it was right about the direction. DENDRAL's champions did not foresee the AI issues. The Fourth Paradigm's advocates did not anticipate that foundation models, not data mining, would become the dominant paradigm. AlphaFold's celebrants did not expect that the next major development would be agents that write papers rather than oracles that answer questions.

The future of AI in science will not be decided by whether the systems get better. It will be decided by which inheritance prevails. The expert system respect for domain knowledge, the robot scientist's insistence on closing the loop, the open infrastructure's commitment to access, or the human copilot faith that judgement cannot be delegated. The chemist's bottleneck of 1965 has become everyone's bottleneck. The question is no longer whether machines will help. It is what we are willing to let them do and what we insist on doing ourselves.

References

1. Baumgart, M., Wegmeth, L., Vente, T., Beel, J.: Evaluating Sakana's AI Scientist for autonomous research: Wishful thinking or an emerging reality towards "Artificial Research Intelligence" (ARI)? ACM SIGIR Forum (2025). arXiv:2502.14297
2. Bran, A.M., et al.: ChemCrow: Augmenting large language models with chemistry tools. Nature Machine Intelligence **6**, 525–535 (2024)
3. Buchanan, B.G., Smith, D.H., White, W.C., et al.: Applications of artificial intelligence for chemical inference. 22. Automatic rule formation in mass spectrometry by means of the Meta-DENDRAL program. Journal of the American Chemical Society **98**(20), 6168–6178 (1976)
4. Chen, C.: CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology **57**(3), 359–377 (2006)
5. Crevier, D.: AI: The Tumultuous History of the Search for Artificial Intelligence. Basic Books, New York (1993)
6. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Magazine **17**(3), 37–54 (1996)
7. FutureHouse: Platform and agent suite: Crow, Owl, Phoenix, Falcon, Finch, Robin (2024–2025). https://platform.futurehouse.org
8. Ginsparg, P.: ArXiv at 20. Nature **476**, 145–147 (2011)
9. GPTZero: GPTZero finds 100 new hallucinations in NeurIPS 2025 accepted papers (2026). https://gptzero.me/news/neurips/

10. Hey, T., Tansley, S., Tolle, K. (eds.): The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond (2009)
11. Hirsch, J.E.: An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. Proceedings of the National Academy of Sciences **102**(46), 16569–16572 (2005)
12. Huang, Y., Chen, Y., Zhang, H., et al.: Deep research agents: A systematic examination and roadmap. arXiv:2506.18096 (2025)
13. Jumper, J., et al.: Highly accurate protein structure prediction with AlphaFold. Nature **596**, 583–589 (2021)
14. Karthik, V., Anand, I.S., Mahanta, U., Sharma, G.: GScholarLens. arXiv:2509.04124 (2025)
15. King, R.D., Rowland, J., Oliver, S.G., et al.: The automation of science. Science **324**(5923), 85–89 (2009)
16. King, R.D., Liakata, M., Lu, C., Oliver, S.G., Soldatova, L.N.: On the formalization and reuse of scientific research. Journal of the Royal Society Interface **8**(63), 1440–1448 (2011)
17. Lander, E.S., et al.: Initial sequencing and analysis of the human genome. Nature **409**, 860–921 (2001)
18. Langley, P., Simon, H.A., Bradshaw, G.L., Zytkow, J.M.: Scientific Discovery: Computational Explorations of the Creative Processes. MIT Press, Cambridge, MA (1987)
19. Liang, W., et al.: Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. arXiv:2403.07183 (2024)
20. Lin, Z.: Hidden prompts in manuscripts exploit AI peer review. arXiv:2507.06185 (2025)
21. Lindsay, R.K., Buchanan, B.G., Feigenbaum, E.A., Lederberg, J.: Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project. McGraw-Hill, New York (1980)
22. Lindsay, R.K., Buchanan, B.G., Feigenbaum, E.A., Lederberg, J.: DENDRAL: A case study of the first expert system for scientific hypothesis formation. Artificial Intelligence **61**(2), 209–261 (1993)
23. Lu, C., Lu, C., Lange, R.T., Foerster, J., Clune, J., Ha, D.: The AI Scientist: Towards fully automated open-ended scientific discovery. arXiv:2408.06292 (2024)
24. Martín-Martín, A., Thelwall, M., Orduna-Malea, E., Delgado López-Cózar, E.: Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. Scientometrics **126**, 871–906 (2021)
25. McKiernan, E.C., et al.: How open science helps researchers succeed. eLife **5**, e16800 (2016)
26. Nelson, A.: Ensuring free, immediate, and equitable access to federally funded research. White House OSTP Memorandum (2022)
27. OpenAI: GPT-4 Technical Report. arXiv:2303.08774 (2023)
28. Open Science Collaboration: Estimating the reproducibility of psychological science. Science **349**(6251), aac4716 (2015)
29. Park, M., Leahey, E., Funk, R.J.: Papers and patents are becoming less disruptive over time. Nature **613**, 138–144 (2023)
30. Retraction Watch: How to juice your Google Scholar h-index, preprint by preprint (2025). https://retractionwatch.com/2025/12/08/
31. Schmidgall, S., et al.: Agent Laboratory: Using LLM agents as research assistants (2025). https://agentlaboratory.github.io/
32. Schmidt, M., Lipson, H.: Distilling free-form natural laws from experimental data. Science **324**(5923), 81–85 (2009)
33. Shortliffe, E.H.: Computer-Based Medical Consultations: MYCIN. Elsevier, New York (1976)
34. TIGER Lab: ScholarCopilot: Training LLMs for academic writing with accurate citations (2025). https://tiger-ai-lab.github.io/ScholarCopilot/
35. UNESCO: UNESCO Recommendation on Open Science (2021). https://unesdoc.unesco.org/ark:/48223/pf0000379949
36. van Eck, N.J., Waltman, L.: Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics **84**(2), 523–538 (2010)
37. van de Schoot, R., et al.: An open source machine learning framework for efficient and transparent systematic reviews. Nature Machine Intelligence **3**, 125–133 (2021)
38. Van Noorden, R.: How big is science's fake-paper problem? Nature **623**(7987), 466–467 (2023). https://doi.org/10.1038/d41586-023-03464-x
39. Villaescusa-Navarro, F., et al.: The Denario project: Deep knowledge AI agents for scientific discovery. arXiv:2510.26887 (2025)
40. Waltman, L., van Eck, N.J.: The inconsistency of the h-index. Journal of the American Society for Information Science and Technology **63**(2), 406–415 (2012)

doi:10.20944/preprints202603.1694.v1

41. Wang, H., et al.: Scientific discovery in the age of artificial intelligence. Nature **620**, 47–60 (2023)

42. Wu, Y.C., et al.: LabClaw: Skill library for autonomous biomedical research (2025). https://github.com/wu-yc/LabClaw

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.